Artificial Intelligence for Enhanced Experimentation in Software Ecosystems: A Systematic Literature Review

https://doi.org/10.63962/IBDN5991

Shady Hegazy
Siemens Technology
Siemens AG
Munich, Germany
shady.hegazy@siemens.com

Christoph Elsner
Siemens Technology
Siemens AG
Munich, Germany
christoph.elsner@siemens.com

Jan Bosch
Department of Computer Science and Engineering
Chalmers University of Technology
Göteborg, Sweden
jan.bosch@chalmers.se

Helena Holmström-Olsson
Department of Computer Science and Media Technology
Malmö University
Malmö, Sweden
helena.holmstrom.olsson@mau.se

***Abstract*— Software ecosystems are networks of interconnected actors co-creating value through a shared technological platform, achieving accelerated growth via network effects. This interconnectedness creates a complex decision space, requiring experimentation-based, data-driven decision making. This study presents a systematic literature review on AI-enhanced experimentation in software ecosystems. Following structured screening and quality assessment, 63 studies were included. Extracted data underwent descriptive, thematic, and cross-analyses. The study offers three contributions: an overview of AI-enhanced experimentation objectives; a phase-aligned analysis; and a framework for AI integration into experimentation pipelines.**

Keywords—artificial intelligence, experimentation, software ecosystems, causal inference

I. Introduction

Software ecosystems (SECOs) have transformed the way value is created in many industries, introducing transformational shifts from vertical integration and firm-confined value-creation to co-creation of value through a network of interconnected actors who collaborate via a shared technological platform within one ecosystem. An example from the mobile industry is the success of ecosystem-oriented approaches of Apple's IOS and Google's Android in attracting around 300,000 third-party developers each in an ecosystem around their application marketplaces, versus the failure of Blackberry's vertically integrated approach which relied on 8000 in-house developers (Wang et al., 2017). However, such growth comes at many costs including additional complexities in orchestration and governance of such multi-sided ecosystems. For instance, TradeLens, a blockchain-based platform ecosystem developed by Maersk and IBM, was ultimately shut down, largely due to governance and adoption challenges as the platform struggled with aligning incentives among highly interdependent and often competing actors, the lack of trust in the platform sponsor's neutrality, hesitancy over data sharing, and difficulties managing standardization and interoperability across legacy systems, which hindered adoption among third-party industry partners (Najati, 2025). Such complexity necessitates adopting a data-driven decision making approach in order to reduce decision risks and uncertainty in an environment with various types of actors who have different, and sometimes conflicting, goals (Liu et al., 2020). While analytics can provide relevant insights to effectively inform the decision making process, online controlled experiments are considered the gold standard for evidence-based decision-making (Liu et al., 2020). Experiments in SECOs "often exhibit complex network effects. Consequently, unless designed carefully, the experiments could suffer from interference" (Candogan et al., 2023). In addition to *interference* and *bias*, network effects in SECOs contribute to the propagation of negative effects of users' exposure to under-performing variants during experiments, which is often known as *regret (Spang et al., 2021)*. The vast amount of data generated in SECOs present opportunities for the utilization of advances in artificial intelligence (AI) to ease these challenges (Charness et al., 2023). While prior secondary studies explored contribution of certain AI methods to experimentation pipelines (Charness et al., 2023), different aspects of online controlled experiments in general (Auer & Felderer, 2018; Quin et al., 2024; Ros et al., 2018), there was no secondary research addressing AI-enhanced experimentation in SECOs. Hence, we conducted this systematic literature review aiming to provide the following contributions: an overview of the major objectives for AI application in experimentation pipelines in SECOs; a phase-aligned analysis of AI-enhanced experimentation; and a framework for AI integration into experimentation pipelines.

II. Methodology

This study follows a systematic literature review methodology to ensure comprehensive and replicable coverage of the existing research.

A. Research Question

We applied the Population, Intervention, Comparison, Outcome, and Context (PICOC) criteria to produce the

following structure to match the study scope (Kitchenham & Charters, 2007).

- Population: SECOs.
- Intervention: Experimentation.
- Comparison: AI-enabled enhancements to experimentation processes.
- Outcome: Enhanced data-driven decision making through more efficient experimentation processes.
- Context: SECOs in business contexts.

Using the above structure, we formulated the following research question: In what ways has AI-supported experimentation been applied in software ecosystems to improve data-driven decision making?

