The Mean-Variance Innovation Tradeoff in AI-Augmented Evaluations

ABSTRACT

Evaluating and selecting among numerous alternative solutions shapes the trajectory and rate of innovation. Central to this process is a fundamental tension between novelty and feasibility that evaluators, operating under bounded rationality, cannot consider simultaneously and therefore rely on heuristics to guide their evaluations. A common heuristic is criteria-sequencing, in which evaluators prioritize alternative criteria at different evaluation stages. Yet, the idiosyncratic ways evaluators sequence these criteria often introduce inconsistencies, creating significant path dependencies in the process. In this paper, we propose that artificial intelligence (AI) offers a potential lever to structure evaluators' criteria-sequencing heuristics. Leveraging a field experiment with 353 evaluators, we investigate how the sequencing of AI recommendations focusing on novelty and feasibility shapes the mean and variance of innovation among selected solutions. Our results reveal a mean–variance innovation tradeoff: a feasibility-then-novelty sequence leads to selections with higher mean innovation, whereas a novelty-then-feasibility sequence yields selections with greater innovation variance. Furthermore, a post hoc analysis uncovers that the format accompanying AI recommendations also matters. A dynamic format (i.e., interactive chatbot) increases the innovation variance among selected solutions but reduces their mean innovation relative to a static format (i.e., fixed explanatory content). Because these effects operate independently, our findings show that in AI-augmented evaluations, both the sequence of criteria and the format accompanying AI recommendations shape the mean–variance innovation tradeoff. These differences have important implications for the composition of innovation portfolios. Our paper contributes to innovation evaluation research and to emerging literature on human–AI collaboration in innovation-related contexts.

Keywords: Innovation evaluation; human-AI collaboration; generative AI and predictive AI; novelty and feasibility; field experiment.

1. Introduction

"*Evaluation is creation* (...) *It is only through evaluation that value exists.*"

(Nietzsche 1883, p. 86)

Evaluating and selecting among numerous alternative solutions shapes the trajectory and rate of innovation across firms, institutions, and the knowledge frontier. Central to this process is identifying solutions that are both *novel* (i.e., departing from established approaches) and *feasible* (i.e., technically viable within existing constraints) (Lane et al. 2025, Mount et al. 2021, Rindova and Petkova 2007). Evaluations unfold under conditions of bounded rationality: evaluators cannot simultaneously consider all solution-relevant information and must rely on heuristics, or rules of thumb, to guide their evaluation (Boudreau et al. 2016, Simon 1977). Yet, such heuristics face growing constraints as solution- and criteria-related information expands (Simon and Newell 1958, Smith 1988). Given this, recent advances in artificial intelligence (AI) hold promise for augmenting human evaluation processes by structuring information processing and mitigating cognitive constraints (Anthony et al. 2023, Doshi et al. 2025, von Krogh 2018). As organizations increasingly integrate AI into their evaluation processes, fundamental questions emerge: How can AI most effectively assist evaluators when multiple criteria must be considered? Should boundedly rational evaluators first receive AI recommendations focusing on novelty, then on feasibility, or the reverse sequence? We examine how the alternative sequencing of criteria shapes the mean and variance of innovation among the solutions that evaluators select.

Our focus on sequencing is grounded in the observation that evaluators naturally rely on *criteria-sequencing*, a heuristic involving the prioritization of alternative criteria at different evaluation stages (Baer and Zhang 2024, Lamont 2012, Terwiesch and Ulrich 2009). Yet, this heuristic often introduces inconsistencies: evaluators tend to shift their priorities across criteria in idiosyncratic ways, shaped by heterogeneous preferences, prior experiences, and the surrounding context (Harvey and Kou 2013, Mitchell et al. 2011, Reitzig and Sorenson 2013). This is evidenced by conflicting empirical observations: some studies indicate that evaluators typically screen for feasibility before novelty (e.g.,

Lane et al. 2025), while others report the opposite sequence (e.g., Sharapov and Dahlander 2025). Since sequencing heuristics introduce significant path dependencies into the evaluation process (Criscuolo et al. 2021, Elhorst and Faems 2021), idiosyncratic criteria-sequencing may result in organizations selecting solutions misaligned with their innovation goals (Böttcher and Klingebiel 2025, Chai et al. 2021, Lane et al. 2022).

A conjecture is that AI may help create greater consistency in evaluations by structuring evaluators' criteria-sequencing heuristics, providing recommendations that selectively focus on specific criteria at different evaluation stages (Bauer and Gill 2024, Chen and Chan 2024). In this perspective, AI recommendations act as levers that allow organizations to deliberately and systematically shape what evaluators prioritize. When recommendations emphasize feasibility, evaluators may prioritize solutions that are viable within existing constraints, benchmarked against past successes (Chai et al. 2021, Just et al. 2022, Rindova and Petkova 2007). By contrast, when recommendations focus on novelty, evaluators may prioritize solutions that depart from known approaches (Falchetti et al. 2022, Harvey and Mueller 2021, Criscuolo et al. 2017).

In this sense, AI recommendations operate much like "spotlights on a stage": they illuminate certain aspects of a solution while leaving others in the dark, subtly structuring the order and weighting of the cues evaluators consider (Balasubramanian et al. 2022, Simon and Newell 1958, Simon 1978). By strategically sequencing these spotlights, organizations can shape the mean and variance of innovation across selected solutions. This is crucial as both factors fundamentally influence innovation: a higher mean ensures that selected solutions are reliably strong options, while a higher variance increases the likelihood that organizations capture atypical, highly innovative solutions at the margins—albeit at the cost of more frequent failures (Augier et al. 2023, Genin et al. 2023, Katila and Ahuja 2002, March 1991).

Currently, we lack an understanding of whether, how, and why sequencing different AI recommendations might shape selection, as most human-AI evaluation studies focus on single-stage, one-shot evaluations (Bell et al. 2024, Dahlander et al. 2023, Lebovitz et al. 2021). Such single-stage focus may evade the complexities of real-world evaluation processes, which typically unfold sequentially

across multiple stages (Criscuolo et al. 2021, Lamont 2012). This raises the critical question of how the sequencing of feasibility- and novelty-focused AI recommendations shapes the mean and variance of innovation in selections.

In this paper, we aim to address this question by examining two AI-augmented sequences: *feasibility-then-novelty* and *novelty-then-feasibility*. Based on these sequences, we hypothesize a *mean-variance innovation tradeoff*. In the feasibility-then-novelty sequence, feasibility-focused AI recommendations initially guide evaluators to focus on exploitative, conventional benchmarks, filtering out solutions that are not clearly viable early on (Lane et al. 2025). Evaluators are next exposed to novelty-focused AI recommendations, leading them to prioritize the originality of these remaining solutions (Boudreau et al. 2016). With fewer options to evaluate, they may engage more deeply in more intuitive judgments about what is truly novel (Mount et al. 2021). This focused evaluation process may yield selections with a higher mean innovation than in the reverse sequence, as novelty is assessed within feasible boundaries.

In contrast, in the novelty-then-feasibility sequence, evaluators are first exposed to novelty-focused AI recommendations, leading them to prioritize the originality of solutions from the start (Li et al. 2025). By emphasizing novelty from the outset, evaluators may be more likely to explore and select atypical solutions that could be discounted if feasibility were assessed first (Chai 2017, Kuhn 1977, March 1991). However, while highly innovative solutions are often atypical, not all atypical solutions are technically viable. Consequently, because feasibility is assessed only after novelty, evaluators must identify sufficiently viable solutions from a pool already shaped by novelty considerations. As a result, this sequence may produce selections with higher variance, but lower mean innovation, than the reverse sequence.

To test our hypotheses, we conducted a preregistered field experiment examining how the sequencing of feasibility- and novelty-focused AI recommendations shapes selection. We partnered with Hackster.io, a leading crowdsourcing platform seeking to integrate AI into its evaluation process. First, to create our dataset of solutions, we launched an innovation contest on the platform, yielding 132

open-source solutions. To operationalize the innovation outcome variable, we relied on two benchmarks: expert evaluations (Baer 2012) and download counts (von Krogh et al. 2003). Second, we developed two AI models producing feasibility- versus novelty-focused 'Pass' or 'Fail' recommendations, accompanied by explanatory content (Lebovitz et al. 2022). Third, we conducted our experiment involving 353 evaluators (two-stage evaluation process; between-subjects design). To facilitate the experiment, we developed a custom web interface and randomly assigned evaluators to one of two sequences: feasibility-then-novelty (Treatment 1) or novelty-then-feasibility (Treatment 2).

Our results support the mean-variance innovation tradeoff: the feasibility-then-novelty sequence led to selections with higher mean innovation, whereas the novelty-then-feasibility sequence yielded selections with greater innovation variance. Additionally, we conducted a post hoc analysis to examine whether the *format* accompanying AI 'Pass' or 'Fail' recommendations, either static explanatory content or a dynamic conversational chatbot, also shaped selection decisions. Interestingly, we find that a similar tradeoff emerged: the dynamic format increased innovation variance but led to lower mean innovation among selected solutions than the static format. Because these effects operate independently, our findings show that, in AI-augmented evaluations, not only the sequence of criteria but also the format accompanying the AI recommendations shape the mean and variance in innovation among selected solutions.

Our paper makes three key contributions. First, we advance innovation evaluation research (Boudreau et al. 2016, Lane et al. 2022). While prior work has examined solution sequencing (i.e., the order in which solutions are evaluated; Criscuolo et al. 2021, Elhorst and Faems 2021), it has left the sequencing of criteria unexplored. We show how criteria-sequencing heuristics—the order in which evaluators prioritize alternative criteria, namely feasibility and novelty—shape selection and give rise to a mean–variance innovation tradeoff.

Second, we contribute to emerging literature on human-AI collaboration in innovation-related contexts (Boussioux et al. 2024, Doshi et al. 2025) by showing how AI recommendations may enable organizations to deliberately and systematically structure evaluators' criteria-sequencing heuristics

(Simon 1977). This provides organizations with new levers to design evaluation processes that align selection outcomes with innovation goals.

Third, our study informs practitioners designing hybrid human-AI collaboration processes (Puranam 2021, von Krogh 2018), revealing how criteria-sequencing heuristics can be shaped with AI to achieve either higher mean innovation to sustain current performance or greater variance to capture atypical solutions that might generate future competitive advantage. Our results further show that the format accompanying AI recommendations (i.e., static versus dynamic) can reinforce these patterns, offering another lever for shaping the composition and risk profile of innovation portfolios.

