Artificial Intelligence in Engineering Design and Systems Engineering: Research Frontiers, Methodological Challenges, and a Community Agenda

Faez Ahmed
Department of Mechanical Engineering,
Massachusetts Institute of Technology,
Cambridge, MA
email: faez@mit.edu

Zoe Szajnfarber
Department of Engineering Management
and Systems Engineering,
George Washington University,
Washington, DC
email: zszajnfa@gwu.edu

Jitesh H. Panchal
School of Mechanical Engineering,
Purdue University,
West Lafayette, IN
email: panchal@purdue.edu

This paper synthesizes the discussions and findings of the NSF Workshop on Artificial Intelligence in Engineering Design and Systems Engineering (AI-EDSE), held at Purdue University in October 2025. The workshop convened eighty participants from academia, industry, and government to identify research priorities at the intersection of AI and engineering. Through framing talks, structured debates, a keynote on human and machine learning, ten early-career lightning talks, and three breakout sessions, the workshop generated several key insights. In the organizers' synthesis, recurring workshop themes included engineering-data scarcity, physics-grounded and uncertainty-aware learning, representation, and education and workforce issues. The paper summarizes these findings into a research agenda with near-term and mid-term priorities, and describes shared infrastructure proposed during the workshop, such as human-centric benchmarks, observation-rich testbeds, curated datasets, and a community grand-challenge specification for design. The workshop's central message is that engineering design and systems engineering should help shape the AI used in engineering and not only adopt tools built for other purposes.

1 Introduction

Artificial intelligence has emerged as a turning point for engineering, prompting a re-evaluation of established practices in engineering design (ED) and systems engineering (SE). Whereas the 2019 NSF workshop on engineering design and systems engineering treated machine learning largely as a promising tool to support design activities [1], the 2025 workshop on Artificial Intelligence in Engineering Design and Systems Engineering (AI-EDSE) recognized AI as a foundational element that is reshaping the roles of designers and engineers and the nature of the design process itself. The workshop's driving objective was to move beyond presenting individual research results toward a collective articulation of the challenges, opportunities, and research priorities that will define the field over the next few years. In doing so, this paper complements the reports from prior NSF-sponsored design workshops that have reviewed the field landscape and helped chart its direction [2,3].

1.1 Motivation: why AI is a turning point for ED and SE. The engineering design and systems engineering communities, served at NSF by the Engineering Design and Systems Engineering (EDSE) program, have periodically convened to define their identity and direction. NSF-supported workshops have periodically taken stock of design research, from a 2008–2009 series on the foundations of design science [3] to gatherings at Clemson (2015), Georgia Tech (2017), Oregon State (2018), and Purdue (2019) [1]. What distinguishes the present moment is the arrival of AI as a broadly capable technology, together with sustained gains in computing. These developments present the EDSE community with a significant opportunity and the responsibility to think critically about how this impacts research and teaching in the field.

Designers have always sought better, faster, and cheaper: to reduce time to market, cut the number of costly iterations, explore vast design spaces, and produce more innovative solutions. The appeal of AI originates from the promises to advance each of these aims while also unlocking new capabilities, such as innovation at scale, the democratization of design expertise, lower barriers to entry for non-experts, and higher productivity for experts. This promise is already visible in the research literature. A bibliometric analysis of five decades of the *Journal of Mechanical Design*, presented in the workshop's engineering design framing talk, claimed a marked and accelerating adoption of machine learning across design research, with a large increase in machine learning terms in titles and abstracts over the past two decades and a sharp rise after the arrival of large language models. By the first half of 2025, a majority of the journal's full-length research papers used machine learning or AI in some form.

1.2 A design-centered perspective on AI in EDSE. A natural question is why the community should pursue AI-EDSE as a research program in its own right. To address this, the workshop was anchored on a design-centered perspective, one that begins from the needs of the designer and the realities of the design process rather than from the capabilities of a particular algorithm. Such a perspective foregrounds concerns that are often peripheral in mainstream AI research but central to engineering: capturing designer intent and constraints, respecting physical laws and lifecycle realities, preserving human judgment and accountability, and integrating AI into established workflows and tools. It also reframes the human as an orchestrator of decisions, whose synthesis, experience, and values remain essential to automation. Adopting this perspective allows the community both to apply AI more effectively to engineering and, reciprocally, to bring engineering rigor to the development of AI itself.

1.3 Paper contributions and organization. This paper synthesizes the discussions and outcomes of the workshop and makes three contributions. First, it distills the workshop's talks, panels, debates, and breakout sessions into a set of key findings on the current state and central challenges of AI-enabled design. Sec-

November 5, 2026

 Copyright © 2026 by ASME

Table 1 Workshop program at a glance. Breakout phases ran as three parallel rooms; lightning talks were delivered by early-career faculty over the two lunches.

Day	Session	Format and speakers
1	Welcome and NSF framing	J. Panchal (workshop objectives); H. Kim (history of NSF design workshops)
1	Framing talks	F. Ahmed (Engineering Design); Z. Szajnfarber (Systems Engineering)
1	Perspectives panel 1: new AI for EDSE	Two-act structured debate: M. Fuge and L. B. Kara
1	Breakout 1: creating AI for EDSE	Representation and physics-informed evaluation · generative methods · designers' requirements
1	Lightning talks (I)	F. Alam, W. Chen, F. Ogoke, L. Wang, S. Sridhara
1	Perspectives panel 2: research literature	Editors-in-chief: C. C. Seepersad (JMD), Y. Wang (JCISE)
1	Breakout 2: using AI for EDSE research	Research methods · supporting humans · design automation
1	Report-out, poster session, reception	Plenary synthesis of breakouts 1–2; evening poster session
2	Keynote: "Learning: Machine and Human"	S. Sarma
2	Perspectives panel 3: applied research	L. Freeman (test and evaluation); A. Raz (INCOSE AI working group)
2	Breakout 3: synthesis of takeaways	Create AI for EDSE · use AI for EDSE · AI and learning
2	Lightning talks (II)	L. Chong, M. Haji, T. Liao, A. K. Ostrowski, T. Topcu
2	Closing	Summary, insights, and path forward

ond, it articulates a research agenda, organized into near- and mid-term priorities together with the community infrastructure and educational changes needed to support them. Third, it advances concrete community actions, including a proposed "AlphaZero" design-specification grand challenge, human-centric benchmarks, and shared data infrastructure. The remainder of the paper is organized as follows. Section 2 describes the workshop's goals, format, and synthesis methodology. Section 3 presents the key findings. Section 4 lays out the research agenda and community recommendations, and Section 5 concludes the paper.

2 Workshop overview and synthesis approach

2.1 Workshop goals, format, and participation. The NSF AI-EDSE Workshop[1] was held on October 16–17, 2025, at Purdue University, West Lafayette, Indiana. Its primary objective was to identify and articulate critical research issues and methodologies at the intersection of artificial intelligence and engineering design and systems engineering. Three specific objectives guided the program:

- to identify emerging opportunities and critical research challenges at the intersection of AI and EDSE;
- to identify new mechanisms for integrating AI-driven EDSE research outcomes across the field's subcommunities; and
- to develop strategies for maximizing the collective impact of the AI-EDSE research community and industry practitioners.

The workshop was attended by eighty registered participants spanning nearly forty institutions and organizations. The roster was predominantly academic, from early-career to senior faculty, journal editors, a few industry-affiliated participants, and the NSF EDSE program director. The organizers invited speakers and panelists from across the AI, ED, and SE communities and publicized the workshop through direct email to several mailing lists. The final speaker list also depended on who accepted and was available, and attendance was not designed as a statistically representative sample of the field.

The program (Table 1) was designed to foster cross-disciplinary dialogue rather than to present finished research. Framing talks and panels served as discussion triggers, and roughly a third of the scheduled time was devoted to structured breakout deliberation.

2.2 Program elements: panels, debates, and breakouts. *Framing presentations and keynote.* The workshop opened with framing talks establishing the current state and specific challenges within both engineering design and systems engineering. On the second day, a keynote address titled "Learning: Machine and Human" by Sanjay Sarma examined human cognition alongside machine learning [4], setting up the workshop's sustained discussion of human-AI collaboration and education.

Panels and debates. Three perspectives panels examined distinct facets of AI integration. The first, on new AI for EDSE, was deliberately staged as a two-act formal debate. The panelists took opposing positions, by coin flip, on (i) whether AI will primarily *automate* existing design processes or *augment* human creativity and judgment, and (ii) whether meaningful progress should be measured through *benchmarks and datasets* or through *real-world case studies*. Each act closed with an explicit synthesis, and the deliberately adversarial format surfaced tensions that recur throughout this report. The format was designed to surface the strongest arguments on each side, not to measure how widely attendees endorsed them. The second panel gathered the perspectives of the editors-in-chief of the *Journal of Mechanical Design (JMD)* and the *Journal of Computing and Information Science in Engineering (JCISE)* on how AI is reshaping the research literature. The third panel brought applied-research perspectives on deploying AI in high-consequence systems, ranging from defense-oriented test and evaluation (framed as AI for T&E (test and evaluation) alongside T&E for AI) to the International Council on Systems Engineering (INCOSE) AI working group.

Poster session. A poster session at the end of the first day showcased ongoing research and emerging concepts, providing a platform for detailed technical exchange and for identifying collaborative opportunities across institutions.

Breakouts. Breakout sessions were the primary mechanism for collecting research priorities from participants. Participants divided into three parallel rooms across three sequential phases: (i) *creating AI for EDSE*, covering design representation and physics-informed evaluation, new deep-learning methods, and designers' requirements for AI tools; (ii) *using AI for EDSE research*, covering AI-enabled research methods, support for human designers, and design automation; and (iii) *synthesis*, charged with identifying the pressing research issues for the next five years. Groups were seeded with driving questions but encouraged to record new research questions throughout; the resulting questions (Table 2) informed the agenda in Sec. 4.