B. Search Strategy

Based on the breakdown of the research question structure, search query terms were selected for each section of the PICOC criteria, as shown in Table I.

TABLE I. SEARCH QUERY TERMS.

Population	”platform ecosystem*” OR ”software ecosystem*” OR ”seco” OR ”digital ecosystem*” OR ”platform*”
Intervention	”experimentation” OR (”experiment*” AND ”design*”) OR ”online experiment*” OR ”online controlled experiment*” OR ”controlled experiment*” OR ”user testing” OR ”a/b testing”
Comparison	
Outcome	”quantitative” OR ”driven” OR ”analy*” OR ”inference” OR ”continuou*” OR ”adapt*” OR ”data” OR ”network*” OR ”graph” OR ”agent*” OR ”sampl*” OR ”cluster*” OR ”classif*” OR ”group*” OR ”causal*” OR ”bias*”
Context	”business*” OR ”industry” OR ”market” OR ”customer” OR ”user”

The search query was executed on the titles and abstracts of research articles in the Scopus, ACM Digital Library, and IEEE Xplore databases, resulting in 1,349, 597, and 113 hits, respectively. The meta-data of the resulting 2,059 studies was retrieved from the corresponding sources.

C. Inclusion and Exclusion Criteria

The set of inclusion and exclusion criteria used for this review are listed in Table II along with the number of exclusion decisions for which each criterion was the most prominent basis.

TABLE II. INCLUSION AND EXCLUSION CRITERIA.

Existence	Criterion	Absence	Exclusions
IC1	A primary study, not a secondary or a tertiary study.	***EC1***	77
IC2	Published in a peer-reviewed journal, conference proceeding, or book chapter not in gray literature reports, blog posts, or non-academic publications.	***EC2***	9
IC3	Published in English.	***EC3***	5
IC4	Not a duplicate or a version of another included study.	***EC4***	421
IC5	Have a significant component related to SECOs.	***EC5***	608
IC6	Discusses experimentation in SECOs with the aim of enabling or enhancing data-driven decision making.	***EC6***	846
IC7	Focuses on business contexts.	***EC7***	5

D. Quality Assessment

The included 88 studies underwent a quality assessment process using a variation of the Standard Quality Assessment Criteria (SQAC) tailored for quantitative research (Kmet et al., 2004). The criteria were scored on a 3-point Likert scale, with zero points for unmet criteria, one point for partially met criteria, and two points for met criteria. The average quality score after removing the 25 disqualified studies was 73.6%.

E. Data Extraction

The following data categories and the underlying data points were extracted from the full texts of the 63 included studies:

- Identifiers: study title; authors; publication year; abstract.
- Software ecosystem: type; number of sides; actor type; industry.
- Experimental Design: method; internal platform; analysis unit; randomization unit; sampling technique; traffic allocation; evaluation metrics.
- Data: volume; types; sources; collection rate; processing rate; AI support.
- AI Support: phase; main techniques; family; task; direct outcome.

III. RESULTS

A. Major Objectives of AI Use in Experimentation Pipelines

1) Maximization of experimentation outcomes

Experimentation results typically help declare a higher performing decision, policy, or variant. However, the use of AI could enable further outcomes by using the resulting data to inform future experiments or generate new insights. For example, the study in (Duivesteijn et al., 2017) used experiment data to dynamically serve different variants to different user groups, through the use of AI methods such as exceptional model mining.

2) Sampling optimization

AI methods were also deployed to construct more informative and balanced samples (Candogan et al., 2023). Studies used graph-based clustering to define interference-aware clusters, matching techniques to reduce covariate imbalance, and segmentation algorithms to stratify populations based on latent characteristics (Brennan et al., 2022).

3) Experimentation cost reduction

Multiple studies applied AI to dynamically reallocate traffic away from underperforming variants, reducing exposure to inferior treatments and shortening the duration required to reach conclusive results (Glynn et al., 2020). Other studies used various AI for the task of offline policy evaluation to reduce the exposure to underperforming variants during live experiments

by predicting higher performing variants and allocating more traffic towards them (Li & Xie, 2020).

4) Intelligent experimentation pipelines

AI methods were used to dynamically revise experimental designs and pipeline parameters based on emerging evidence. For instance, the study in (Bojinov et al., 2023) described the use of AI to predict carryover bias from treatments in *switchback* experiments in order to dynamically space sequential treatments accordingly.