2. Theory

As organizations innovate and generate vast numbers of potential solutions, a fundamental challenge lies in evaluating and selecting those worth pursuing (Berg 2016, Boudreau et al. 2016, Lane et al. 2021, Terwiesch and Ulrich 2009). At the core of this process is the assessment of two essential criteria: *novelty* and *feasibility* (Arrighi et al. 2015, Boudreau et al. 2016). Novelty reflects departures from existing knowledge and approaches (Lane et al. 2022, Johnson and Proudfoot 2024), whereas feasibility concerns the technical viability of solutions within existing constraints, typically assessed based on alignment with established benchmarks of what has worked before (Lane et al. 2025, Phene et al. 2006). This 'essential tension' (Kuhn 1977) between relying on past knowledge and remaining open to novel possibilities illustrates one of the most persistent challenges in management regarding how organizations evaluate and select solutions (Abernathy and Clark 1985, Agarwal and Helfat 2009, Baer 2012, Criscuolo et al. 2017, Knudsen and Levinthal 2007, March 1991, Nelson and Winter 1982, Mount et al. 2021, von Hippel and von Krogh 2016).

Evaluating novelty and feasibility is especially challenging because selection decisions are made under conditions of bounded rationality (March and Simon 1958, Simon 1947). As evaluators cannot process all solution-relevant information, they must rely on heuristics, or simple rules of thumb, to guide their judgment (Simon and Newell 1958, Simon 1977). A common heuristic is *criteria-sequencing*, whereby evaluators assign different weights to specific criteria at various stages of the evaluation process

(Baer and Zhang 2024, Dahlander et al. 2023, Elhorst and Faems 2021, Reitzig and Sorenson 2013). Lamont (2012, p. 212) highlighted the prevalence of this heuristic in her review of the evaluation literature, suggesting that "evaluators are easily led to privilege different [criteria] at different times." In a similar vein, Baer and Zhang (2024, p. 395) suggested that "separating discussions of novelty and [feasibility] is to fit the human tendency to process information sequentially."

Yet, the specific sequence in which criteria are prioritized often unfolds idiosyncratically, shaped by factors such as evaluators' individual preferences, prior experiences, professional backgrounds, and the evaluation context (Cole et al. 1981, Harvey and Kou 2013, Harvey and Mueller 2021). This is evidenced by conflicting empirical observations on the sequence evaluators typically follow. For example, in a study of NASA-led innovation contests, Lane et al. (2025) found that evaluators typically screened solutions for feasibility first, checking whether they met minimum technical viability thresholds based on established benchmarks, before assessing novelty. By contrast, Sharapov and Dahlander (2025) reported the opposite sequence in accelerator programs, where investment boards tended to first screen proposals based on their novelty before considering feasibility.

Such idiosyncrasies in criteria-sequencing heuristics across evaluators introduces inconsistencies into the evaluation process, shaping which solutions ultimately advance and the level of innovation they embody (Elhorst and Faems 2021, Kheirandish and Mousavi 2018, Lane et al. 2022, Mitchell et al. 2011, Lamont 2012). The order in which novelty and feasibility are assessed acts as an initial filter, meaning that solutions excluded early based on the prioritization of one criterion never receive assessment on the other, regardless of their overall potential (Baer and Zhang 2024, Bastani et al. 2025). These variations can, in turn, shift the composition and risk profile of innovation portfolios, shaping both the mean and variance of selected solutions. A higher mean innovation ensures that selected solutions are reliably strong options, whereas a higher variance increases the likelihood of capturing boundary-pushing solutions that could drive breakthroughs—albeit at the cost of more frequent failures (Augier et al. 2023, Böttcher and Klingebiel 2025, Genin et al. 2023, Katila and Ahuja 2002, Klingebiel et al. 2022, March 1991).

Given the differential effect of sequencing on the mean and variance of innovation among selected solutions, this raises the central question of how prioritizing feasibility before novelty, or vice versa, shapes selection, and what levers can systematically structure evaluators' sequencing heuristics to ensure that selection decisions align with innovation goals.

2.1. AI-Augmented Evaluation

The search for new levers that can systematically structure evaluators' sequencing heuristics becomes particularly important in the age of artificial intelligence (AI), as evaluation processes are increasingly structured and augmented by AI recommendations that influence evaluators' judgment (Amabile 2020, Anthony et al. 2023, Balasubramanian et al. 2022, Choudhary et al. 2023, Puranam 2021, von Krogh 2018). AI systems may be leveraged to generate recommendations that emphasize novelty and feasibility at different stages (Bauer and Gill 2024, Chen and Chan 2024). For example, recommendations could highlight feasibility early in the evaluation process or rather focus on novelty first. This opens new opportunities to leverage AI to guide evaluators toward sequences of novelty and feasibility assessments likely to yield selections aligned with their innovation goals.

AI systems designed to support evaluation vary considerably in their underlying logic and capabilities; thus, it is helpful to distinguish between predictive and generative AI systems (Levinthal 2025, Raisch and Fomina 2025, Russell and Norvig 2021). Predictive AI systems—such as those based on supervised machine learning, Bayesian inference, and reinforcement learning—analyze past training data to identify patterns that explain previous successes and then extrapolate these patterns to predict future outcomes (Agarwal et al. 2019, Choudhury et al. 2021, Lazar et al. 2025). Because feasibility is often evaluated against what has worked before, predictive systems are particularly well-suited to generate feasibility-focused recommendations, drawing on concrete, quantifiable parameters such as technical viability, documentation completeness, and component specifications (Bell et al. 2024, Senoner et al. 2022). While these systems could, in principle, flag solutions that deviate from historical patterns, such outliers do not necessarily constitute novelty per se (Li et al. 2025, Lou and Wu 2021, Shrestha et al. 2021). Instead, novelty requires recombining knowledge in ways not previously explored (Kuhn 1977,

Schumpeter 1939, Uzzi et al. 2013). In effect, predictive AI systems are often more effective for feasibility-focused than for novelty-focused recommendations.

In contrast, generative AI systems—such as those based on large language models (LLMs), generative adversarial networks, and diffusion models—are designed to create new content by recombining existing knowledge in novel ways (Boussioux et al. 2024, Goodfellow et al. 2014, Zhou and Lee 2024). For instance, although LLMs technically predict the next token, they are considered generative because they transform input tokens through a learned vector space and decode them into new sequences, producing content that extends beyond memorized patterns in the training data (Brown et al. 2020, Vaswani et al. 2017). This generative capacity makes systems particularly well-suited to produce novelty-focused recommendations, as they can surface unconventional recombinations of knowledge that human evaluators might overlook (Chen and Chan 2024, Doshi et al. 2025). Yet, assessing feasibility typically depends on structured, domain-specific data with clearly defined performance criteria—precisely the contexts where predictive systems excel. Generative AI systems, by contrast, are designed to model complex data distributions and generate new samples from them, enabling exploration of broad and open-ended search spaces based on learned associations rather than explicit outcome optimization (Brown et al. 2020, Vert 2023). As a result, their recommendations may lack grounding in the precise, context-specific constraints or benchmarks that determine whether a solution is likely to work in practice.

Thus, generative AI is particularly well-suited for novelty-focused recommendations, while predictive AI excels at feasibility-focused ones, highlighting their potential complementarity in evaluation processes. Their recommendations may serve as distinct levers guiding evaluators to prioritize alternative criteria at different evaluation stages (Bauer and Gill 2024, Chen and Chan 2024). Metaphorically, AI recommendations operate like “spotlights on a stage” that can be directed to illuminate certain aspects of a solution first and bring others into focus later (Simon 1978, Smith 1988). By strategically sequencing these spotlights, evaluators can structure their heuristics throughout the process, ultimately shaping the mean and variance of innovation across selected solutions. This complementarity between predictive and

generative AI thus provides a compelling lens to examine how the sequencing of feasibility and novelty assessments shapes selection.

Despite these promising complementarities, little is known empirically about how predictive and generative AI might be leveraged together. Existing research on human-AI collaboration in evaluation largely focuses on either generative AI (e.g., Csaszar et al. 2024, Doshi et al. 2025, Teo et al. 2025, Zhong 2025) or predictive AI (e.g., Bell et al. 2024, Gaessler and Piezunka 2023, Krakowski et al. 2025, Senoner et al. 2022), but not both in tandem. This lopsided focus stems from the fact that most human-AI evaluation research has concentrated on single-stage assessments, even though real-world evaluation typically unfolds across multiple, sequential stages (Dahlander et al. 2023). Most studies examine one-shot AI-assisted tasks, particularly in medical diagnostics (e.g., Lebovitz et al. 2021, 2022, Mei et al. 2020, Teo et al. 2025, Yin et al. 2025), with few investigating AI integration at specific stages of multi-stage processes, including early screening (e.g., Bell et al. 2024, Just et al. 2024) or later-stage selection (e.g., Csaszar et al. 2024, Dell'Acqua et al. 2025).

As a result, we lack a scholarly understanding of whether, how, and why feasibility-focused recommendations from predictive AI and novelty-focused recommendations from generative AI may be sequenced to augment human evaluators throughout the evaluation process from early-stage screening to late-stage assessments. Gaining such understanding is particularly important because the sequence in which criteria are prioritized may fundamentally shape selection, particularly the mean and variance of innovation among selected solutions.

2.2. Sequencing in AI-Augmentated Evaluation

We use feasibility-focused recommendations from predictive AI and novelty-focused recommendations from generative AI to hypothesize selection outcomes under two sequences: *feasibility-then-novelty* and *novelty-then-feasibility*. We theorize that these AI-augmented sequences give rise to a *mean-variance innovation tradeoff* in evaluation.

When AI recommendations are presented in a feasibility-then-novelty sequence, we propose that evaluators are likely to select solutions with higher mean innovation than in the reverse sequence. The

feasibility-then-novelty sequence closely aligns with evaluators' natural decision-making heuristics (see Lane et al. 2025). In many evaluation contexts, decision-makers adopt an elimination-by-aspects approach when faced with multiple criteria: they first apply objective constraints to narrow the choice set, then explore the remaining options more subjectively (Tversky 1972). While this sequence may not be universal (see Sharapov and Dahlander 2025), it is commonly embedded in innovation stage-gate processes, where feasibility hurdles typically precede novelty assessments (Cooper 1990, Ettlie and Elsenbach 2007, Li et al. 2025). Presenting AI recommendations in this familiar sequence may help evaluators integrate them seamlessly and intuitively into their evaluation processes.