2.3 Synthesis approach, evidence base, and limitations. To ensure comprehensive capture of the proceedings, all formal talks, panels, and the keynote were video-recorded, and the breakout discussions were audio-recorded at the room and table level. All recordings were transcribed, and participants' written research questions were collected and curated. Participants were informed that sessions were being recorded and that comments would not be attributed to individuals; the few quotations reproduced in this paper are unattributed accordingly.

This report is the organizers' interpretation of that record rather than the product of a formal qualitative analysis. In preparing it, the organizers watched the session videos, read the transcripts and rapporteur notes, and used a commercial AI tool (Google NotebookLM) on the transcriptions to summarize sessions and locate recurring themes, which entailed uploading the de-identified transcripts to the tool provider. The selection and framing of all findings are the authors' own. Themes are reported as workshop find-

[1] https://engineering.purdue.edu/ME/AI-EDSE/

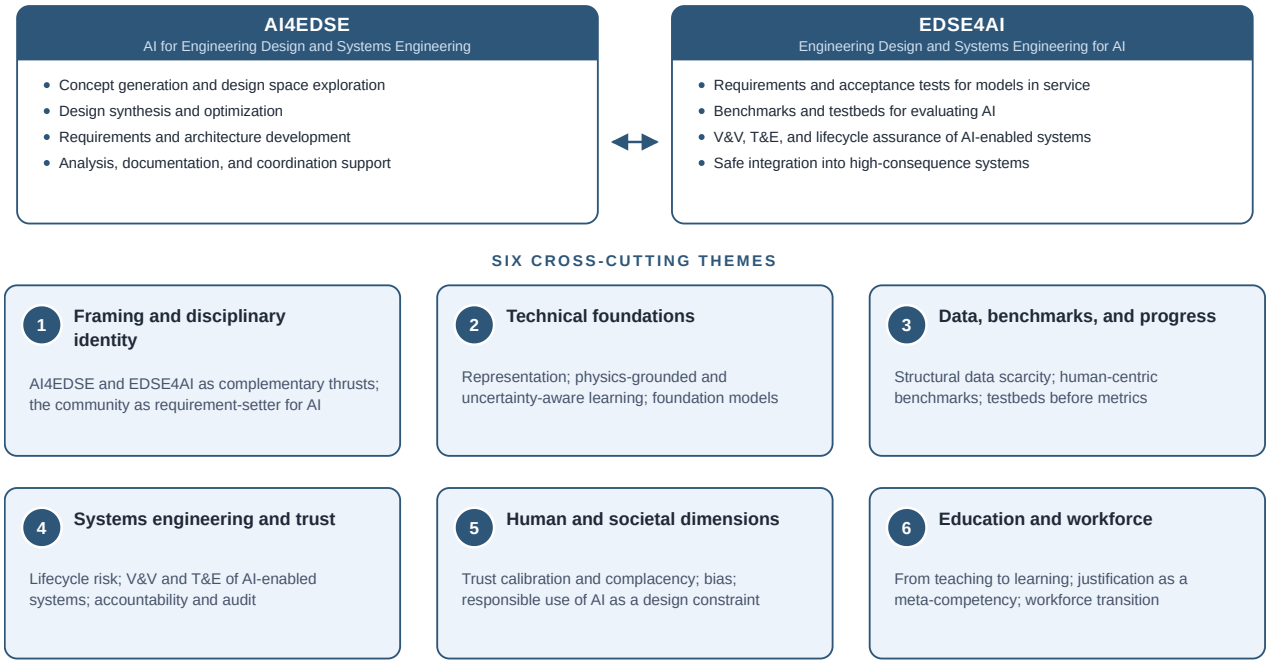


Fig. 1 The workshop's framing and findings: the AI4EDSE and EDSE4AI thrusts as defined in Sec. 3.2 (top), and the six cross-cutting themes elaborated in the subsections that follow (bottom).

ings when the organizers observed them recurring across multiple sessions, but no formal coding procedure was applied, and what is emphasized here, and how it is framed, reflects the organizers' judgment; Appendix A traces each finding and agenda area to the sessions in which it arose.

Three further conventions govern the text. First, where specific results or figures from individual presentations are reported, they are workshop testimony rather than established findings unless a peer-reviewed source is cited; the plenary presentations are public (see Data Availability), so such reports can be checked against the recordings. Second, no vote or poll was taken at the workshop. Statements about recurring views reflect the organizers' reading of the sessions, and where the text notes that a question was contested, that reflects disagreement observed in discussion rather than a measured split. Third, because the organizers set the objectives, delivered the framing talks, and seeded the breakouts with driving questions, the reported themes may partly reflect the workshop's design. However, the clearest evidence that the outcomes of the workshop are not wholly an artifact of workshop design is education, which recurred in every breakout phase despite being an assigned prompt only in the final phase.

As with any community-driven synthesis, the perspectives presented here reflect the composition and scope of the workshop. The emphasis on engineering design and systems engineering necessarily limits coverage. Participants were primarily based in the United States, which may shape priorities related to funding structures and educational systems, and industry was underrepresented relative to academia, a gap that participants themselves flagged, alongside the near-absence of quantum and alternative computing from the program. The synthesis is therefore best read as a snapshot of the community's thinking at a pivotal moment.

3 Key findings from the NSF AI-EDSE workshop

3.1 The state of the field: a cross-section of recent advances. The breadth of recent activity is striking. Beyond the adoption statistics noted in Sec. 1, the editors' panel mapped where AI now appears across the *Journal of Mechanical Design*: concept generation with deep generative models, graph neural networks for interdependent engineering systems, human-AI "co-pilot" collaboration, AI-enabled design for additive manufacturing, automated design space exploration (a lineage running through decades of surrogate modeling and metamodeling), and early-stage design support built on large language models mining customer needs. Multimodal learning that fuses text, geometry, and performance data is increasingly prevalent. The panel also described two recurring failure patterns in submissions: the "genetic algorithm effect," in which a popular method is applied where it is not appropriate, and a "Lego approach" that stacks methods for apparent novelty. These can be seen as symptoms of a field racing to absorb new tools.

The workshop's ten lightning talks, delivered primarily by early-career faculty invited by the organizers, offered an illustrative cross-section of current research, and they surfaced the recurring themes that structure this report: the primacy of representation, the constraint of data scarcity, the danger of out-of-distribution failure, and the need to keep the human in the loop. We mapped these findings onto six cross-cutting themes, as shown in Figure 1.

Several talks addressed geometric representation and generative design. Work on foundation models for generative computational design argued that the central obstacle is not model capacity but representation: three-dimensional mechanical shapes lack the mature representations that text and images enjoy, and meshes discard the parametric history that could capture engineering intent. The proposed path recasts geometry as code, a sequence of CAD operations or an executable script, so that 3D generation becomes a language-modeling problem that preserves design history and editability; systems such as GenCAD reconstruct full CAD command sequences from images or sketches [5]. A complementary talk introduced continuous representations of discrete engineering catalogs, reporting that a typical car comprises roughly 30,000 parts, of which some 25,000 are selected from catalogs and only about a fifth are continuous geometries; systems-level design therefore requires embedding vast discrete catalog data (fasteners, bearings, materials) into continuous latent spaces so that assemblies can be optimized continuously and then snapped back to real, manufacturable components [6].

Other talks emphasized physics grounding and the limits of data-driven models. A talk on inverse design of nacre-inspired

composites showed that conditional generative models silently fail on out-of-distribution targets (no generated design satisfied targets outside the training envelope) and that simply diversifying the dataset does not help, because designers control the distribution of designs but not the resulting distribution of responses [7]. The proposed remedy trains an uncertainty-aware predictive model that assigns each candidate a likelihood of satisfying the target and, critically, declines to generate anything for infeasible targets, on the principle that models should recognize the limits of their own competence. A talk on additive manufacturing tackled the multi-fidelity gap in laser powder-bed fusion: because mapping cheap, coarse simulations to expensive, fine-resolution ones is ill-posed, a deterministic model predicts the large-scale melt-pool structure while a generative model adds realistic fine-scale detail, yielding sharp field estimates with calibrated confidence bounds that widen where the physics is unstable. Work on AI-enabled design of material systems demonstrated co-design of material, architecture, and stimuli (mechanical cloaking, 4D-printed liquid-crystal-elastomer networks, and untethered magnetic soft robots), often via differentiable simulation or a hybrid of AI modules within conventional tools, with the pointed observation that the value of a method lies in solving the problem, whether or not it is nominally or entirely "AI." A talk on multidisciplinary design optimization for offshore renewable systems reached the same conclusion from the systems side: the enabling investment was rewriting hydrodynamic solvers to be differentiable, after which learning methods become far more useful.

A third group of talks examined systems engineering and digital engineering. A study of digital-engineering transformation showed, through discrete-event simulation of a naval design organization, that moving to an authoritative source of truth can substantially reduce design cycle time before any AI is added, while in a companion experiment the text classifiers evaluated did not reliably distinguish large-language-model-generated systems-engineering artifacts from human-written ones, even though expert review under the study's protocol identified substantive deficiencies: a failure mode that the automated checks tested there missed, with direct implications for V&V.

Finally, talks on human-centered design examined how designers actually use these tools. Studies of trust and reliance found that confidence in an AI recommender is easily lost and hardly regained and, more troublingly, that although designers' self-confidence was initially insensitive to the AI's accuracy, an unannounced change in that accuracy, in either direction, reduced their confidence in their own abilities [8], while users leaning on generative tools often overlooked the best-balanced designs. A modular AI design-assistant testbed suggested that emotional recognition and rapport may partly offset weaker agent performance. Participatory studies of workplace robots and AI agents showed that employees co-designing agent roles converge on careful, calm, practical agent "personalities," that embodiment shapes preferred autonomy, and that long-term users can become dangerously complacent, motivating deliberate "designing against complacency." Together, these talks by junior faculty foreshadow the themes developed below.