5) Causal inference enhancement

As ideal experimental conditions are hard to achieve in interconnected SECOs, AI was commonly applied to enhance the robustness of causal inference. For instance, several studies leveraged deep learning based causal inference, ensemble learning, and causal forests to estimate treatment effects in the presence of heterogeneity, interference, or selection bias (Ye et al., 2023). In addition, AI methods, such as analysis, were used to construct more accurate overall evaluation criteria, thus indirectly enhancing causal inference (Hornback et al., 2023).

B. Phase-Based AI Support in Experimentation Pipelines

1) Design phase

AI support in the planning and design phases focused on structuring experimental units, forecasting outcomes, and policy evaluation. AI was also used for causal forecasting and offline policy evaluation to prioritize high-performing treatments and reduce experimentation risks and costs.

2) Execution phase

In this phase, AI methods were used to dynamically reallocate traffic, forecast responses, and adjust experiment structures based on emerging evidence. AI was commonly applied to minimize regret and optimize policy rollout under budget and engagement constraints. While no end-to-end AI-orchestrated pipelines were observed, AI support in this phase enabled adaptive, intelligent experimentation with substantial efficiency gains and regret reduction.

3) Evaluation phase

AI support in the post-experiment phase focused on causal inference, bias and interference correction, heterogeneity analysis, and insight generation for future experiments. Its value was especially evident in ecosystems with noisy, high-dimensional data such as networking ecosystems. For instance, causal forests and ensemble methods were used to model heterogeneous effects in complex ecosystems.

C. A Framework for AI-Enhanced Experimentation Pipelines

Through our analysis of the reviewed studies we developed a framework, described in Table III, that formalizes the introduction of AI methods in experimentation pipelines by connecting phases, objectives, and methods.

IV. Conclusion

Through a systematic literature review, we investigated how artificial intelligence can enhance experimentation in software ecosystems. From an initial pool of 2,059 studies, 63 met the inclusion and quality criteria. We analyzed the integration of AI across experimentation pipelines, identifying key goals such as bias mitigation, regret minimization, and experimentation cost–benefit optimization. Our phase-aligned analysis showed AI support is most common in the design and evaluation phases, with frequent use of methods like graph clustering, reinforcement learning, and causal forests. We propose a structured framework linking AI techniques to experimentation enhancement objectives and phases, providing a foundation for formalization of AI integration in experimentation pipelines.

TABLE III. A framework for AI-enhanced experimentation pipelines (AIEXP).

	Design	**Execution**	**Evaluation**
Sampling optimization	Graph cluster randomization; hierarchical clustering.	Online bias/variance balancing.	Regression-adjusted CUPED; covariate regression.
Experimentation cost reduction	Monte Carlo simulations; Bayesian machine learning.	Sequential policy learning; multi-armed bandits.	Meta-learning; experiment recommender systems.
Intelligent experimentation pipelines	Adaptive segmentation.	Automated early stopping; Dynamic guardrails exploration.	Generative AI for reporting.
Causal inference enhancement	Predictive power analysis.	Delayed feedback prediction; deep Q-learning.	Debiased machine learning; Deep learning causal inference; sentiment analysis; ensemble learning; transfer learning.
Outcome maximization	Offline policy learning.	Contextual bandits; uplift modelling.	Exceptional model mining; exploratory data analysis.

References

[1] Auer, F., & Felderer, M. (2018). Current State of Research on Continuous Experimentation: A Systematic Mapping Study. *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 335–344.

[2] Bojinov, I., Simchi-Levi, D., & Zhao, J. (2023). Design and Analysis of Switchback Experiments. In *Management Science* (Vol. 69, Issue 7, pp. 3759–3777). https://doi.org/10.1287/mnsc.2022.4583

[3] Brennan, J., Mirrokni, V., & Pouget-Abadie, J. (2022). Cluster randomized designs for one-sided bipartite experiments. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 37962–37974.