Consequently, this alignment with existing heuristics may help mitigate cognitive load (Simon 1977), enabling evaluators to engage more deeply with AI recommendations at each stage. When evaluators are first exposed to feasibility-focused recommendations from predictive AI, they may efficiently identify solutions that meet minimum feasibility thresholds, helping them screen out those that lack technical viability or sufficient documentation. This aligns with prior work demonstrating the effectiveness of predictive AI in feasibility filtering (see Bell et al. 2024, Just et al. 2024). In turn, doing so significantly reduces uncertainty about which solutions can realistically be implemented (Senoner et al. 2022). Since predictive AI draws on concrete, quantifiable features to make these assessments, the resulting 'feasibility floor' may appear more objective, verifiable, and trustworthy (Choi et al. 2025, Vanneste and Puranam 2024). With the potential downside constrained by this confirmed feasibility, evaluators may be more willing to take greater risks in the subsequent novelty stage with recommendations from generative AI. In short, because novelty is assessed within feasible boundaries, evaluators can engage more deeply with the remaining solutions to identify which are truly original and unique, raising the overall mean innovation of their selections compared to the reverse sequence. Thus, we hypothesize:

Hypothesis 1. *In AI-augmented evaluations, evaluators who follow a* ***feasibility-then-novelty*** *sequence will select solutions with a* ***higher mean innovation*** *than those who follow a novelty-then-feasibility sequence.*

By contrast, a novelty-then-feasibility sequence may cast a wider initial net, allowing evaluators to explore a broader solution space and surface atypical solutions that a feasibility-then-novelty sequence might otherwise overlook (Chai 2017, Kuhn 1977). During initial novelty assessments, generative AI may play a key role by highlighting atypical patterns, creative interpolations, and unconventional recombinations across the solution space (Boussioux et al. 2024, Chen and Chen 2024). However, such atypical solutions, featuring new, unconventional approaches and combinations (Ferguson and Carnabuci 2017, Vakili and Kaplan 2015), may carry a higher risk of failing to meet basic feasibility requirements (Lane et al. 2025).

Consequently, a novelty-then-feasibility sequence may create a more heterogeneous pool for subsequent feasibility assessment. Critically, by the time feasibility is considered, evaluators have already invested cognitive resources in exploring the novelty underlying each solution. Because novelty is naturally viewed as a potential gain, a high reward that people tend to overweight, it holds greater pull than feasibility, which is viewed merely as meeting basic requirements or constraints (Kahneman and Tversky 1979, Gilovich et al. 2002). By focusing on novelty first—a criterion inherently more subjective than concrete feasibility assessments (Falchetti et al. 2022)—evaluators may develop greater investment in the solutions they identify, shaping their subsequent judgments (Tversky and Kahneman 1974). When later exposed to feasibility-focused recommendations from predictive AI, they may be inclined to preserve the potential gains they have identified, making them more likely to interpret implementation challenges as surmountable rather than disqualifying. In other words, feasibility assessments become more relaxed and porous: evaluators may be invested in maintaining their diverse, atypical solutions that stricter, feasibility-focused gate-keeping standards would typically filter out in a feasibility-then-novelty sequence (Chai 2017, Lane et al. 2025).

This porous filtering may help retain highly innovative solutions, often too atypical and unconventional to fit standard feasibility requirements, but it may also allow a greater number of barely viable solutions to remain in the selection (Chai et al. 2021, Lane et al. 2022, Rindova and Petkova 2007). As the resulting selection contains a broad mix of both highly innovative solutions and less viable ones,

this sequence may ultimately produce selections with higher innovation variance, but lower mean innovation, than the reverse sequence. We hypothesize:

Hypothesis 2. *In AI-augmented evaluations, evaluators who follow a* ***novelty-then-feasibility*** *sequence will select solutions with* ***greater innovation variance*** *than those who follow a feasibility-then-novelty sequence.*

3. Method

3.1. Empirical Context

We partnered with Hackster.io,[1] a crowdsourcing platform with over 2 million users across 150 countries that runs innovation contests for companies. Solutions on Hackster.io are tangible, hardware-based prototypes accompanied by descriptions, images, and videos that detail their working principles and assembly procedures (Ghaleb et al. 2022). Hackster.io seeks to integrate AI into its evaluation processes to better serve these companies, which often struggle to evaluate the large volume of solutions generated by these contests. Addressing this challenge, prevalent in innovation contests (Terwiesch and Ulrich 2009), is central to Hackster.io's strategy for enhancing its value proposition for scalable growth.

To explore how AI may be integrated into Hackster.io's evaluation processes and test out hypotheses, we conducted a preregistered field experiment.[2] The experiment involved a two-stage evaluation process designed to test how feasibility-focused and novelty-focused recommendations from AI, integrated via different sequences, shape selection decisions. As detailed below, we first generated a dataset of solutions for evaluation and developed two systems to produce feasibility-focused (using predictive AI) and novelty-focused (using generative AI) recommendations. We then conducted a field experiment with 353 evaluators and analyzed the resulting data.

3.2. Evaluation Dataset and Innovation Benchmarks

[1] See https://www.hackster.io (accessed on May 13, 2025).
[2] Anonymized pre-registration link: https://osf.io/pa2qm?view_only=45f56956dc884d378eb7468a06398814. The study has also been approved by the ethics commission of our university.

To generate a dataset of solutions for evaluation, we partnered with Hackster.io to launch a six-month innovation contest, from March 1 to September 2024, supported by several major tech companies (Arm, Blues, PCBWay, M5Stack, Seeed Studio, and DFRobot). We focused the contest on innovation in the disability space because it is resource-constrained, rife with critical evaluation bottlenecks, and home to significant unmet needs. This space offers potential for both innovation and social impact (Park et al. 2023).

Before the contest launch, we consulted with Hackster.io's CEO and representatives to refine the contest design and determine appropriate financial incentives, with a competitive prize pool of approximately $20,000. With this approach, we ensured transparency and control over the solution generation process, which was essential for determining whether the AI-generated recommendations produced were reliable and aligned with the contest's expectations. Doing so also allowed us to align the contest design with Hackster.io's need for ecological validity, while maintaining ownership of the dataset and promoting meaningful social impact.

Overall, the contest generated a dataset of 132 solutions. To operationalize the innovation variable, we relied on two benchmarks: expert evaluations of innovation and download counts. On the one hand, for expert evaluations (Baer 2012), we engaged 21 experts in the disability innovation space. Each expert evaluated an average of 25 solutions to ensure a manageable cognitive load (Boudreau et al., 2016), with each solution rated by four experts. Experts evaluated solutions' innovation, measured as the mean of novelty and feasibility (Mount et al. 2021, Mueller et al. 2018). As expert evaluations were reliable (ICC(1, 4) for consistency = 0.910, ICC(2,4) for absolute agreement = 0.836), we averaged their evaluations to represent their group-level evaluation of innovation (*Mean* = 2.508, *SD* = 0.929) (Criscuolo et al. 2017).

On the other hand, as the solutions from our innovation contest were open source, we assessed their reach beyond the innovation contest by tracking their download count. Download count is an established innovation proxy for user innovation and open-source research (Grewal et al. 2006, von Krogh et al. 2003). At the time of writing—nine months after the contest—the open-source solutions had

been downloaded over 30,000 times, underscoring the contest's broad impact (von Hippel 2006). Given the presence of outliers and the high standard deviation (*Mean* = 255.864, *SD* = 528.672), we applied interquartile (IQR)-based discretization (see Fox 2015, Tukey 1977) to rescale download counts to ordinal categories reflecting their position relative to the lower quartile (= 1), interquartile (= 2), and upper quartile (= 3), thereby reducing the impact of extreme values and skewness for more robust comparisons.

For both expert evaluations and download counts, we split solutions into two groups: above the median ('Pass') and below the median ('Fail'). These labels were moderately correlated (Pearson's r = 0.300, $p < 0.001$), showing some overlap between the two benchmarks while also capturing distinct aspects of innovation. This split allowed us to compare AI 'Pass' or 'Fail' recommendations against these two benchmarks (see Section 3.3.3 for details).

3.3. Feasibility-focused and Novelty-focused AI Recommendations

We developed two systems producing recommendations: a predictive AI system, trained on extensive Hackster.io data, producing *feasibility-focused* recommendations extrapolating identified patterns in past successes (i.e., implementation-based features that historically increased the likelihood of winning), and a generative AI system, given only a few examples and without access to that training data, producing *novelty-focused* recommendations (i.e., concept-level cues that highlight atypical, creative approaches). Recommendations were accompanied by an overall rating ('Pass' or 'Fail') and explanatory content (see the Appendix, Table A1 for examples). If the recommendation indicated 'Pass', this meant that the system suggested selecting the solution, while 'Fail' meant the system advised against selecting it.

3.3.1. Feasibility-focused Recommendations with Predictive AI. Predictive AI recommendations focused on feasibility-related aspects such as the degree of implementation, workability, and documentation, extrapolating identified patterns in training data to generate forecasts. Naturally, the training data did not include solutions from our innovation contest. For training data, we relied on publicly available archival data from Hackster.io, comprising 6,516 solutions submitted across 112 contests. For feature engineering, we extracted the following attributes from each solution: description length, number and type of visuals (images, videos, their duration, CAD files, and

schematics), number of components used (i.e., bill of materials size), presence of a code file, and number of lines of code (adapted from Bell et al. 2024). The target variable was whether the solution won an innovation prize in its respective contest, with 387 of 6,516 solutions (6%) receiving a prize. Following these steps, our training data was split into a training set and a test set at an 80:20 ratio (Choudhury et al. 2021). Then, we tested a range of models (e.g., k-nearest neighbors, random forest, gradient boosting, support vector machines, single- or double-layer neural networks), and identified multi-layer perceptron (MLP)—a feedforward neural network with multiple hidden layers (Rumelhart et al. 1986)—as the top-performing model based on established metrics (i.e., accuracy, recall, F1-score).

The system outputs a continuous score between 0 and 1, indicating the feasibility of a solution based on the features described above. Based on these scores, solutions in the top half were labeled 'Pass,' while those in the bottom half were labeled 'Fail.' In addition to the 'Pass' or 'Fail' recommendations, the system produced static, feasibility-focused explanatory content comparing each solution's "description and bills of materials" and "visuals, code, and other documentation" with those of other solutions submitted to the contest to justify the recommendations (see the Appendix, Table A2 for details on the development of the system).