3.2 Framing: AI4EDSE and EDSE4AI. The workshop's central framing drew, deliberately, on the systems engineering community's experience organizing research at this intersection. Two complementary research thrusts were codified in that community's research roadmap [9] and have been developed through the annual AI4SE & SE4AI research workshop convened by the Systems Engineering Research Center (SERC) and its partners since 2020, a series that grew from 20 people in a video call to a capped registration of 250 in 2025, with a mix of academia, industry, and government throughout [10]. *AI for systems engineering* (AI4SE) applies AI and machine learning to improve the practice of systems engineering itself: requirements elicitation, classification, and analysis; assistance for model-based systems engineering; knowledge integration across the digital thread; and tradespace exploration and documentation support. *Systems engineering for AI* (SE4AI) applies systems engineering methods, processes, and tools to the engineering of AI-enabled systems: specification and architecture of learning-enabled components, verification and validation, test and evaluation, safety assurance, and governance across the lifecycle (Fig. 2). The two are complementary rather than sequential, and the workshop treated them as parallel thrusts that serve different needs. The same dual pattern is spreading to neighboring specialties; the applied-research panel framed test and evaluation the same way, as AI for T&E alongside T&E for AI. As the field has matured, these thrusts have proven not merely complementary but mutually reinforcing: systems-engineering principles are increasingly used to assure and re-scope AI4SE tools, and AI is increasingly used to make SE4AI itself more efficient. The systems engineering community now treats these feedback loops between the thrusts as at least as important as either thrust in isolation [9], and frames the relationship less as a dichotomy than as a virtuous cycle.

The workshop extended this duality to engineering design and to the EDSE community as a whole, as *AI4EDSE* and *EDSE4AI*. The terms were used at the workshop without formal definition; we define them here. AI4EDSE, AI for engineering design and systems engineering, is the application of AI and machine learning to improve the practice of engineering design and systems engineering across the lifecycle: concept generation and design space exploration, design synthesis and optimization, requirements and architecture development, and the analysis, documentation, and coordination work that surrounds them. EDSE4AI, engineering design and systems engineering for AI, is the application of design and systems engineering methods to the creation, evaluation, and integration of AI itself: setting requirements and acceptance tests for models used in engineering service, designing the benchmarks and testbeds against which they are judged, and bringing verification, validation, and lifecycle assurance to AI-enabled systems (Fig. 1, top). On the AI4EDSE side, participants argued that the field is maturing past the stage where the routine application of existing machine learning methods to engineering problems constitutes significant intellectual merit; the frontier lies in methods shaped by the structure of design problems themselves. On the EDSE4AI side, the field is still in its nascency. The community's opportunity and responsibility is to act as requirement-setter and benchmark-definer for the AI used in engineering: defining the acceptance tests any model must pass to be trusted in engineering service, whoever builds it, and contributing the engineering rigor that safe integration of AI into high-consequence systems demands. The latter aligns with a 2025 National Academies study of machine learning for safety-critical applications, which calls for a new engineering discipline drawing from machine learning and systems engineering [11].

At the same time, participants sounded a note of caution: AI should be applied where it genuinely helps, and much valuable work in the community neither requires nor benefits from machine learning. Several sessions tempered expectations with evidence from practice: surrogate models whose errors have plateaued well above the thresholds practitioners need, generative CAD limited to simple shapes, and language models whose apparent competence collapses under adversarial or circular evaluation. The task ahead is to close such gaps honestly rather than to celebrate demonstrations.

3.3 Technical bottlenecks: data scarcity, hallucination, and physics-grounding. Across sessions, participants repeatedly raised three bottlenecks that currently limit the reliability and scientific pace of AI-enabled EDSE: scarce and fragmented data, plausible-but-wrong generation, and insufficient grounding in physics and real-world constraints.

3.3.1 Data scarcity as a structural condition. Unlike text- or image-centric AI, EDSE learning is constrained by the economics and governance of engineering data. Where language models train on trillions of tokens, engineering datasets rarely exceed a hundred

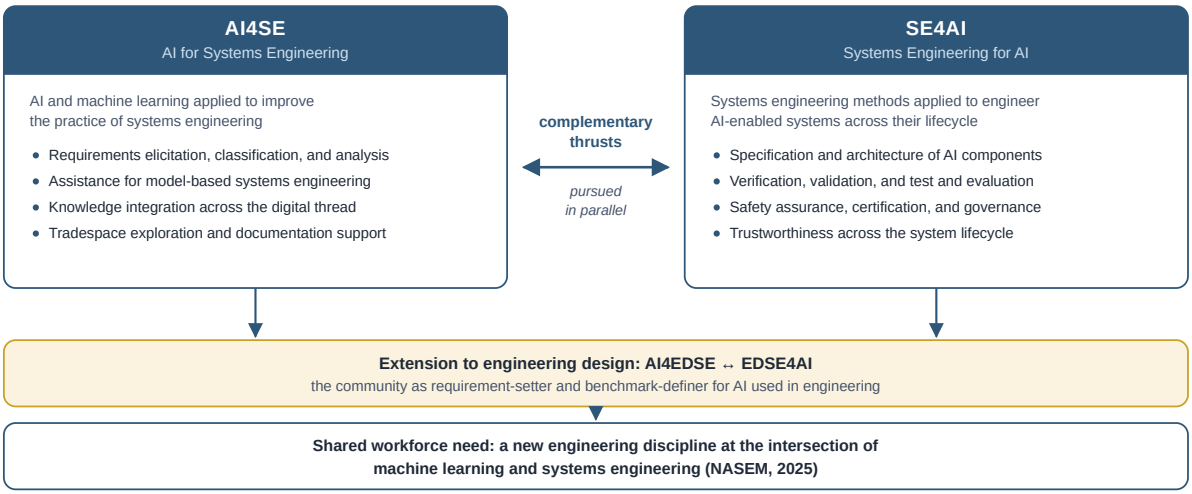


Fig. 2 The complementary AI4SE and SE4AI research thrusts as codified by the systems engineering community [9,10], and the workshop's extension of the duality to engineering design as AI4EDSE and EDSE4AI. The thrusts proceed in parallel and converge on the need for a new engineering discipline at the intersection of machine learning and systems engineering [11].

thousand diverse items, and producing those items is expensive: by the framing talk's account, a single high-fidelity aerodynamic dataset of eight thousand car geometries consumed roughly three million CPU-hours and thirty-nine terabytes of storage for one boundary condition [12]. The framing talk on engineering design argued that this scarcity is also a market failure: firms hold data they have a disincentive to share, the data they hold encode superseded designs stripped of tacit knowledge, and intellectual-property and export-control boundaries block aggregation. Participants noted that even when data exist, they are rarely organized for reuse: boundary conditions, meshing decisions, solver settings, and modeling assumptions are often missing, which makes results hard to reproduce or compare.

The workshop added two refinements to the familiar scarcity narrative. First, *granularity*: most datasets capture whole-task outcomes (a final geometry, a performance label) rather than the sub-task decisions from which engineers actually work, entangling what practitioners like and dislike about a design; process-rich data (requirements evolution, rejected alternatives, design rationale, and the reasons concepts failed) are the scarcest and most valuable of all, and industry rarely records them. Second, *information content*: not all data are equally informative. A heavy-truck dataset dominated by steady highway driving carries little new information however large it grows, while rare events such as maintenance and diagnostics carry the signal.

These observations motivate research on resource-optimal data generation that maximizes learning value per expensive simulation or experiment, and on data augmentation defined at the level of governing equations and also capturing engineering realities.

3.3.2 Hallucination and silent failure in safety-relevant contexts. Generative models can produce outputs that look coherent and professional while being incorrect, infeasible, or unsafe. In EDSE, this failure mode is particularly dangerous because it can be silent: an AI-generated requirement, design rationale, or analysis summary may pass an initial human scan yet violate physical laws, omit constraints, or misrepresent uncertainty. The workshop record is rich in concrete examples, from generated systems-engineering artifacts that fooled text classifiers while being substantively poor, to statistical agencies discovering that fluent model-written reports contained fabricated figures, to inverse-design models confidently proposing structures for targets they had no basis to reach. Participants emphasized that hallucination in engineering is not merely a language problem: it is a systems risk that propagates downstream through the lifecycle.

Participants highlighted three families of countermeasures: *calibration and uncertainty communication*, so users can distinguish confident claims from speculative ones; *retrieval and provenance*, linking generated statements to traceable sources such as requirements baselines, validated models, and test reports; and *out-of-distribution stress testing*, to reveal brittleness before deployment in high-consequence workflows. A recurring proposal was that journals and reviewers should expect papers to report the failure envelope, meaning where physics is violated, which constraints are fragile, and how errors grow under distribution shift, rather than in-distribution accuracy alone.

3.3.3 Physics-grounding as a pathway to trust and robustness. To address both data scarcity and hallucination, the workshop repeatedly returned to physics-informed and constraint-aware learning. Participants argued that grounding is not only about accuracy; it is a foundation for assurance. Approaches discussed include embedding conservation laws and invariances, enforcing hard constraints during inference rather than as soft penalties, hybridizing learned models with simulators and solvers, and operator-learning methods that generalize across input functions and boundary conditions [13–15].