[4] Auer, F., & Felderer, M. (2018). Current State of Research on Continuous Experimentation: A Systematic Mapping Study. *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 335–344. https://doi.org/10.1109/SEAA.2018.00062

[5] Bojinov, I., Simchi-Levi, D., & Zhao, J. (2023). Design and Analysis of Switchback Experiments. In *Management Science* (Vol. 69, Issue 7, pp. 3759–3777). https://doi.org/10.1287/mnsc.2022.4583

[6] Brennan, J., Mirrokni, V., & Pouget-Abadie, J. (2022). Cluster randomized designs for one-sided bipartite experiments. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 37962–37974.

[7] Candogan, O., Chen, C., & Niazadeh, R. (2023). Correlated Cluster-Based Randomized Experiments: Robust Variance Minimization.

Proceedings of the 24th ACM Conference on Economics and Computation, 411. https://doi.org/10.1145/3580507.3597820

[8] Charness, G., Jabarian, B., & List, J. (2023). Generation Next: Experimentation with AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4574623

[9] Duivesteijn, W., Farzami, T., Putman, T., Peer, E., Weerts, H. J. P., Adegeest, J. N., Foks, G., & Pechenizkiy, M. (2017). Have It Both Ways—From A/B Testing to A&B Testing with Exceptional Model Mining. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 10536 LNAI* (pp. 114–126). https://doi.org/10.1007/978-3-319-71273-4_10

[10] Glynn, P., Johari, R., & Rasouli, M. (2020). Adaptive experimental design with temporal interference: A maximum likelihood approach. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 15054–15064.

[11] Hornback, A., Buckley, S., Kos, J., Bunin, S., An, S., Joyner, D., & Goel, A. (2023). A Scalable Architecture for Conducting A/B Experiments in Educational Settings. *Proceedings of the Tenth ACM Conference on Learning @ Scale*, 373–377. https://doi.org/10.1145/3573051.3596190

[12] Kitchenham, B. A., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (No. EBSE 2007-001). Keele University and Durham University Joint Report / Keele University.

[13] Kmet, L. M., Lee, R. C., & Research, A. H. F. for M. (2004). *Standard Quality Assessment Criteria for Evaluating Primary Research Papers from a Variety of Fields*. Alberta Heritage Foundation for Medical Research.

[14] Li, G., & Xie, H. (2020). Auxiliary Decision-making for Controlled Experiments based on Mid-term Treatment Effect Prediction: Applications in Ant Financial's Offline-payment Business. 2, 19–30. https://doi.org/10.5220/0009770500190030

[15] Liu, C. H. B., Chamberlain, B. P., & McCoy, E. J. (2020). What is the Value of Experimentation and Measurement?: Quantifying the Value and Risk of Reducing Uncertainty to Make Better Decisions. In *Data Science and Engineering* (Vol. 5, pp. 152–167). https://doi.org/10.1007/s41019-020-00121-5

[16] Najati, I. (2025). Exploring the failure factors of blockchain adopting projects: A case study of tradelens through the lens of commons theory. *Frontiers in Blockchain*, *8*, 1503595. https://doi.org/10.3389/fbloc.2025.1503595

[17] Quin, F., Weyns, D., Galster, M., & Silva, C. C. (2024). A/B testing: A systematic literature review. *Journal of Systems and Software*, *211*, 112011. https://doi.org/10.1016/j.jss.2024.112011

[18] Ros, R., Ros, R., Runeson, P., & Runeson, P. (2018). Continuous experimentation and A/B testing: A mapping study. *Null*. https://doi.org/10.1145/3194760.3194766

[19] Spang, B., Hannan, V., Kunamalla, S., Huang, T.-Y., McKeown, N., & Johari, R. (2021). Unbiased experiments in congested networks. *Proceedings of the 21st ACM Internet Measurement Conference*, 80–95. https://doi.org/10.1145/3487552.3487851

[20] Wang, H., Liu, Z., Guo, Y., Chen, X., Zhang, M., Xu, G., & Hong, J. (2017). An Explorative Study of the Mobile App Ecosystem from App Developers' Perspective. *Proceedings of the 26th International Conference on World Wide Web*, 163–172. WWW '17: 26th International World Wide Web Conference. https://doi.org/10.1145/3038912.3052712

[21] Ye, Z., Zhang, Z., Zhang, D. J., Zhang, H., & Zhang, R. (2023). Deep Learning Based Causal Inference for Large-Scale Combinatorial Experiments: Theory and Empirical Evidence. *Proceedings of the 24th ACM Conference on Economics and Computation*, 1160. https://doi.org/10.1145/3580507.3597718