3.3.2. Novelty-focused Recommendations with Generative AI. Generative AI recommendations focused on novelty-related aspects of solutions (content-level features). We prompted OpenAI's model GPT-4.1,[3] at the time a state-of-the-art large language model (LLM), to generate 'Pass' or 'Fail' recommendations accompanied by explanatory content to justify its recommendation (see the Appendix, Table A3 for the full prompt text). The explanatory content emphasized the "novelty of the solution" as well as the "usefulness of the solution," included to contextualize novelty by indicating whether solutions, however novel, addressed the needs of people with disabilities—the focus of our innovation contest. We adopted a few-shot learning approach (see Brown et al. 2020, Boussioux et al. 2024), embedding five input–output examples directly in the prompt to demonstrate the desired evaluation format and focus on novelty, enabling the LLM to generate appropriate responses without

[3] Version: gpt-4_1-2025-04-14. See: https://platform.openai.com/docs/models/gpt-4.1 (accessed on May 13, 2025).

additional training. To avoid data leakage and ensure the integrity of our evaluation, we sourced these examples from the previous iteration of this disability-focused innovation contest on Hackster.io, ensuring examples were entirely separate from our dataset of solutions used for the evaluation experiment. We iteratively refined our prompt to ensure the recommendations were evenly split, with half labeled 'Pass' and half labeled 'Fail,' and manually reviewed each generated content.

Moreover, to capture the full potential of generative AI recommendations, we examined two formats it now enables: *static*, through fixed explanatory content, and *dynamic*, through interactive conversations with an LLM-based chatbot (see the Appendix, Table A3 for the full prompt text). To avoid confusion, evaluators were randomly assigned to one format only. In the static format, each 'Pass' or 'Fail' recommendation was accompanied by explanatory content (as described above). In contrast, in the dynamic format, evaluators received the same 'Pass' or 'Fail' recommendations but interacted with a chatbot to exchange and solicit advice about the evaluated solution instead of reading fixed, static explanatory content. To guide the conversation while keeping the focus on novelty, the chatbot began with the message: "*What is your overall evaluation of the solution, especially regarding novelty and usefulness? I am happy to share my thoughts with you—feel free to ask any questions about the solution*."

3.3.3. AI Recommendations' Alignment with Benchmarks. The overall alignment of the AI systems was calculated as the proportion of correct classifications, where the systems' 'Pass' or 'Fail' recommendations matched whether expert-rated innovation scores or download counts were above or below the median. Our systems demonstrated high alignment: for expert evaluations of innovation, predictive AI reached 80% alignment, while generative AI achieved 82% (see Bell et al. 2024, Just et al. 2024, Doshi et al. 2025). Using download counts as the benchmark, both systems reached 82% alignment. Overall, the systems' outputs were strongly, albeit not perfectly, correlated (Pearson's $r = 0.758$, $p < 0.001$), suggesting that the two systems' recommendations captured distinct aspects of innovation (see the Appendix, Figure A1, for a bar chart of alignment results). This high correlation likely stems from the fact that highly innovative solutions tend to be both novel and feasible, while poor solutions are often neither, which leads to overlapping 'Pass' or 'Fail' recommendations, even though the AI systems capture distinct

criteria, novelty versus feasibility, of innovation. Naturally, the explanatory content accompanying the 'Pass' or 'Fail' recommendations—whether emphasizing novelty or feasibility—differed markedly between the two AI systems.

3.3.4. Explanatory Content Standardization. To isolate the effects of recommendation type, we ensured that the explanatory content was comparable in structure, length, and sentiment valence. First, for structure, both types of recommendations consistently included two bullet points: predictive AI emphasized the feasibility and degree of implementation, while generative AI focused on novelty and its relevance to the needs of the disability community. Second, we ensured similar word length for the explanatory content: generative AI (*Mean* = 246.250, *SD* = 23.015) and predictive AI (*Mean* = 247.220, *SD* = 34.941) ($M_{\text{difference}}$ = -0.970, $U(130)$ = 8521.5, 95% CI = [-31.141; 29.201], p = 0.759, d = -0.033). Third, we controlled for sentiment valence (ranging from -1 to +1, with 0 representing a neutral sentiment): generative AI (*Mean* = 0.092, *SD* = 0.040) and predictive AI (*Mean* = 0.096, *SD* = 0.077) ($M_{\text{difference}}$ = -0.004, $U(130)$ = 9140.0, 95% CI = [-0.017; 0.010], p = 0.491, d = -0.063).

3.4. Experimental Sample, Design, and Procedures

To facilitate the pre-registered field experiment (between-subjects design), we developed a custom web interface (screenshots of the interface are provided in the Appendix, Table A4). As noted earlier, our two-stage experimental design assessed how the timing of integrating feasibility-focused (using predictive AI) versus novelty-focused (using generative AI) recommendations shaped evaluation outcomes. Hence, our experiment included two conditions, with the following recommendations: *Treatment 1: feasibility-then-novelty*; *Treatment 2: novelty-then-feasibility*. As mentioned above, the evaluators were further divided into receiving either static or dynamic formats (i.e., fixed explanatory content or an interactive, conversational chatbot) when receiving novelty-focused generative AI recommendations.

Data collection occurred over two sessions (April 30 and May 5, 2025). Evaluators were engineering and computer science Bachelor students from an elite Indian university, specialized in those technical fields. It goes without saying that evaluators were blind to our manipulation and hypotheses. We

selected this sample of evaluators because we were interested in how individuals with technical expertise in hardware, but not in the disability space, would incorporate AI recommendations into their evaluation decisions.

Prior to the sessions, we consulted university representatives to design appropriate incentives. Based on their input, we provided evaluators with signed certificates of participation from our U.S.- and Europe-based institutions. In addition, after the experiment, we also delivered lectures on recent advances in AI research, followed by Q&A sessions—a rare opportunity for students to interact with international scholars. Furthermore, to motivate performance, evaluators were informed that the top 40 would receive certificates of excellence and a monetary reward of about 4,000 Indian Rupees (about $50; half of the rewards allocated to each condition for fairness). Evaluators were ranked based on how closely their selected solutions matched expert innovation evaluations. Thus, evaluators were incentivized to identify the "best" solutions.

In total, 353 evaluators enrolled in the study, completed control variable measures (e.g., demographics) via the web interface, and reviewed the study instructions (summary statistics in Table 1). Then, they were randomly assigned to one of the two conditions (i.e., sequences). To familiarize evaluators with the web interface and evaluation process, all evaluators completed a tutorial consisting of three sample solutions from the previous iteration of the contest on Hackster.io, along with example recommendations similar to those they would encounter in the first stage of their assigned condition.

[Insert Table 1]

In each condition, evaluators evaluated a subset of 20 solutions, randomly sampled from the pool of 132 solutions, resulting in 7,060 evaluator-solution pairs. In the first stage, evaluators had to pass half of the solutions and fail the other half. Then, in the second stage, they re-evaluated the 10 solutions they had passed in the first stage, and again passed half and failed the other half, resulting in a selection of five solutions they considered the "best." By fixing the number of solutions evaluators could pass or fail at each stage, our design ensured greater consistency and comparability across evaluators compared to a fully discretionary approach. Evaluators were instructed to spend between four and seven minutes per

solution, had the opportunity to revise their decisions at the end of each stage, and were encouraged to select solutions that were feasible and novel. To prevent bias toward or against the recommendations, evaluators were told: "*Use your own critical judgment. Do not ignore or overly rely on the AI-generated recommendations*" (see the Appendix, Table A4 for the full instructions).

3.5. Variables, Measures, and Estimation Approach

3.5.1. Dependent Variables. The dependent variables captured the mean and variance of innovation ratings of the selected solutions. Variance referred to the spread of innovation scores, indicating how much selected solutions differ from one another in terms of their innovativeness. Both dependent variables were operationalized using two innovation-related benchmarks: expert evaluations and download counts (see Section 3.2).

3.5.2. Independent Variables. Our main independent variable, *Randomized Sequence*, was a binary variable corresponding to the assigned experimental condition, capturing the sequence of recommendations (T1: feasibility-then-novelty, that is, predictive AI followed by generative AI = 1, or T2: novelty-then-feasibility, that is, generative AI followed by predictive AI = 0). Additionally, we included *Randomized Generative AI*, a binary variable corresponding to the assigned generative AI format (static, fixed explanatory content = 0; dynamic, interactive chatbot = 1; see Section 3.3.2).

3.5.3. Control Variables. Based on existing research and input from the university representatives, we controlled for gender (0 = male, 1 = female), age (continuous), level of education (0 = Bachelor's 1st or 2nd year, 1 = Bachelor's 3rd or 4th year), AI expertise (based on field of education, 0 = not computer science, 1 = computer science) (Jeppesen and Lakhani 2010), domain expertise (having close family member(s) or friend(s) with disabilities, 0 = no, 1 = yes) (Park et al. 2023). Finally, we added total completion time as a post hoc control variable.

3.5.4. Estimation Approach.

[Note: This section has been omitted due to the Strategy Science Conference 35-page page limit]

4. Results

4.1. Randomization and Manipulation Checks

4.1.1. Randomization Check. We found no statistical evidence of imbalance in our control variables across conditions (*ps* > .511; see Table 2 for statistics), indicating that our randomization was effective.

[Insert Table 2]

4.1.2. Manipulation Check. Since evaluation requires human insight, our AI recommendations were designed to inform rather than replace judgment; thus, we first examined whether this was indeed the case. Overall, evaluators' compliance ranged from 64% to 74%, which is well justified given the AI systems' strong alignment with expert evaluations and download counts (see Section 3.3.3). Still, evaluators overrode AI recommendations in 26% of cases in Stage 1 and 36% in Stage 2, indicating areas of disagreement and confirming that human judgment remains central (Lebovitz et al. 2022, von Krogh 2018). Compliance did not vary by sequence, generative AI format (static versus dynamic), or alignment with benchmark recommendations (see Appendix Table A5 for mixed-effects models predicting AI compliance). Together, these results demonstrate that our manipulation worked as intended: AI recommendations informed rather than replaced evaluators' judgments.

Next, we turn to testing our two main hypotheses at the evaluator–solution level.

4.2. Hypothesis Testing

4.2.1. Hypothesis 1. *Hypothesis 1* theorized that evaluators following a feasibility-then-novelty sequence (i.e., first receiving feasibility-focused AI recommendations, followed by novelty-focused AI recommendations) would select solutions with a higher mean innovation than those assigned to the novelty-then-feasibility sequence (estimation approach in Section 3.5.4). Figure 1 illustrates the results, and Table 3 presents the results of our mixed-effects models predicting solution selection.