The breakout on representation and physics-informed evaluation sharpened this into a methodological thesis: *designs are not images*. Image-space similarity does not encode the physical significance of geometric change: a perturbation that is negligible by such measures can transform structural performance, and two structures far apart in image space can be identical in property space. Tools imported from image-centric ML therefore carry the wrong inductive biases, and constraint satisfaction, which mainstream generative modeling treats as a soft preference, is in engineering the difference between a design and a rendering. The same session urged intellectual honesty about baselines and argued that data-driven and physics-based per-instance methods are complementary rather than competing: a generative model can act as a preconditioner that identifies the low-dimensional subspace in which differentiable, simulator-in-the-loop refinement then operates. Related work on minimal representations suggests why: familiar parameterizations often overstate a problem's intrinsic dimensionality (an airfoil family that engineers parameterize with hundreds of variables may occupy a low-dimensional manifold),

and compressing toward that intrinsic dimension measurably reduces hallucination.

3.4 Representations and foundation models for engineering design. Many workshop discussions framed representation as the bottleneck between modern foundation models and engineering workflows. EDSE artifacts span multiple modalities and tools: natural-language requirements, functional decompositions, architecture graphs, CAD feature histories, meshes and boundary conditions, bills of materials, manufacturing plans, test data, and fielded operational logs. Today these artifacts are fragmented, inconsistently structured, and weakly linked across the lifecycle. Participants observed that "representation" means different things to different audiences: representations suited to AI models, to engineering software (finite-element, CFD, and dynamics solvers), and to human designers impose distinct requirements, and the central challenge is integration among them so information flows without loss.

3.4.1 From geometry to intent. Engineering representations must go beyond geometry. They must encode *designer intent* (constraints, interfaces, allowable variation), *context* (why trade-offs were made), and *lifecycle realism* (manufacturability, inspection, maintenance, sustainability). Representations that omit intent may support shape generation, but they do not support design accountability or modification under evolving requirements.

3.4.2 Toward universal, auditable representations. A recurring aspiration was a "programming language for the 3D world": a shared, auditable substrate structured enough for constraint reasoning and simulation coupling, yet flexible enough to represent assemblies, multi-material parts, and process-specific detail, together with the "compiler" that translates it into machine-executable form, an ambition with roots in the design-compiler research of the late 1980s and in chemistry's molecular string representations that made large-scale molecular machine learning possible. Participants acknowledged a harder fallback position: if no universal representation exists, the community should organize existing representations hierarchically and build translations between branches. In parallel, systems engineering demands representations connecting requirements, architectures, interfaces, and verification evidence; participants advocated tighter integration between model-based systems engineering artifacts and design-generation tools [16], noting the irony that some organizations are drifting back from model-based to document-based artifacts precisely because language models read text more readily than models, at the risk of losing the human's full-system view.

3.4.3 Foundation models for EDSE: opportunities and prerequisites. A foundation model is trained on broad data and adapted to many downstream tasks [17]; foundation models for engineering offer a pathway to reusable design intelligence across retrieval, synthesis, generation, critique, and repair. The workshop's discussion was notably measured about what this requires. Participants stressed that EDSE foundation models will be trustworthy only if paired with cross-modal alignment among requirements, architecture, geometry, and analysis; provenance and versioning consistent with digital engineering; constraint-aware decoding to prevent "pretty but impossible" solutions; and evaluation protocols that reward robustness and lifecycle feasibility rather than novelty alone.

The question of scope was contested: how many distinct physics can one model plausibly encompass, and is the right architecture a single monolith or composable single-physics "building blocks," each validated in its own domain and combined through mixture-of-experts mechanisms? Participants speculated that composing privately trained expert models might also let firms contribute capability without exposing proprietary data, while noting that this would require dedicated federated-training and confidentiality mechanisms beyond the architecture itself. Several participants argued for getting single-physics foundation models right first; others noted that for many practitioners a well-characterized domain-specific model is all the foundation they need.

Participants also stressed prerequisites beyond the models. Realizing these capabilities requires rethinking legacy design tools (CAD, CAM) for AI-centric workflows and investing in AI-native infrastructure: automatic differentiation, GPU-native design software, and open, reproducible platforms. Like the frameworks underlying modern machine learning, these are not algorithms but preconditions for algorithm research. A recurring open question was whether such models can move beyond interpolating among prior designs toward extrapolation and novelty, and how that novelty would be evaluated; Section 3.9 returns to the associated fixation risk.

Much of the workshop's discussion on foundation models focused on geometry and physical behavior (e.g., CAD, FEA, CFD). Participants were careful to situate the cross-disciplinary ambition in its lineage: linking requirements, architecture, geometry, and analysis into a coherent whole is the problem that model-based systems engineering (MBSE) has pursued for decades and that digital engineering is now addressing in part [16]. It was argued that foundation-model approaches should build on that tradition.

3.5 Systems engineering: requirements, uncertainty, and lifecycle risk. AI should be treated as a component of a complex system, which means that systems-engineering perspectives must be brought in at the developmental stage to establish AI trustworthiness with rigor. This is consequential because most AI models today operate at the level of single components, whereas the hardest EDSE problems are *lifecycle* problems: requirements evolve, environments shift, and systems must remain safe and maintainable under uncertainty. This framing places specification, traceability, verification and validation, and governance at the center of AI-enabled design.

The systems engineering framing talk offered a three-level structure that recurred throughout the workshop: opportunities exist at the *task* level, the *workflow* level, and the *system* level, and the levels do not reduce to one another [10]. A fielded AI tool can perform its task well yet deliver no value because the surrounding workflow never acts on its outputs; the fix in one government case study was not a better model but a redesigned communication loop between the humans the tool was meant to serve. Conversely, replacing a human role with AI at the system level displaces ancillary functions that the role silently absorbed (the driver who also loads luggage and reassures passengers), so the entire task architecture must be re-partitioned. And when AI performs system-level functions no human ever performed, such as market-scale matching, the interpersonal mechanisms by which people calibrated trust in other people no longer apply. Each level demands its own research: new architecture patterns that exploit AI capabilities rather than inheriting assumptions from heritage systems, new test-and-evaluation strategies for behaviors that emerge from many-human, many-AI interaction, and deliberate design of workforce transitions as part of the system.

The applied-research panel emphasized that benchmark accuracy and safety-critical reliability are not commensurate quantities: benchmark accuracy aggregates prediction outcomes over a sampled dataset, whereas system-level risk must be defined per demand, hour, or mission within a specified operational domain and hazard model. Translating component-level predictive performance into system-level risk is itself an open engineering problem, and participants argued that more data alone does not solve it without physics, structure, and rigorous test design. Test and evaluation itself is being reshaped from the classical V-model toward tightly coupled, continuous design-and-test practices, and AI is entering on both sides: as the object of new T&E strategies (operating-envelope characterization, drift monitoring, designed experiments that isolate which environmental factors actually drive failures) and

as a tool for T&E (test-case generation, simulation environments, and automated experiment design).

Two further threads recurred. First, requirements were described as living artifacts that accumulate *technical debt*, ambiguous or shortcut requirements whose costs surface downstream, for which AI-assisted cross-checking, quality analysis, and detection of that debt are promising because requirements language is semi-structured. Participants from the systems-engineering community cautioned, however, that requirements are highly contextual. Tools that aim to generate requirements often miss precisely the context that makes them correct, so the more robust near-term opportunity is to validate, cross-check, and connect requirements to design models rather than to author them using AI. Second, participants pointed to digital engineering and the maintenance of an *authoritative source of truth* across the lifecycle [16,18], including proposals to integrate the human into the digital twin, which one workshop session termed the "digital triplet," together with verification, validation, and uncertainty-quantification methods suited to non-deterministic, adaptive components.

3.6 Emerging modes of practice: agentic workflows, digital engineering readiness, and virtual environments. Looking forward, participants repeatedly used software engineering as a leading indicator of how design practice may reorganize around AI. In software, assistants have progressed within a few years from code completion to agents that plan, execute, and check multi-step tasks with intermittent human direction; several participants argued that engineering design will follow, with agents operating CAD, meshing, simulation, and optimization tools under a designer's supervision. The workshop record is specific about what this transition requires. It requires machine-operable interfaces to engineering tools (participants pointed to the standardized tool-interface protocols emerging in the software ecosystem as a template for CAD and CAE interoperability), representations that preserve design intent (Sec. 3.4), and the deliberate task- and workflow-level re-architecture discussed in Sec. 3.5: an agent that automates one task reshapes the workflow around it, and supervising agents (delegating, checking, and knowing when to intervene) is itself a skill that must be taught (Sec. 3.10). It also carries specific risks: the jagged frontier makes an agent's competence boundary hard to anticipate, and the longitudinal workplace studies presented at the workshop suggest that sustained exposure breeds exactly the complacency that agentic autonomy makes dangerous (Sec. 3.9).

Two further forward-looking threads emerged. First, AI's value depends on *digital engineering readiness*. The discrete-event study of a design organization discussed in Sec. 3.1 found that establishing an authoritative source of truth substantially reduces design cycle time before any AI is introduced; participants argued, conversely, that organizations without a coherent digital thread will not realize AI's benefits, because the artifacts an AI system needs (requirements, models, analyses, decisions) are scattered, inconsistent, or trapped in documents [16,18]. Investment in digital engineering is thus a precondition for, not an alternative to, AI adoption. Second, participants pointed to rich virtual environments, up to and including world-model simulators of operational contexts, as the substrate on which AI-enabled engineered systems can be trained, stress-tested, and evaluated before deployment, connecting the testbed agenda of Sec. 3.8 to the test-and-evaluation needs of Sec. 3.5. On a longer horizon, participants flagged *self-organizing and self-redefining systems*, engineered systems that restructure their own architecture over their lifecycle, as a coming object of study for which today's design theory offers little guidance.

3.7 AI in design research methodology. Beyond its role as a design tool, AI is reshaping the *methods* by which design research itself is conducted, and breakout discussions repeatedly returned to both the opportunities this creates and the norms it demands. Because large language models (LLMs) are now broadly accessible, the community can ask questions that were previously impractical, but only if it establishes shared standards for how AI is used and disclosed in the research process.