[Insert Figure 1 and Table 3]

First, we find no evidence that the sequence influenced the likelihood of a solution being selected ('Randomized Sequence' row; Model 1: β = -0.004; p = 0.952), even after controlling for solution innovation, the interaction between sequence and innovation, generative AI format, and other covariates,

both when benchmarking against expert evaluations (Models 2-4: $ps > 0.391$) and download counts (Models 5-7: $ps > 0.526$).

Second, we find that, overall, more innovative solutions were more likely to be selected by evaluators ('Solution Innovation Rating' row) both when benchmarking against expert evaluations (Model 2: odds ratio = 2.465; $\beta = 0.902$; $p < 0.001$; Model 3 controlling for the interaction between sequence and innovation: odds ratio = 2.221; $\beta = 0.798$; $p < 0.001$; Model 4 additionally controlling for generative AI format and other covariates: odds ratio = 2.489; $\beta = 0.912$; $p < 0.001$) and download counts (Model 5: odds ratio = 2.921; $\beta = 1.072$; $p < 0.001$; Model 6 controlling for the interaction between sequence and innovation: odds ratio = 2.633; $\beta = 0.968$; $p < 0.001$; Model 7 additionally controlling for generative AI format and other covariates: odds ratio = 2.784; $\beta = 1.024$; $p < 0.001$). Thus, evaluators engaged seriously with the experiment and tended to recognize value.

Third, supporting our hypothesis, we find that evaluators assigned to the feasibility-then-novelty sequence selected more innovative solutions than those in the reverse sequence ('Randomized Sequence × Solution Innovation Rating' row), both when benchmarking against expert evaluations (Model 3: odds ratio = 1.242; $\beta = 0.217$; $p = 0.004$) and download counts (Model 5: odds ratio = 1.237; $\beta = 0.213$; $p = 0.022$). This effect remains consistent after controlling for generative AI format and other covariates, both when benchmarking against expert evaluations (Model 4: odds ratio = 1.237; $\beta = 0.213$; $p = 0.005$) and download counts (Model 7: odds ratio = 1.236; $\beta = 0.212$; $p = 0.023$).

To obtain more granular insights into results, testing our hypothesis, we compare the interaction between sequence and innovation at each stage (Models 8–11, all of which control for the interaction between sequence and innovation and for generative AI format and additional covariates). For Stage 1 (i.e., evaluators passing 10 solutions out of the 20 initially received; 'Randomized Sequence × Solution Innovation Rating' row), we find that more innovative solutions were more likely to be selected in the feasibility-then-novelty sequence compared to the reverse sequence, both when benchmarking against expert evaluations (Model 8: odds ratio = 1.298; $\beta = 0.261$; $p < 0.001$) and download counts (Model 9: odds ratio = 1.290; $\beta = 0.255$; $p = 0.001$). Yet, surprisingly, for Stage 2 (i.e., evaluators selecting five final

solutions out of the 10 already passed in Stage 1), we find no evidence that evaluators' final selections within the already passed pool differed by condition, when benchmarking against expert evaluations (Model 10: $\beta = 0.102$; $p = 0.259$) as well as download counts (Model 11: $\beta = 0.086$; $p = 0.434$). This suggests that differences in innovation of final selections are primarily attributed to differences in Stage 1 selections.

We find support for *Hypothesis 1*. Evaluators exposed to the feasibility-then-novelty sequence were more likely to select solutions with a higher mean innovation than those in the reverse sequence, with differences emerging primarily in Stage 1.

4.2.1. Hypothesis 2. *Hypothesis 2* theorized that evaluators following a novelty-then-feasibility sequence (i.e., first receiving novelty-focused AI recommendations, followed by feasibility-focused AI recommendations) would select solutions with greater innovation variance than those assigned to the feasibility-then-novelty sequence (estimation approach in Section 3.5.4). Figure 2 illustrates the results.

[Insert Figure 2]

Benchmarking against expert evaluations, in Stage 1, we find that evaluators assigned to the novelty-then-feasibility sequence showed greater innovation variance among selected solutions ($Var_{Mean} = 0.692$, $Var_{SD} = 0.274$) than those in the feasibility-then-novelty sequence ($Var_{Mean} = 0.590$, $Var_{SD} = 0.246$) (17% increase in variance) ($Var_{mean\text{-}difference} = 0.101$, $W(19998) = 114.967$, 95% CI = [0.3611; 0.417], $p < 0.001$, $d = 0.389$). The results are similar in Stage 2 though the effect is dampened (12% increase in variance, compared to 17% in Stage 1): the novelty-then-feasibility sequence again led to greater innovation variance ($Var_{Mean} = 0.534$, $Var_{SD} = 0.351$) compared to the reverse sequence ($Var_{Mean} = 0.479$, $Var_{SD} = 0.335$) ($Var_{mean\text{-}difference} = 0.056$, $W(19998) = 29.340$, 95% CI = [0.135; 0.190], $p < 0.001$, $d = 0.163$). Patterns of results were similar when benchmarking against download counts ($p < 0.001$ in both stages, see Appendix Figure A2).

Thus, we find support for *Hypothesis 2*, consistent with the proposed mean-variance innovation tradeoff: although the feasibility-then-novelty sequence produced a steeper increase in mean innovation (*Hypothesis 1*), it also reduced innovation variance among selected solutions. Importantly, both the effects

on mean and variance were stronger in Stage 1 than in Stage 2, suggesting that criteria-sequencing has its greatest impact during initial evaluations. Evaluators following the feasibility-then-novelty sequence appeared more selective when screening solutions, focusing their selections on fewer, often more innovative ones.

Although our theorizing centered on the effects of criteria-sequencing heuristics, the generative AI format represents another important design feature that could shape evaluation outcomes in our experiment. To explore this possibility, we conducted a post hoc analysis to examine whether, how, and why the dynamic versus static format influenced the mean and variance in innovation among selected solutions.

4.3. Post Hoc Analysis: Dynamic versus Static Generative AI Format

[Note: This section has been omitted due to the Strategy Science Conference 35-page page limit]

5. Discussion

This study investigates how criteria-sequencing heuristics—specifically, whether evaluators prioritize feasibility before novelty or vice versa—shape selection in AI-augmented evaluation processes. Drawing on a field experiment with 353 evaluators, each assessing subsets of 20 solutions randomly sampled from a pool of 132 submissions to an innovation contest, we compared two AI-augmented sequences: feasibility-then-novelty and novelty-then-feasibility. This design allowed us to directly examine how the sequencing of AI recommendations emphasizing different criteria, feasibility and novelty, shaped the mean and variance of innovation among the solutions that evaluators select.

Overall, our experimental results provide compelling evidence of a *mean-variance innovation tradeoff*: evaluators in the feasibility-then-novelty sequence selected solutions with a higher mean innovation than those assigned to the novelty-then-feasibility sequence (supporting *Hypothesis 1*), but at the cost of reduced innovation variance (supporting *Hypothesis 2*).

Following the experimental results, we conducted a post hoc analysis to examine whether the generative AI format—static (i.e., fixed explanatory content) versus dynamic (i.e., interactive

chatbot)—influenced selection. Our results reveal that the dynamic format increased the innovation variance among selected solutions compared to the static format, but led evaluators to select solutions with a lower mean innovation. Thus, the mean-variance innovation tradeoff observed between the sequences also emerged when comparing the static and dynamic formats of generative AI. Of course, although these effects appear similar, they operate independently, as one concerns the evaluation sequence, while the other pertains to the formats used within a given stage. Our analysis of chatbot conversations and mouse-tracking data revealed that, in the absence of explanatory content, evaluators relied more on their own judgment, exploring a broader and more diverse set of options but ultimately selecting solutions with a lower mean innovation than those assigned to the static format.

5.1. Theoretical Contributions

Our paper makes three contributions to the literature. First, it advances innovation evaluation research (e.g., Boudreau et al. 2016, Lane et al. 2022, Mount et al. 2021) by demonstrating how criteria-sequencing heuristics fundamentally shape selection. While prior work has examined sequencing in terms of the order in which solutions are evaluated (e.g., Criscuolo et al. 2021, Dahlander et al. 2023, Elhorst and Faems 2021) or developed (e.g., Helfat et al. 2000, Klingebiel et al. 2022, Loch et al. 2001, Thomke and Bell 2001), research on innovation evaluation has largely treated multiple criteria as being evaluated concurrently. Recent studies propose that evaluators often implicitly sequence their assessments of novelty and feasibility (see Baer and Zhang 2024, Lane et al. 2025, Sharapov and Dahlander 2025), and that this sequencing varies idiosyncratically across evaluators (Lamont 2012, Reitzig and Sorenson 2013). As a result, we lack a theoretical understanding of how the sequencing of these assessments shapes which solutions are ultimately selected. Our study provides a systematic framework for a process that has previously been largely unstructured and ad hoc.

The mean-variance tradeoff we uncover has downstream implications for the composition and risk profile of organizations' innovation portfolios. Effective innovation management requires balancing two complementary goals: maximizing the average innovation of selected solutions (i.e., optimizing for the mean) while maintaining atypical deviations and uncertain breakthroughs (i.e., optimizing for the

variance) (Jeppesen and Lakhani 2010, March 1991, Terwiesch and Ulrich 2009). In contexts where maximizing the mean innovation and ensuring rapid, focused technological advancement is essential, such as in R&D-intensive sectors like pharmaceuticals (Tranchero 2023), healthcare (Lebovitz et al. 2022), or aerospace (Chai et al. 2021), prioritizing feasibility before novelty may help evaluators identify reliably stronger solutions—crucial to maintain current competitive advantages. In contrast, in highly uncertain environments where maximizing innovation variance and fostering exploration are critical, such as in emerging markets (Jue-Rajasingh 2025), early-stage ventures (Åstebro and Elhedhli 2006), or complex social innovation projects (Fayard 2024), beginning with novelty assessments may enable organizations to surface a broader range of possibilities. Doing so may increase the likelihood of uncovering highly novel, atypical solutions, but also carries a higher risk of failure (Klingebiel et al. 2022, Chai 2017). These dynamics matter because they shape how organizations navigate the tension between exploration and exploitation (March 1991), balancing the 'essential tension' between reliance on established benchmarks to assess a solution's feasibility with the openness to the novel possibilities it may introduce (Kuhn 1977).