A first opportunity is *scaling qualitative inquiry*. Participants described pipelines in which a human researcher abductively codes a subsample of open-ended responses, trains language models to replicate that coding, and applies the models to tens of thousands of responses with strong inter-coder reliability, an approach rigorous enough to satisfy review at top management-science venues. This makes it feasible to analyze design conversations, interviews, and think-aloud protocols at a scale manual coding cannot reach. A second opportunity is *synthetic personas and simulated users*: encoding design theories or user segments into LLM agents to probe design cognition, pilot interview protocols, and rehearse user research before engaging real participants. A recurring caution is that such agents inherit the biases of their training data; one pilot that simulated clinic nurses in a Global South setting surfaced colonial framings and fabricated local practices, which participants reframed as a valuable cautionary example for teaching students to interrogate rather than trust model outputs. A third, more mundane, opportunity is *acceleration of research instrumentation* itself: study interfaces that once consumed a dissertation year can now be prototyped in weeks.

These opportunities are inseparable from questions of *research norms and integrity*. Participants debated where the boundaries lie for using AI to brainstorm hypotheses, design a study, or draft a manuscript, and how such use should be reported, distinguishing, by analogy to authorship conventions, between assistance that warrants acknowledgment and assistance that would amount to co-authorship. Concerns included the erosion of peer review by AI-inflated submissions, fabricated or mismatched references propagating through the literature, the difficulty of reproducing results built on continuously changing commercial models, and the risk that heavy reliance on tools trained on the average of prior work pushes research toward the mean. Participants also noted a genuine tension over whether writing can be cleanly separated from intellectual contribution, and how disclosure norms should accommodate non-native English speakers who benefit most from language assistance. The community identified a need for explicit, discipline-specific guidelines (automated reference-integrity checking, structured disclosure of which research stages used AI, and artifact-evaluation incentives adapted from computer science) and for publication venues, such as short methods notes, in which methodological contributions of this kind can be documented and vetted.

3.8 Measuring progress: benchmarks, testbeds, and case studies. How should the field define and measure meaningful progress? The workshop's second staged debate confronted benchmarks with case studies, and the tension proved productive precisely because both sides are right. Benchmarks enable standardized, reproducible comparison, lower the barrier for new entrants, and connect EDSE to the broader machine-learning community [19]; as one panelist noted, whoever builds a benchmark first effectively sets the field's direction: "for better or worse, benchmarks shape a field" [20]. Yet any single benchmark captures a narrow slice of a skill, invites overfitting and leaderboard-chasing, and can enshrine toy problems whose edge cases are exactly what matters in practice: the debate's case studies ranged from image-recognition advances that failed to transfer to driving safety, to catastrophic system failures that no component benchmark would have caught. Systems engineers added a disciplinary caution: in much of SE the goal is *discovery* of requirements and operating envelopes, not optimization of a fixed metric, so benchmark thinking must be complemented by characterizing where a method works and where it fails.

Three constructive positions emerged. First, the community lacks, and should build, *human-centric* benchmarks for generative design and 3D-shape understanding, analogous to human-evaluation benchmarks in computer vision [21]; participants could point to no widely used standard that measures whether generated

designs are meaningful to the humans who must accept, modify, and certify them. Second, several participants argued that observation-rich *testbeds and sandboxes* should precede formal metrics: rich environments for watching human-AI design workflows reveal what is worth measuring, whereas premature metrics risk institutionalizing the measurable at the expense of the meaningful; "measuring the wrong thing can be worse than not measuring at all." Third, measurement requires shared vocabulary: the community must first agree on what terms such as "benchmark," "efficacy," and even "design" and "creativity" mean before meaningful comparison is possible. Alongside these, participants called for stronger reproducibility and reporting norms, which Sec. 4.3 carries into the agenda.

3.9 Human and socio-technical dimensions: trust, accountability, and ethics. Throughout the workshop the discussion returned to the balance between augmentation (helping humans) and automation (replacing them), the subject of the first staged debate. The automation position marshaled self-play systems that exceeded human competence without human data [22], the general trend of compute-leveraging methods outperforming hand-crafted knowledge [23], and the rapid agentic transformation of software engineering. The augmentation position answered with two durable observations: humans cannot fully prescribe what they want, because requirements are tacit, dynamic, and revised in response to what AI produces; and responsibility is not delegable: in safety-critical engineering, accountability must ultimately rest with a person. The synthesis was not a compromise but a reframing consistent with the keynote's warning against the "Turing trap" of building AI that imitates rather than complements people [24]: decide deliberately, task by task, what to automate and what to keep human, and design the interfaces between them.

Trust emerged as a measurable, dynamic quantity rather than a slogan [25]. Human-subject studies presented at the workshop show that confidence in AI advice erodes quickly after observed failures and recovers slowly, and that an unannounced change in an AI's accuracy, in either direction, can depress designers' confidence in themselves [8,26]. The failure modes are two-sided: under-reliance wastes capable tools, while over-reliance breeds complacency; seasoned operators of autonomous systems were observed extending interpersonal trust to machines, leading participants to propose designing *against* complacency, including deliberately imperfect behavior that keeps humans critically engaged, and psychology-informed tools that adapt to a user's confidence and cognitive load.

As AI becomes ubiquitous, the community faces questions of accountability, bias, and responsible use. A recurring point was the absence of audit mechanisms for AI-inflected design decisions: because an AI system cannot be held liable, engineering practice needs to know where and when decisions were made, by whom or what, and who is answerable; these are questions of certification and liability that existing regimes for licensed engineers do not yet cover. Participants translated this gap into an actionable research direction: *decision-provenance infrastructure* that records which design decisions were made, on what evidence, and with what degree of AI involvement, so that AI-inflected design processes remain auditable after the fact; and certification pathways that extend established safety-case practice to learning-enabled components rather than exempting them. Both are areas where the EDSE community can lead, because auditability is ultimately a requirement to be engineered into the design process itself. Participants documented how AI can amplify existing bias: a generative tool trained on bicycle-seat data reflecting predominantly male riders reproduces that bias for all users, and medical and mobility datasets over-represent particular populations and geographies. Beyond these risks, participants sought to define and amplify what should be preserved for the human designer (synthesis, experience, empathy, and embodied skills) rather than treating human-AI collaboration as an end in itself. Finally, the community framed the responsible use of AI itself, given its consumption of energy, water, and land, as a first-order design concern, one on which mechanical engineers, whose discipline includes thermal and energy systems, are unusually well positioned to act. Engineers, participants argued, are the bridge between advances in mathematics and computer science and the needs of society, making the humanities and social sciences full partners as the field develops.

3.10 Learning and workforce in the age of AI. Education recurred in every breakout, even when it was not the assigned prompt, and the keynote placed it at the center of the field's future. Its argument, in brief: a century and a half of industrial-era schooling optimized humans for routine cognitive and fine-motor work (in effect, "humans were the robots"), and now that machines occupy that ground [27], education must shift from *teaching* as explicit knowledge injection to *learning* as the formation of a world model, restoring the implicit, coached, experience-driven channel that industrial schooling stripped away [4]. The skills worth cultivating are those complementary to AI: empathy and emotional intelligence, presence and connection, judgment and ethics, creativity, and vision, along with comfort with ill-posed, inverse problems that demand a lifetime of context. The keynote also supplied the workshop's most-cited cautions about the transition: the "jagged frontier" of AI capability [28], along which AI excels at one task and fails at its neighbor, and the "cliff of adoption," where humans over-rely on automation precisely in the regime where it is almost, but not quite, dependable.

Breakout participants translated these themes into concrete educational concerns. Students often treat AI as a calculator, but the analogy fails: a calculator's scope is knowable, while a language model's competence boundary is invisible, and cognitive offloading of judgment risks eroding the foundational understanding needed to evaluate AI outputs; participants hypothesized that it amplifies rather than relieves classic failure modes such as design fixation [29]. These concerns crystallized into open research questions: how to ensure students achieve the same depth of understanding they would attain without AI; which skills are durably worth retaining; to what extent AI inhibits or facilitates intrinsically motivated learning within grade-oriented systems; and how to design learning environments that harness AI in service of specific competencies.

Participants distinguished *competencies* (obtaining the answer) from *meta-competencies* (justifying it), advocating assessment that rewards the justification of design decisions (oral and in-person formats, process milestones, and post-mortems that grant credit for rigorous failure analysis) over correct final answers alone, in part because industry participants reported that graduates increasingly can solve problems but cannot justify their decisions. The motif of "touching grass" captured a second recurring theme: physical prototyping and experiential work provide authentic, hard-to-fake evidence of learning and a natural check on AI over-reliance. A third framing cast AI literacy as a "defense against the dark arts": where a finite-element code is controllable magic with known assumptions, a general-purpose language model guides the user's thinking in ways that must be actively resisted, so students must be taught to control and interrogate these tools rather than be controlled by them. Participants also called for an open repository of best practices for teaching AI for designers, collaboration with K–12 education, and longitudinal studies of how AI use affects learning.

The workforce concern extends beyond the classroom. As entry-level tasks, where practitioners have traditionally learned their craft, are automated, the community will need new mechanisms to help both students and experienced engineers adapt: supervising AI and taking guidance from it are new skills, the disappearing "middle roles" of maturing technologies should inform what we train people for, and organizations that treat operators as part of the engineered system will shape whether deployed AI is used at all. Participants also emphasized broadening participation, supporting novice designers and lowering barriers for newcomers, and strengthening community structure itself: more joint design-and-

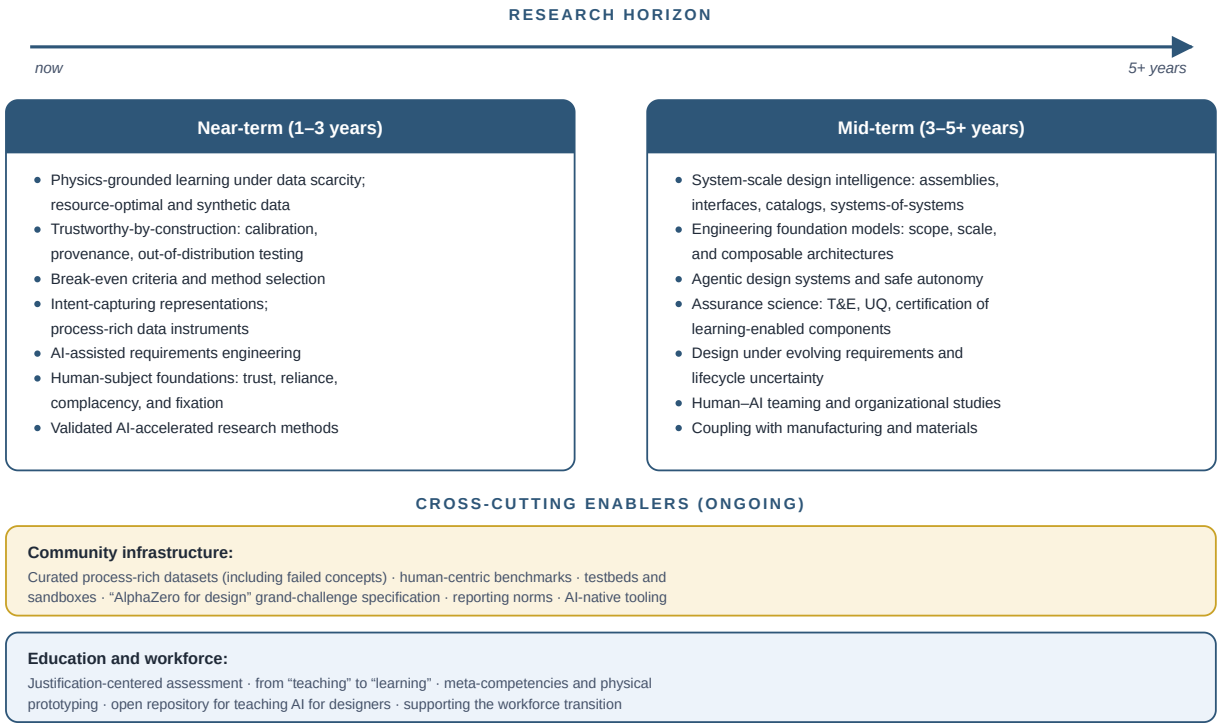


Fig. 3 The community's research agenda: near- and mid-term priorities along the research horizon, resting on a foundation of cross-cutting infrastructure and education efforts that must proceed in parallel.

systems-engineering events, deeper engagement with the machine-learning community beyond ASME venues, and candidate themes for future workshops including self-organizing and self-redefining systems, the human-computer partnership, and an explicitly sociotechnical, human-centered focus.

4 Research agenda and community recommendations

The findings above translate into a time-phased agenda, summarized in Fig. 3; Table 2 records representative research questions in the participants' own framing, drawn from the breakout record. The agenda items are the themes the organizers observed recurring across distinct sessions and formats (Appendix A); assignment to the near-term horizon indicates work the organizers judge achievable with existing data, methods, and human-subject techniques, while the mid-term horizon collects goals that depend on shared infrastructure or sustained community coordination. Ordering within each horizon implies no ranking.

Several sessions framed the agenda's ambition as the question of what a "ChatGPT moment" for EDSE would require: the conjunction of enablers that allowed language modeling to compound: a representation as universal as text, data at scale, benchmarks that focus a community's effort, and tooling that makes experimentation cheap. The engineering analogs of these enablers do not yet exist, and much of what follows can be read as the community's plan to build them, with one difference stressed throughout the workshop: engineering's bar is physical correctness and safety under uncertainty, not plausibility.

4.1 Near-term research priorities (1–3 years).

Seven near-term thrusts emerged from the breakout record, each actionable with today's methods and infrastructure.

Physics-grounded learning under data scarcity. Because high-fidelity simulations and experiments are expensive, proprietary, and typically outcome-only rather than process-rich, near-term work should advance physics-informed and constraint-aware learning that reduces the data burden [13], synthetic and resource-optimal data-generation strategies that maximize information gained per simulation or experiment, and domain-appropriate augmentation defined at the level of governing equations and geometry rather than borrowed image transforms. Hard-constraint satisfaction at inference time, rather than as a soft penalty, is a specific and tractable target (Sec. 3.3).

Trustworthy-by-construction methods. Calibrated uncertainty communication, retrieval and provenance linking generated outputs to authoritative sources, out-of-distribution stress testing, and routine reporting of the failure envelope, so that models know their limits and decline to generate for targets they have no basis to reach, together with the reporting norms that make these practices visible.

Break-even criteria and method selection. Understanding when a data-driven approach outperforms established design methods for a given class of problems, including honest comparisons against trivial baselines, and consolidating that understanding into the shared method-selection guidance called for in Sec. 4.5, so practitioners know when, and when not, to use AI at all.

Representations and process-rich data capture. Representation research that captures design constraints and intent in computable form (Sec. 3.4), paired with instrumenting design tools and design curricula to capture process-rich records (design histories, rationale, and abandoned alternatives) so that tomorrow's datasets describe how engineers actually work rather than only what they produced.

AI-assisted requirements validation. The most robust near-term gains are in requirements classification, quality checking, cross-checking for consistency, and the detection of technical debt in evolving requirement sets, and in connecting requirements to higher-fidelity design models earlier in the process (Sec. 3.5).

Human-subject foundations for human-AI design. The trust, reliance, self-confidence, and fixation effects documented in

Secs. 3.9 and 3.10 can be studied now with established design-cognition methods [8,25,26,29]: measuring and calibrating reliance, designing against complacency, countering confirmation bias, and identifying the interaction designs that preserve designer judgment; these are foundations the field needs in place before agentic tools arrive at scale.

Validated AI-accelerated research methods. Establishing the validity, reliability, and disclosure standards that let LLM-scaled qualitative analysis and synthetic-persona studies (Sec. 3.7) be trusted, reviewed, and reused as legitimate design-research instruments.

4.2 Mid-term research priorities (3–5+ years). Over a three-to-five-year horizon, participants pointed toward seven more foundational and systemic goals.

System-scale design intelligence. Foremost is the shift from single components toward assemblies and full systems: methods and representations that let AI reason about large, coupled, multi-physics systems; identify interfaces and decompositions automatically; incorporate the discrete catalog components that dominate real assemblies (Sec. 3.1); and manage the emergent behavior and long lifecycles of complex engineered systems. This thrust includes integrating learned models into multidisciplinary design optimization, where generative models can serve as priors and preconditioners for simulator-in-the-loop refinement rather than as replacements for it (Sec. 3.3).

Foundation models for engineering analysis and design. The open questions of whether, at what scale, and in what architecture: how many distinct physics one model can encompass; whether the goal is best realized as composable, physics-based building blocks combined through mixture-of-experts approaches rather than a single monolith; and the cross-modal alignment, constraint-aware decoding, and lifecycle-oriented evaluation that would make such models trustworthy (Sec. 3.4), together with the differentiable, GPU-native analysis tools that are their co-requisite.

Agentic design systems and safe autonomy. Building on Sec. 3.6: machine-operable interfaces to engineering tools, models of agent supervision and intervention, deliberate task- and workflow-level re-architecture, and a principled characterization of which design tasks admit what degree of autonomy at what risk.

Assurance science for AI-enabled engineered systems. The maturation of the SE4AI agenda within EDSE: verification, validation, and uncertainty quantification for non-deterministic, adaptive components; test-and-evaluation strategies for socio-technical emergence, including operating-envelope characterization, drift monitoring, and designed experiments that isolate the factors driving failure; certification pathways and the decision-provenance infrastructure of Sec. 3.9; and methods that translate component-level predictive performance into system-level risk within a specified operational domain and hazard model [9,11].

Design under evolving requirements and lifecycle uncertainty. Systems that reason probabilistically, update beliefs over time, and remain safe and maintainable as environments shift; living-requirements management; digital twins that incorporate the human [16,18], termed the "digital triplet" at the workshop; and, on the far horizon, design theory for the self-organizing and self-redefining systems flagged in Sec. 3.6.

Human-AI teaming at maturity. Defining and measuring what success in human-AI collaboration means; role and task architectures that deliberately allocate automation and augmentation, including the roles and abilities that should be preserved for the human designer; and field studies at the workflow and organizational levels, where Sec. 3.5 shows value is realized or lost, including longitudinal effects of AI adoption on designers, teams, and organizations.

Coupling with manufacturing and materials. Participants repeatedly connected the design agenda to manufacturing: inverse design carried to production fidelity, multi-fidelity process–structure–property models with calibrated uncertainty, and the co-design of material, architecture, and function demonstrated in the lightning talks (Sec. 3.1).

Collectively, these priorities span the core scope of NSF's EDSE program while connecting naturally to investments in education and research cyberinfrastructure, suggesting opportunities for cross-cutting programs.

4.3 Community infrastructure: datasets, benchmarks, reporting norms, and testbeds. Realizing these priorities requires shared infrastructure that no single laboratory can build alone. Participants repeatedly called for curated, openly available *datasets* [19] that capture not only final designs but also intermediate artifacts, design rationale, and failed concepts with the reasons they were abandoned, which are rarely recorded and highly valuable. Because the scarcity has the structure of a market failure, participants emphasized incentives as much as repositories: treating curated datasets and benchmarks as citable scholarly contributions; adopting artifact-evaluation practices from computer science; and developing partnership and federation approaches, such as composing privately trained models under appropriate federated-training and confidentiality safeguards, that would let industry contribute capability without surrendering proprietary data.