Second, our paper contributes to the emerging literature on human-AI collaboration in innovation contexts (e.g., Berg et al. 2023, Krakowski et al. 2025, Lazar et al. 2025). While some studies have begun to examine how AI assists human evaluators, they have focused mainly on one-shot AI-assisted tasks (e.g., Lebovitz et al. 2021, Doshi et al. 2024, Csaszar et al. 2024), even though evaluations in real-world settings typically unfold across multiple stages. Crucially, most studies have examined how generative AI affects the scale and efficiency of solution generation, rather than evaluation (e.g., Boussioux et al. 2024, Dell'Acqua et al. 2025, Doshi and Hauser 2024). These works highlight that while AI can enhance efficiency, it may also suppress diversity by exerting a homogenizing influence on the content generated (Doshi and Hauser 2024, Hsu and Bechky 2024, Yan et al. 2024). A reduction in content diversity is often seen as negative for breakthrough innovation because such innovation depends on recombining a wide range of knowledge sets (Ferguson and Carnabuci 2017, Kaplan and Vakili 2015).

Our findings offer a new perspective by shifting attention from AI's effects on content diversity in solution generation to its potential to homogenize (i.e., structure) the heuristics that humans rely on in evaluations. Recent work distinguishes between the literature on human-AI collaboration, in which humans use AI to manage organizational tasks, and algorithmic management, in which humans are managed by AI (see Hillebrand et al. 2025). Our study reveals that AI-augmented evaluation embodies both dimensions simultaneously. From a human-AI collaboration perspective, organizations can deliberately use AI to manage evaluation processes: by sequencing feasibility- and novelty-focused recommendations, they systematically shape evaluators' criteria-sequencing heuristics to align selection outcomes with innovation goals. AI recommendations function as "spotlights," directing the order and weighting of criteria that evaluators consider. From an algorithmic management perspective, this same structure constrains how evaluators behave—sequencing different "spotlights" influences their judgments in ways they may not fully control. This dual nature reveals a fundamental design challenge: organizations must balance their desire for consistent, goal-aligned evaluation processes against evaluators' need for subjective judgment and autonomy. Rather than viewing AI's homogenizing influence as inherently detrimental, our findings suggest it creates a new organizational lever—but one that requires careful calibration to preserve the human expertise that makes evaluation valuable (Baer and Zhang 2024, Lamont 2012, Simon 1977).

Third, these findings have broader implications: in the age of ubiquitous AI, organizations must consider not only how, when, and whether to integrate AI into existing evaluative workflows, but also whether AI can be leveraged to fundamentally reorganize these workflows (Amabile 2020, Anthony et al. 2023, Hinds and von Krogh 2024, Mullainathan and Ramachandran 2025). Most existing studies implicitly assume that AI functions primarily as a tool to improve existing processes, shaping the questions researchers ask about delegation (e.g., Balasubramanian et al. 2022), task allocation (e.g., Fügener et al. 2025), trust (e.g., Bauer and Gill 2024), or performance (e.g., Dell'Acqua et al. 2025). Departing from this assumption, we examine whether AI can do more than augment existing processes—specifically, whether it can actively shape evaluators' heuristics to create new, more

consistent, evaluative workflows. Because decision-making is path dependent, initial AI recommendations can influence subsequent decisions, potentially altering the trajectory of innovation. Our results indicate that AI may fundamentally reorganize the very evaluation processes evaluators rely on to identify promising solutions, rather than merely enhancing their efficiency. This perspective points to a new research agenda: instead of treating AI as a tool for improving existing innovation processes, future research could investigate how its evolving capabilities could reorganize the entire innovation process—from ideation to evaluation—reconfiguring the temporal and cognitive structure of innovation itself.

5.2. Practical Implications

[Note: This section has been omitted due to the Strategy Science Conference 35-page page limit]

5.3. Limitations and Future Directions

[Note: This section has been omitted due to the Strategy Science Conference 35-page page limit]

REFERENCES

Abernathy WJ, Clark KB (1985) Innovation: Mapping the winds of creative destruction. *Res. Policy* 14(1):3–22.

Agarwal R, Helfat CE (2009). Strategic renewal of organizations. *Organ. Sci.* 20(2):281–293.

Amabile T (2020) Creativity, artificial intelligence, and a world of surprises. *Acad. Manag. Disc.* 6:351–354.

Anthony C, Bechky BA, Fayard AL (2023) "Collaborating" with AI: Taking a system view to explore the future of work. *Organ. Sci.* 34(5):1672-1694.

Arrighi PA, Le Masson P, Weil B (2015) Addressing constraints creatively: How new design software helps solve the dilemma of originality and feasibility. *Creat. Innov. Manag.* 24(2):247–260.

Åstebro T, Elhedhli S (2006) The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Manag. Sci.* 52(3):395–409.

Augier M, March JG, Marshall AW (2015) Perspective—the flaring of intellectual outliers: An organizational interpretation of the generation of novelty in the RAND corporation. *Organ. Sci.* 26(4):1140-1161.

Baer M (2012) Putting creativity to work: The implementation of creative ideas in organizations. *Acad. Manag. J.* 55(5):1102–1119.

Baer M, Zhang JH (2024) Discerning creativity: a group process perspective on idea selection. *Innovation*, 26(3):387–400.

Balasubramanian N, Ye Y, Xu M (2022) Substituting human decision-making with machine learning: Implications for organizational learning. *Acad. Manag. Rev.* 47(3):448–465.

Bastani H, Bastani O, Sinchaisri WP (2025) Improving human sequential decision making with reinforcement learning. *Manag. Sci.* (forthcoming).

Bauer K, Gill A (2024) Mirror, mirror on the wall: Algorithmic assessments, transparency, and self-fulfilling prophecies. *Inf. Syst. Res.* 35(1):226–248.

Bell JJ, Pescher C, Tellis GJ, Füller J (2024) Can AI help in ideation? A theory-based model for idea screening in crowdsourcing contests. *Marketing Sci.* 43(1):54–72.

Berg JM (2016) Balancing on the creative highwire: Forecasting the success of novel ideas in organizations. *Admin. Sci. Quart.* 61(3):433–468.

Berg JM, Raj M, Seamans R (2023) Capturing value from artificial intelligence. *Acad. Manag. Discov.* 9(4):424-428.

Bergner AS, Hildebrand C, Häubl G (2023) Machine Talk: How Verbal Embodiment in Conversational AI Shapes Consumer–Brand Relationships. *J. Consum. Res.* 50(4):742–764.

Böttcher L, Klingebiel R (2025) Organizational selection of innovation. *Organ. Sci.* 36(1):387-410.

Boudreau KJ, Guinan EC, Lakhani KR, Riedl C (2016) Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Manag. Sci.* 62(10):2765–2783.

Boussioux L, Lane JN, Zhang M, Jacimovic V, Lakhani KR (2024) The crowdless future? Generative AI and creative problem-solving. *Organ. Sci.* 35(5):1589–1607.

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. (2020) Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33:1877–1901.

Chai S (2017) Near misses in the breakthrough discovery process. *Organ. Sci.* 28(3):411–428.

Chai S, Doshi AR, Silvestri L (2021) How catastrophic innovation failure affects organizational and industry legitimacy: The 2014 Virgin Galactic test flight crash. *Organ. Sci.* 33(3):1068-1093.

Chen Z, Chan J (2024) Large language model in creative work: The role of collaboration modality and user expertise. *Manag. Sci.* 70(12):9101–9117.

Chernick MR, LaBudde RA (2014) *An introduction to bootstrap methods with applications to R*. (John Wiley and Sons, New York).

Choi S, Kang H, Kim N, Kim J (2025) How does artificial intelligence improve human decision‐making? Evidence from the AI‐powered Go program. *Strat. Manag. J.* 46(6):1523–1554.

Choudhary V, Marchetti A, Shrestha YR, Puranam P (2025) Human-AI ensembles: When can they work? *J. Manag.* 51(2):536–569.

Choudhury P, Allen RT, Endres MG (2021) Machine learning for pattern discovery in management research. *Strat. Manag. J.* 42(1):30–57.

Cole S, Cole JR, Simon GA (1981) Chance and consensus in peer review. *Science*, 214(4523):881-886.

Cooper RG (1990) Stage-gate systems: A new tool for managing new products. *Bus. Horizons* 33(3):44–54.

Criscuolo P, Dahlander L, Grohsjean T, Salter A (2017) Evaluating novelty: The role of panels in the selection of R&D projects. *Acad. Manag. J.* 60(2):433–460.

Criscuolo P, Dahlander L, Grohsjean T, Salter A (2021) The sequence effect in panel decisions: Evidence from the evaluation of research and development projects. *Organ. Sci.* 32(4):987–1008.

Csaszar FA, Ketkar H, Kim H (2024) Artificial intelligence and strategic decision-making: Evidence from entrepreneurs and investors. *Strategy Sci.* 9(4):322–345.

Daehn IS, Croxson PL (2021) Disability innovation strengthens STEM. *Science* 373(6559): 1097-1099.

Dahlander L, Beretta M, Thomas A, Kazemi S, Fenger MH, Frederiksen L (2023) Weeding out or picking winners in open innovation? Factors driving multi-stage crowd selection on LEGO ideas. *Res. Policy* 52(10):104875.

Davison AC, Hinkley DV (1997) *Bootstrap methods and their application* (No. 1) (Cambridge University Press, UK).

Dell'Acqua F, Ayoubi C, Lifshitz H, Sadun R, Mollick E, Mollick L, Lakhani K (2025) The cybernetic teammate: A field experiment on generative AI reshaping teamwork and expertise (No. w33641). *NBER*.

Doshi AR, Hauser OP (2024) Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Adv.* 10(28):eadn5290.

Doshi AR, Bell JJ, Mirzayev E, Vanneste BS (2025) Generative artificial intelligence and evaluating strategic decisions. *Strat. Manag. J.* 46(3):583–610.

Elhorst P, Faems D (2021) Evaluating proposals in innovation contests: Exploring negative scoring spillovers in the absence of a strict evaluation sequence. *Res. Policy* 50(4):104198.

Ettlie JE, Elsenbach JM (2007) Modified stage‐gate regimes in new product development. *J. Product Innov. Manag.* 24(1):20–33.

Falchetti D, Cattani G, Ferriani S (2022) Start with "why," but only if you have to: The strategic framing of novel ideas across different audiences. *Strat. Manag. J.* 43(1):130–159.

Fayard AL (2024) Making time for social innovation: How to interweave clock time and event time in open social innovation to nurture idea generation and social impact. *Organ. Sci.* 35(3):1131–1156.