On *benchmarks and testbeds*, the agenda follows directly from Sec. 3.8: build the missing human-centric benchmarks for generative design and 3D-shape understanding [21], and let observation-rich testbeds precede formal metrics so the community first learns what is worth measuring [30]. A concrete proposal that recurred across sessions, first raised in the opening panel debate, was an "AlphaZero for design" [22]: a large-scale, community-authored

Table 2 Representative research questions recorded by workshop participants, organized by the six themes of Fig. 1. Condensed from the curated breakout outputs.

Theme	Representative open questions from the breakouts
Framing and disciplinary identity	Who should work at the AI–SE intersection: AI experts learning SE, systems engineers learning AI, or a new third community? What requirements and acceptance tests should EDSE set for any AI model used in engineering service? Where does AI adoption end and intellectual merit begin?
Technical foundations	What representations of design enable AI integration while capturing intent and constraints? Should the field build foundation models for engineering, and if so at what scale, spanning how many physics, monolithic or composable? What are break-even criteria for data-driven design optimization versus established methods?
Data, benchmarks, and progress	What would a large-scale *human-centric* benchmark for generative design and 3D-shape understanding look like? How should sub-task, process-rich design data be captured and shared? When should testbeds and sandboxes precede formal metrics?
Systems engineering and trust	How should non-deterministic, adaptive AI components be verified, validated, and certified? What audit mechanisms make AI-inflected design decisions accountable? How can AI identify interfaces and decompositions in systems-of-systems, and detect technical debt in evolving requirements?
Human and societal dimensions	What defines success in human-AI collaboration, and which human roles must be preserved? How should trust be calibrated and complacency designed against? How can implicit bias in training data be prevented from shaping design outcomes, and AI's energy, water, and land footprint be treated as a design constraint?
Education and workforce	How can students reach the same depth of understanding with AI as without? Which skills and meta-competencies are durable? To what extent does AI inhibit or enable intrinsically motivated learning? What mechanisms support practicing engineers as entry-level tasks are automated?

design-requirement specification that would convene researchers worldwide around one grand-challenge problem spanning optimization, sustainability, and human interpretability, giving the field a shared object of study the way landmark challenge problems have organized other communities. Participants also stressed *reporting norms* (disclosure of AI usage, mandatory out-of-distribution testing, reference-integrity checking, and reproducibility and artifact-sharing standards adapted from computer science) and *AI-native tooling*: automatic differentiation, GPU-native design software, and open platforms that are not AI algorithms but are prerequisites for AI research in engineering.

4.4 Implications for education and workforce development. Education was framed as inseparable from the research agenda, and the findings of Sec. 3.10 translate into three commitments. First, reform assessment to reward the justification of design decisions through oral, in-person, and process-based formats, with physical prototyping as authentic evidence of learning. Second, redesign learning environments so that students build foundational competence with AI rather than around it, guarding against cognitive offloading, skill atrophy, and the hypothesized AI amplification of design fixation [29]. Third, build shared resources: an open repository of best practices for teaching AI for designers, collaboration with K–12 education, and longitudinal studies of AI's effect on learning. Underlying all three is the workforce question: as the entry-level tasks where practitioners traditionally learned their craft are automated, the community must develop new mechanisms to support the transition of both students and experienced engineers into an AI-rich practice.

4.5 Community actions and future directions. The workshop closed with a set of actions that require coordination more than funding. First, tighter coupling between the engineering design and systems engineering communities themselves: the two fields approach the same AI questions from complementary ends, with design research anchored in academia and systems engineering grounded in industry and government practice, and participants called for more joint venues, pointing to the strong mixed participation of academia, industry, and government in the AI4SE & SE4AI workshop series as a model [10]. Second, deliberate cross-pollination with the machine-learning community beyond ASME venues, both to import methods and to export problems: engineering's constraint-rich, safety-critical, data-scarce setting is a regime that mainstream machine learning research underexplores, and framing it well is a contribution in itself. Third, engagement with industry on real-world validation: participants argued that the long-term measure of the community's success is not benchmark performance but meticulously tested deployment in real engineered systems, which requires partnership models that give researchers access to operational contexts and data. Fourth, practical guidance for adopters: the field currently lacks a shared taxonomy that maps classes of AI methods to classes of design problems (which method, at what fidelity, for which task, and when none is warranted), and systematic survey and best-practice efforts would immediately raise the floor of practice while the research frontier advances. Finally, participants sketched horizons for future gatherings: self-organizing and self-redefining systems, the deepening human–AI partnership in design, an explicitly sociotechnical and human-centered treatment that brings the humanities and social sciences into the room, and the near-absence of quantum and alternative computing from the present workshop's field of view.

5 Conclusion

The 2025 NSF AI-EDSE workshop marked a clear shift from the 2019 gathering, at which machine learning was a promising but peripheral tool [1]. A view that recurred across sessions was that AI is now a core enabler reshaping what it means to practice engineering design and systems engineering, and that the field must respond deliberately rather than by inertia. The most consequential framing to emerge was the workshop's embrace of both halves of the AI4SE/SE4AI duality, extended to design as AI4EDSE and EDSE4AI: using AI to strengthen engineering design and systems engineering practice while, in parallel, bringing engineering rigor to AI-enabled systems, together with a call for a new engineering discipline at the intersection of machine learning and systems engineering capable of integrating AI safely into high-consequence systems [9,11].

Three recommendations recurred with enough force to serve as rallying points. The first is an "AlphaZero" grand challenge: a large, community-authored design-requirement specification that would unite researchers around a shared problem spanning optimization, sustainability, and human interpretability. The second is shared infrastructure to overcome the field's acute data scarcity: open data repositories that capture process and failure, human-centric benchmarks for generative design and 3D-shape understanding, observation-rich testbeds, and AI-native tooling. The third is a sustained focus on the human: preserving the judgment, synthesis, and experience that distinguish designers; calibrating trust while designing against complacency; reforming education from teaching toward learning; and treating the responsible use of AI, including its consumption of energy, water, and land, as a first-order design concern.

Participants also sounded a note of caution: AI should be applied where it genuinely helps, and much valuable work in the community neither requires nor benefits from it. Realizing the opportunities ahead will depend not only on algorithmic advances but on new benchmarks, reporting norms, and educational models, and on closer collaboration between the engineering-design and systems-engineering communities and with the machine-learning community beyond. If the field succeeds, engineering design and systems engineering will not merely adopt AI but help shape it, serving as the bridge between advances in mathematics and computer science and the needs of society.

Acknowledgments

This workshop was funded by the National Science Foundation under award CMMI-2535707. The organizers thank Harrison Kim (NSF Program Director) for his support; the panelists, keynote speaker, lightning-talk presenters, and poster presenters for setting the workshop's intellectual course; the student volunteers who recorded, transcribed, and documented the sessions; and all eighty participants for two days of candid, energetic discussion. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Google NotebookLM tool was applied to the session transcriptions to assist with transcript analysis, as described in Sec. 2.3. This paper was created by the authors, who take full responsibility for the content.

Data availability

The workshop's agenda, video recordings of the talks, panels, and keynote are publicly available at https://engineering.purdue.edu/ME/AI-EDSE/Agenda. Breakout audio and transcripts are not shared, consistent with the commitment to participants that their comments would not be attributed.

Appendix A: Traceability of findings and agenda areas to workshop sessions

Table 3 lists, for each finding of Sec. 3 and each part of the agenda in Sec. 4, the workshop sessions in which the organizers observed the theme. It documents provenance, not consensus: appearance in a session means the theme was raised and discussed there, and the mapping reflects the organizers' reading of the record described in Sec. 2.3. Recordings of the framing talks, panels, keynote, and lightning talks are public (see Data Availability);

breakout sessions are identified by phase and room as listed in Sec. 2. Table 4 lists selected quantitative and high-salience empirical reports discussed in the text. Timestamps are omitted, but the named sessions' recordings are public, so each report can be located and checked against its source.

References

[1] 2019, "2019 NSF Workshop for Engineering Design and Systems Engineering (EDSE): Positioning EDSE Research for Sustained Societal Impact—Workshop Summary," Workshop report, Purdue University, West Lafayette, IN, NSF Grant No. 1916088, accessed 2026-08-11, https://engineering.purdue.edu/ME/EDSE/NSF_EDSE_2019_Workshop_Summary.pdf

[2] Simpson, T. W. and Martins, J. R. R. A., 2011, "Multidisciplinary Design Optimization for Complex Engineered Systems: Report From a National Science Foundation Workshop," ASME Journal of Mechanical Design, **133**(10), p. 101002.

[3] Papalambros, P. Y., 2015, "Design Science: Why, What and How," Design Science, **1**, p. e1.

[4] Sarma, S. and Yoquinto, L., 2020, *Grasp: The Science Transforming How We Learn*, Doubleday, New York.

[5] Alam, M. F. and Ahmed, F., 2025, "GenCAD: Image-Conditioned Computer-Aided Design Generation with Transformer-Based Contrastive Representation and Diffusion Priors," Transactions on Machine Learning Research, arXiv:2409.16294.

[6] Sridhara, S. and Suresh, K., 2025, "A Visual Representation of Engineering Catalogs Using Variational Autoencoders," Journal of Mechanical Design, **147**(4), p. 041708.

[7] Mu, H. D., Tang, M., Gao, W., and Chen, W., 2025, "GUIDe: Generative and Uncertainty-Informed Inverse Design for On-Demand Nonlinear Functional Responses," arXiv preprint arXiv:2509.05641.

[8] Chong, L., Raina, A., Goucher-Lambert, K., Kotovsky, K., and Cagan, J., 2023, "The Evolution and Impact of Human Confidence in Artificial Intelligence and in Themselves on AI-Assisted Decision-Making in Design," ASME Journal of Mechanical Design, **145**(3), p. 031401.