Ferguson JP, Carnabuci G (2017) Risky recombinations: Institutional gatekeeping in the innovation process. *Organ. Sci.* 28(1):133–151.

Fox J (2015) *Applied Regression Analysis and Generalized Linear Models* (Sage Publications, Thousand Oaks, CA).

Fügener A, Walzner DD, Gupta A (2025). Roles of Artificial Intelligence in Collaboration with Humans: Automation, Augmentation, and the Future of Work. *Manag. Sci.* (forthcoming).

Gaessler F, Piezunka H (2023) Training with AI: Evidence from chess computers. *Strat. Manag. J.* 44(11):2724–2750.

Gelman A, Hill J (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, UK).

Genin A, Ma W, Bhagwat V, Bernile G (2023) Board experiential diversity and corporate radical innovation. *Strat. Manag. J.* 44(11):2634-2657.

Ghaleb TA, da Costa DA, Zou Y (2022) On the popularity of Internet of Things projects in online communities: An empirical study of Hackster.io. *Inform. Systems Frontiers* 24(5):1601–1634.

Gilovich T, Griffin D, Kahneman D (2002) *Heuristics and biases: The psychology of intuitive judgment* (Cambridge University Press, UK).

Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. (2014) Generative adversarial nets. *Adv. Neural Inform. Process. Systems* 27.

Grewal R, Lilien GL, Mallapragada G (2006) Location, location, location: How network embeddedness affects project success in open source systems. *Manag. Sci.* 52(7):1043–1056.

Harvey S, Kou CY (2013) Collective engagement in creative tasks: The role of evaluation in the creative process in groups. *Adm. Sci. Q.* 58(3):346–386.

Harvey S, Mueller JS (2021) Staying alive: Toward a diverging consensus model of overcoming a bias against novelty in groups. *Organ. Sci.* 32(2):293–314.

Helfat CE, Raubitschek RS (2000) Product sequencing: Co-evolution of knowledge, capabilities, and products. *Strat. Manag. J.* 21(10–11):961–979.

Hill J, Ford WR, Farreras IG (2015) Real conversations with artificial intelligence: A comparison between human‐human online conversations and human‐chatbot conversations. *Comput. Hum. Behav.* 49:245-250.

Hillebrand L, Raisch S, Schad J (2025) Managing with artificial intelligence: An integrative framework. *Acad. Manag. Annals* 19(1):343-375.

Hinds P, von Krogh G (2024) Generative AI, Emerging Technology, and Organizing: Towards a theory of progressive encapsulation. *Organ. Theory* 5(4):26317877241293478.

Hox J, Moerbeek M, Van de Schoot R (2017) *Multilevel Analysis: Techniques and Applications* (Routledge, London, UK).

Hsu G, Bechky BA (2024) Exploring the digital undertow: How generative AI impacts social categorizations in creative work. *Organ. Theory* 5(3):26317877241275118.

Hutchby I, Wooffitt R (2008) *Conversation analysis* (Polity Press, UK).

Johnson W, Proudfoot D (2024) Greater variability in judgements of the value of novel ideas. *Nature Hum. Behav.* 8(3):471–479.

Jue-Rajasingh D (2025) Second-order knowledge intermediaries and multi-country entrepreneurial entry into a nascent industry. *Organ. Sci.* (forthcoming).

Just J, Ströhle T, Füller J, Hutter K (2024) AI-based novelty detection in crowdsourced idea spaces. *Innovation* 26(3):359–386.

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–291.

Katila R, Ahuja G (2002) Something old, something new: A longitudinal study of search behavior and new product introduction. *Acad. Manag. J.* 45(6): 1183-1194.

Kheirandish R, Mousavi S (2018) Herbert Simon, innovation, and heuristics. *Mind & Society* 17(1):97-109.

Knudsen T, Levinthal DA (2007) Two faces of search: Alternative generation and alternative evaluation. *Organ. Sci.* 18(1):39–54.

Krakowski S, Haftor D, Luger J, Pashkevich N, Raisch S (2025) Human-Centered Artificial Intelligence: A Field Experiment. *Manag. Sci.* (forthcoming).

Kuhn TS (1977) *The essential tension: Selected studies in scientific tradition and change* (University of Chicago Press, IL).

Kuznetsova A, Brockhoff PB, Christensen RH (2017) lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* 82:1–26.

Lamont M (2012) Toward a comparative sociology of valuation and evaluation. *Annu. Rev. Sociol*. 38(1):201–221.

Lane JN, Ganguli I, Gaule P, Guinan E, Lakhani KR (2021) Engineering serendipity: When does knowledge sharing lead to knowledge production? *Strateg. Manag. J.* 42(6):1215–1244.

Lane JN, Teplitskiy M, Gray G, Ranu H, Menietti M, Guinan EC, Lakhani KR (2022) Conservatism gets funded? A field experiment on the role of negative information in novel project evaluation. *Manag. Sci.* 68(6):4478–4495.

Lane JN, Szajnfarber Z, Crusan J, Menietti M, Lakhani KR (2025) Beyond feasibility filters: How expertise heterogeneity enables innovation recognition. *Strat. Manag. J.* (forthcoming).

Lazar M, Lifshitz H, Ayoubi C, Emuna H (2025) Would Archimedes shout "Eureka" with algorithms? The hidden hand of algorithmic design in idea generation, the creation of ideation bubbles, and how experts can burst them. *Acad. Manag. J.* (forthcoming).

Lebovitz S, Levina N, Lifshitz-Assaf H (2021) Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Q.* 45(3):1501–1526.

Lebovitz S, Lifshitz-Assaf H, Levina N (2022) To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organ. Sci.* 33(1):126–148.

Levene H (1960) Robust Tests for Equality of Variances. In: *Olkin, I., Ed., Contributions to Probability and Statistics*, 278-292 (Stanford University Press, CA).

Levinthal, D. A. (2025). Navigating more or less: AI and resource allocation on the intensive and extensive margins. *J. Org. Design* 1-4.

Li D, Raymond L, Bergman P (2025) Hiring as exploration. *Rev. Econ. Stud*. rdaf040.

Loch CH, Terwiesch C, Thomke S (2001) Parallel and sequential testing of design alternatives. *Manag. Sci.* 47(5):663–678.

Lou B, Wu L (2021) AI on drugs: Can artificial intelligence accelerate drug development? Evidence from a large-scale examination of bio-pharma firms. *MIS Q.* 45(3):1451–1482.

March JG, Simon HA (1958) *Organizations* (Wiley, NY).

March JG (1991) Exploration and exploitation in organizational learning. *Organ. Sci.* 2(1):71–87.

Mei X, Lee HC, Diao KY, Huang M, Lin B, Liu C … Yang Y (2020) Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nature Med*. 26(8):1224–1228.

Mitchell RJ, Shepherd DA, Sharfman MP (2011) Erratic strategic decisions: When and why managers are inconsistent in strategic decision making. *Strateg. Manag. J.*, 32(7), 683-704.

Mount MP, Baer M, Lupoli MJ (2021) Quantum leaps or baby steps? Expertise distance, construal level, and the propensity to invest in novel technological ideas. *Strateg. Manag. J.* 42(8):1490–1515.
Mueller J, Melwani S, Loewenstein J, Deal JJ (2018) Reframing the decision-makers' dilemma: Toward a social context model of creative idea recognition. *Acad. Manag. J.* 61(1):94–110.
Nelson RR, Winter SG (1982) The Schumpeterian tradeoff revisited. *Am. Econ. Rev.* 72(1):114–132.
Nietzsche FW (1883) *Also sprach Zarathustra: Ein Buch für Alle und Keinen* [*Thus Spoke Zarathustra: A Book for Everyone and No One*], 1. Teil (Ernst Schmeitzner, Chemnitz).
Park CH, von Krogh G, Stadtfeld C, Meboldt M, Shrestha YR (2023) Healthcare hackathons as open innovation. *Nature Rev. Bioeng*. 1(9):610–611.
Phene A, Fladmoe‐Lindquist K, Marsh L (2006) Breakthrough innovations in the US biotechnology industry: the effects of technological space and geographic origin. *Strateg. Manag. J.* 27(4):369–388.
Puranam P (2021) Human–AI collaborative decision-making as an organization design problem. *J. Org. Des.* 10(2):75–80.
Reitzig M, Sorenson O (2013) Biases in the selection stage of bottom‐up strategy formulation. *Strateg. Manag. J.* 34(7):782–799.
Rindova VP, Petkova AP (2007) When is a new thing a good thing? Technological change, product form design, and perceptions of value for product innovations. *Organ. Sci.* 18(2):217–232.
Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088):533–536.
Russell SJ, Norvig P (2021) *Artificial intelligence: A modern approach* (4th ed.) (Pearson, NY).
Schumpeter JA (1939) *Business cycles: A theoretical, historical, and statistical analysis of the capitalist process*, Vol. 1 (McGraw-Hill, NY).
Senoner J, Netland T, Feuerriegel S (2022) Using explainable artificial intelligence to improve process quality: evidence from semiconductor manufacturing. *Manag. Sci.* 68(8):5704–5723.
Simon HA (1947) *Administrative behavior: A study of decision-making processes in administrative organization* (The Macmillan Company, New York).
Simon HA, Newell A (1958) Heuristic problem solving: The next advance in operations research. *Operations Res.* 6(1):1–10.
Simon, HA (1977) *The logic of heuristic decision making. Models of discovery: And other topics in the methods of science* (Springer, Netherlands).
Simon HA (1978) Rationality as process and as product of thought. *Am. Econ. Rev.*, 68(2), 1-16.
Sharapov D, Dahlander L (2025) Selection regimes and selection errors. *Organ. Sci.* (forthcoming).
Shrestha YR, Ben-Menahem SM, von Krogh G (2021) Organizational decision-making structures in the age of AI. *Calif. Manag. Rev.* 63(3):46–68.
Smith GF (1988) Towards a heuristic theory of problem structuring. *Manag. Sci.* 34(12):1489-1506.
Terwiesch C, Ulrich K (2009) *Innovation tournaments: Creating and selecting exceptional opportunities* (Harvard Business Press, MA).
Thomke S, Bell DE (2001) Sequential testing in product development. *Manag. Sci.* 47(2):308–323.
Tukey JW (1977) *Exploratory Data Analysis* (Addison-Wesley, MA).
Tversky A (1972) Elimination by aspects: A theory of choice. *Psychol. Rev.* 79(4):281–299.
Tversky A, Kahneman D (1974) Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science* 185(4157):1124-1131.
Vanneste BS, Puranam P (2024) Artificial intelligence, trust, and perceptions of agency. *Acad. Manag. Rev.* 50(4):726-744.
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, ... Polosukhin I (2017) Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30.
von Hippel E (2006) *Democratizing Innovation* (MIT Press, Cambridge, MA).
von Hippel E, von Krogh G (2016) Crossroads—Identifying viable "need–solution pairs": Problem solving without problem formulation. *Organ. Sci.* 27(1):207–221.
von Krogh G. Spaeth, S., & Lakhani, K. R. (2003) Community, joining, and specialization in open source software innovation: a case study. *Res. Policy* 32(7):1217-1241.

von Krogh G (2018) Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Acad. Manag. Discov.* 4(4):404–409.
Zhong H (2025) Optimal Integration: Human, Machine, and Generative AI. *Manag. Sci.* (forthcoming).
Zhou E, Lee D (2024) Generative artificial intelligence, human creativity, and art. *PNAS Nexus* 3(3):pgae052.