[9] McDermott, T., DeLaurentis, D., Beling, P., Blackburn, M., and Bone, M., 2020, "AI4SE and SE4AI: A Research Roadmap," INSIGHT, **23**(1), pp. 8–14.

[10] Szajnfarber, Z., McDermott, T. A., and Sitterle, V., 2026, "Challenges and Opportunities for Systems Engineering in the Age of AI: SERC Perspective," INSIGHT, **29**(3), pp. 10–14.

[11] National Academies of Sciences, Engineering, and Medicine, 2025, *Machine Learning for Safety-Critical Applications: Opportunities, Challenges, and a Research Agenda*, The National Academies Press, Washington, DC.

[12] Elrefaie, M., Morar, F., Dai, A., and Ahmed, F., 2024, "DrivAerNet++: A Large-Scale Multimodal Car Dataset with Computational Fluid Dynamics Simulations and Deep Learning Benchmarks," *NeurIPS 2024 Datasets and Benchmarks Track*, Vancouver, Canada, https://openreview.net/forum?id=c4NnhBi4oM

[13] Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L., 2021, "Physics-Informed Machine Learning," Nature Reviews Physics, **3**, pp. 422–440.

[14] Lu, L., Jin, P., Pang, G., Zhang, Z., and Karniadakis, G. E., 2021, "Learning Nonlinear Operators via DeepONet Based on the Universal Approximation Theorem of Operators," Nature Machine Intelligence, **3**(3), pp. 218–229.

[15] Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A., 2021, "Fourier Neural Operator for Parametric Partial Differential Equations," *International Conference on Learning Representations (ICLR)*, https://openreview.net/forum?id=c8P9NQVtmnO

[16] Madni, A. M., Madni, C. C., and Lucero, S. D., 2019, "Leveraging Digital Twin Technology in Model-Based Systems Engineering," Systems, **7**(1), p. 7.

[17] Bommasani, R., Hudson, D. A., Adeli, E., et al., 2021, "On the Opportunities and Risks of Foundation Models," arXiv preprint arXiv:2108.07258.

[18] Grieves, M. and Vickers, J., 2017, "Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems," *Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches*, F.-J. Kahlen, S. Flumerfelt, and A. Alves, eds., Springer, Cham, pp. 85–113.

[19] Regenwetter, L., Curry, B., and Ahmed, F., 2022, "BIKED: A Dataset for Computational Bicycle Design With Machine Learning Benchmarks," ASME Journal of Mechanical Design, **144**(3), p. 031706.

[20] Patterson, D. A., 2012, "For Better or Worse, Benchmarks Shape a Field: Technical Perspective," Communications of the ACM, **55**(7), p. 104.

[21] Zhou, S., Gordon, M. L., Krishna, R., Narcomey, A., Fei-Fei, L., and Bernstein, M. S., 2019, "HYPE: A Benchmark for Human eYe Perceptual Evaluation of Generative Models," *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Vancouver, Canada, pp. 3449–3461.

[22] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D., 2018, "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go Through Self-Play," Science, **362**(6419), pp. 1140–1144.

[23] Sutton, R. S., 2019, "The Bitter Lesson," Blog essay, Incomplete Ideas, accessed 2026-08-11, http://www.incompleteideas.net/IncIdeas/BitterLesson.html

[24] Brynjolfsson, E., 2022, "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence," Daedalus, **151**(2), pp. 272–287.

[25] Lee, J. D. and See, K. A., 2004, "Trust in Automation: Designing for Appropriate Reliance," Human Factors, **46**(1), pp. 50–80.

[26] Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., and Cagan, J., 2022, "Human Confidence in Artificial Intelligence and in Themselves: The Evolution and Impact of Confidence on Adoption of AI Advice," Computers in Human Behavior, **127**, p. 107018.

[27] Moravec, H., 1988, *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press, Cambridge, MA.

[28] Dell'Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R., 2026, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality," Organization Science, **37**(2), pp. 403–423.

[29] Jansson, D. G. and Smith, S. M., 1991, "Design Fixation," Design Studies, **12**(1), pp. 3–11.

[30] McGreivy, N. and Hakim, A., 2024, "Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations," Nature machine intelligence, **6**(10), pp. 1256–1269.

Table 3 Sources of findings and agenda areas. FT-ED/FT-SE = engineering design / systems engineering framing talks; K = keynote; P1–P3 = perspectives panels; LT = lightning talks (topic in parentheses); B*p.r* = breakout phase *p* (1–3), room *r* (1–3, in the order listed in Sec. 2.2); RO = day-one report-out; RQ = participants' written research questions; RN = rapporteur notes; CL = closing discussion.

Finding / agenda area	Sessions in which the theme was observed
3.1 State of the field	FT-ED; P2; LT (all ten); B1.2; B3.2
3.2 AI4EDSE and EDSE4AI framing	FT-SE; P3; B3.1; CL
3.3 Data scarcity, hallucination, physics-grounding	FT-ED; P1; LT (inverse design; additive manufacturing); B1.1; B1.3; RQ
3.4 Representations and foundation models	B1.1; B1.2; LT (CAD generation; material systems; catalogs); FT-ED; B3.1
3.5 Systems engineering: requirements, lifecycle	FT-SE; P3; LT (digital engineering); B1.3; B3.2; RQ
3.6 Agentic workflows, digital readiness, virtual environments	P1; K; LT (digital engineering; workplace agents); B1.2; B3.1
3.7 AI in design research methodology	B2.1; P2; LT (participatory design)
3.8 Benchmarks, testbeds, case studies	P1; FT-ED; B1.2; B3.1; P3; RN; RQ
3.9 Trust, accountability, ethics	P1; K; LT (trust and reliance; human-AI testbed; workplace agents); B2.2; B3.2; RN
3.10 Learning and workforce	K; B1.1; B1.2; B2.2; B3.3; RO; RQ (education arose in all three breakout phases)
4.1 Near-term priorities	Distilled from the sources of 3.3, 3.4, 3.5, 3.7, and 3.9, with B3 (all rooms) and RQ
4.2 Mid-term priorities	Distilled from the sources of 3.1, 3.2, 3.4, 3.5, 3.6, and 3.9, with B3 (all rooms) and RQ
4.3 Community infrastructure	P1; B1.3; B3.1; RO; RQ; CL
4.4 Education and workforce development	K; B3.3; RQ
4.5 Community actions	CL; RO; B3 (all rooms); RQ

Table 4 Selected empirical reports appearing in the text, their session of origin, and their evidentiary status.

Reported claim	Session	Status
Bibliometric trends in the *Journal of Mechanical Design*: rising machine learning terms (framing talk); a majority of papers published in the first half of 2025 using machine learning or AI (editors' panel count)	FT-ED; P2	Workshop testimony
Aerodynamic dataset of 8,000 car geometries requiring millions of CPU-hours of simulation for one boundary condition	FT-ED	Workshop testimony
CAD command-sequence reconstruction from images or sketches (GenCAD)	LT (CAD generation)	Peer-reviewed [5]
Roughly 30,000 parts per vehicle, most selected from catalogs	LT (catalogs)	Workshop testimony
Out-of-distribution failure of conditional generative models in inverse design	LT (inverse design)	Workshop testimony
Multi-fidelity melt-pool prediction with calibrated confidence bounds	LT (additive manufacturing)	Workshop testimony
Co-design demonstrations in material systems: mechanical cloaking, 4D-printed liquid-crystal-elastomer networks, magnetic soft robots	LT (material systems)	Workshop testimony
Differentiable rewrite of hydrodynamic solvers as the enabling investment for learning-based design optimization	LT (offshore design)	Workshop testimony
Text classifiers evaluated did not reliably distinguish LLM-generated systems-engineering artifacts; close inspection found them to be of poor quality	LT (digital engineering)	Workshop testimony
Substantial design cycle-time reduction from an authoritative source of truth (process-simulation study)	LT (digital engineering)	Workshop testimony
Confidence in an AI recommender easily lost and hardly gained; users of generative tools overlooked the best-balanced designs	LT (trust and reliance)	Workshop testimony
Self-confidence effects of unannounced changes in AI accuracy	LT (trust and reliance)	Peer-reviewed [8,26]
Emotion recognition and social connection suggested as possibly compensating for weaker performance in a modular design-assistant testbed	LT (human-AI testbed)	Workshop testimony
Workplace-agent studies: convergence on calm, practical agent personalities; embodiment shaping preferred autonomy; long-term complacency; interpersonal trust extended to machines	LT (workplace agents)	Workshop testimony
LLM-scaled qualitative coding of roughly 10,000 responses with strong inter-coder reliability, reported to have passed review at a top management journal	B2.1	Workshop testimony
Simulated clinic-nurse persona pilot surfacing colonial framings and fabricated local practices	B2.1	Workshop testimony
Sketching and CAD study interfaces set up in about a month using generative models, versus a Ph.D.-scale effort previously	B2.1	Workshop testimony
Surrogate-model errors plateauing above practitioner-required thresholds	B1.2	Workshop testimony
Generative CAD limited to simple shapes in current practice	B1.2	Workshop testimony
Language-model competence collapsing under circular evaluation	B3.2	Workshop testimony
Familiar parameterizations overstating intrinsic dimensionality (airfoil example); compression reducing hallucination	B1.1	Workshop testimony
Heavy-truck dataset dominated by steady highway driving, carrying little information	B1.3	Workshop testimony
Bicycle-seat training data reflecting seats designed predominantly for men's anatomy	B3.2	Workshop testimony
Company complaints, relayed by a participant, of graduates who solve problems but cannot justify decisions	B2.2	Workshop testimony
Scale and composition of the AI4SE & SE4AI workshop series	FT-SE	Workshop testimony; see also [10]