FIGURES AND TABLES

Table 1. Summary statistics for the 353 evaluators who participated in the experiment.

Control variable	Operationalization	Results
Gender	0 = Male 1 = Female	0: 55.241% 1: 44.759%
Age	Continuous	18: 7.932%; 19: 47.592% 20: 32.861%; 21: 10.482% 22: 0.850%; 23: 0.283%
Level of education	0 = Bachelor's 1st/2nd year 1 = Bachelor's 3rd/4th year	0: 52.125% 1: 47.875%
AI expertise (field of education)	0 = Not computer science 1 = Computer science	0: 58.357% 1: 41.643%
Domain expertise (having close family member(s) or friend(s) with disabilities)	0 = Not having 1 = Having	0: 87.819% 1: 12.181%

Table 2. Randomization of evaluators into treatments: T1: *feasibility-then-novelty* versus T2: *novelty-then-feasibility*. The operationalization of the control variables is described in **Table 1**, with Mann-Whitney *U* tests for continuous variables and Chi-squared (χ^2) tests for categorical variables. Additionally, completion time is added as a post hoc control variable.

Control variables	T1: Feasibility-then-novelty ($N_{evaluator}$ = 179)	T2: Novelty-then-feasibility ($N_{evaluator}$ = 174)	Results
Gender	*Mean* = 0.436 (*SD* = 0.497)	*Mean* = 0.460 (*SD* = 0.500)	$\chi^2(1) = 0.120$, $p = 0.729$
Age	*Mean* = 19.503 (*SD* = 0.883)	*Mean* = 19.489 (*SD* = 0.795)	$U(351) = 15455.0$, $p = 0.895$
Level of education	*Mean* = 0.469 (*SD* = 0.500)	*Mean* = 0.489 (*SD* = 0.501)	$\chi^2(1) = 0.065$, $p = 0.799$
AI expertise	*Mean* = 0.397 (*SD* = 0.491)	*Mean* = 0.437 (*SD* = 0.497)	$\chi^2(1) = 0.431$, $p = 0.511$
Domain expertise	*Mean* = 0.123 (*SD* = 0.339)	*Mean* = 0.121 (*SD* = 0.327)	$\chi^2(1) = 0.000$, $p = 1.000$
Post hoc			
Completion time (in minutes)	*Mean* = 19.503 (*SD* = 0.883)	*Mean* = 19.489 (*SD* = 0.795)	$U(351) = 15455.0$, $p = 0.895$

Table 3. Mixed-Effects Models Predicting Selection of Solutions (for *Hypothesis 1*).

	Dependent Variable: Solution Selected (Pass = 1)										
	Overall (20 → 5 solutions)							Stage 1 (20 → 10)		Stage 2 (10 → 5)	
	Sequence	Expert Evaluations			Download Counts			Expert	Download	Expert	Download
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Randomized Sequence *(1 = Feasibility-then-Novelty)*	-0.004 (0.061)	-0.002 (0.061)	-0.055 (0.064)	-0.050 (0.064)	-0.005 (0.061)	-0.040 (0.063)	-0.038 (0.063)	0.022 (0.054)	0.016 (0.054)	-0.062 (0.076)	-0.050 (0.075)
Solution Innovation Rating (*Mean-centered*)		0.902*** (0.091)	0.798*** (0.097)	0.912*** (0.106)	1.072*** (0.126)	0.968*** (0.134)	1.024*** (0.142)	0.812*** (0.089)	0.972*** (0.120)	0.524*** (0.095)	0.559*** (0.124)
Randomized Sequence × Solution Innovation Rating			0.217*** (0.075)	0.213*** (0.075)		0.213** (0.093)	0.212** (0.094)	0.261*** (0.060)	0.255*** (0.078)	0.102 (0.091)	0.086 (0.111)
Randomized Generative AI *(1 = Dynamic chatbot)*				0.048 (0.064)			0.015 (0.063)	-0.036 (0.055)	-0.034 (0.055)	0.081 (0.076)	0.057 (0.075)
Randomized Generative AI × Solution Innovation Rating				-0.212*** (0.075)			-0.109 (0.093)	-0.262*** (0.060)	-0.243*** (0.078)	-0.065 (0.090)	0.022 (0.110)
Constant	-1.429*** (0.116)	-1.343*** (0.089)	-1.319*** (0.089)	-0.871 (0.952)	-1.349*** (0.095)	-1.333*** (0.095)	-0.924 (0.953)	0.457 (0.836)	0.400 (0.838)	-0.501 (1.111)	-0.537 (1.114)
Observations	7060	7060	7060	7060	7060	7060	7060	7060	7060	3530	3530
Controls	N	N	N	Y	N	N	Y	Y	Y	Y	Y
Log likelihood (and df)	-3420.9 (df = 4)	-3383.9 (df = 5)	-3379.8 (df = 6)	-3375.6 (df = 13)	-3391.8 (df = 5)	-3389.2 (df = 6)	-3388.3 (df = 13)	-4194.2 (df = 13)	-4210.2 (df = 13)	-2320.1 (df = 13)	-2326.4 (df = 13)

Notes. This table reports the estimates from mixed-effects logistic regression (generalized linear mixed modeling) predicting the probability of solution selection. All models account for random effects at the evaluator and solution levels. Controls include gender, age, level of education, technical expertise, and domain expertise. The intercept (constant) represents the baseline log-odds of selection when all predictors are zero. To convert it to a baseline probability, use the logistic function: *Probability = exp(Intercept) / (1 + exp(Intercept))*. Coefficients (β) can also be converted to odds ratios using: *Odds Ratio = exp(β)*, which indicates how many times more likely the outcome is for a one-unit increase in the predictor. Standard errors are in parentheses. df, degrees of freedom.

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$;

Figure 1. Selection gains, defined as the increase in average innovation from the initial pool to the selected solutions at each stage, across conditions: *T1* feasibility-then-novelty ($N_{\text{evaluator-solution}}$ = 1790 in Stage 1, 985 in Stage 2) and *T2* novelty-then-feasibility ($N_{\text{evaluator-solution}}$ = 1740 in Stage 1, 870 in Stage 2). Error bars represent ±1 standard error. Results support *Hypothesis 1*, relying on benchmarks: **A)** expert evaluations and **B)** download counts.

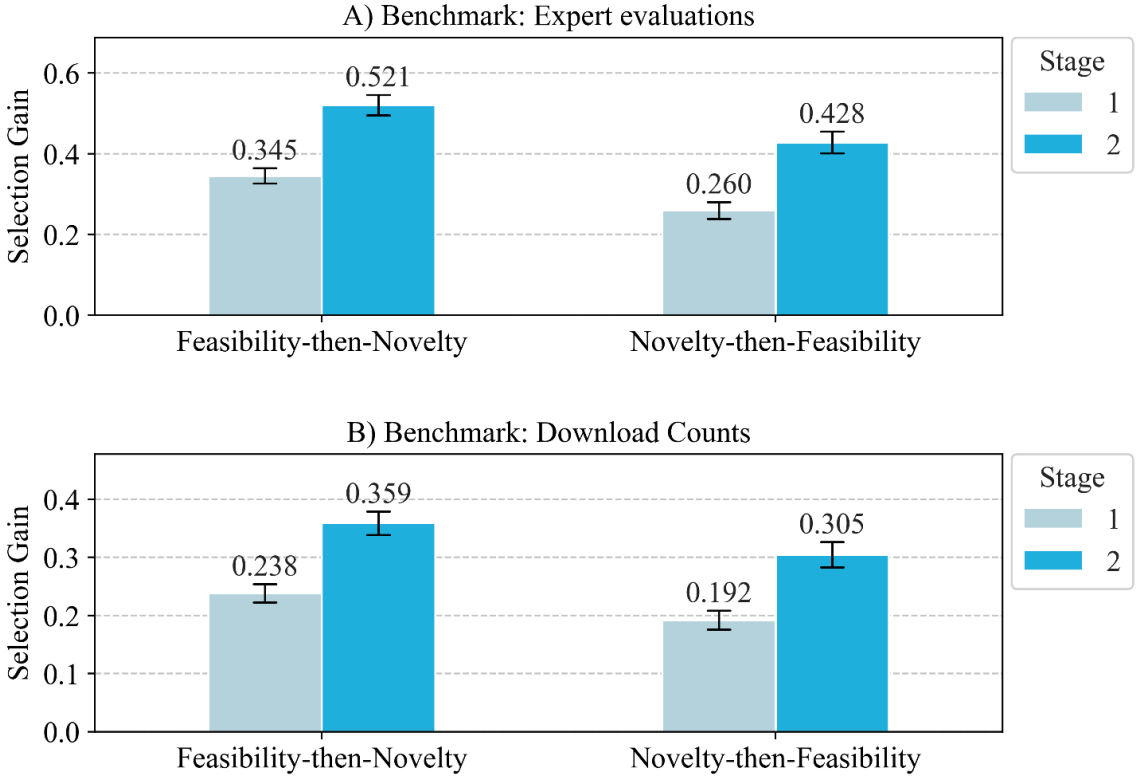

Figure 2. Density distributions and innovation variance in both stages (Stage 1 and Stage 2), across conditions (*T1* feasibility-then-novelty, *T2* novelty-then-feasibility), benchmarking against expert evaluations. Variances and their means (in dotted lines) are obtained from clustered bootstrap samples of passed solutions (i.e., 10 in Stage 1, 5 in Stage 2; see Section 3.5.4). Results support *Hypothesis 2*. See the Appendix, Figure A2, for a similar plot and results, relying on download counts as the benchmark.