AI-ARENA
Ethical AI Agents Framework

Financira
Europska unija
NextGenerationEU



Framework for Ethical Use of AI Agents in Multidisciplinary Research of the Digital Society

D3.2. - Framework Objectives Definition

This project was funded by NextGenEU

Principal investigator:

Markus Schatten

Copyright © 2026 Artificial Intelligence Laboratory

PUBLISHED BY ARTIFICIAL INTELLIGENCE LABORATORY,
FACULTY OF ORGANIZATION AND INFORMATICS, UNIVERSITY OF ZAGREB

HTTP://DRAGON.FOI.HR:8888/AI-ARENA

Licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License (the "License"). You may not use this file except in compliance with the License. You may obtain a copy of the License at http://creativecommons.org/licenses/by-nc/3.0. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

Technical Report No. AIL202602 – First release, March 2026, edit November 2026

Document compiled by: Markus Schatten with inputs from other project team members. ChatGPT 5 has been used as a writing tool. All text and references have been checked by the author and the author takes full responsibility for the content of this report.

AI-ARENA is financed in full by the European Union - Next Generetion EU under the project AI-ARENA FOI-IIP-2025-04.

Contents

1 Project Description 1
1.1 Abstract 1
1.2 Introduction 1
1.3 Team Members 3

2 Introduction 5
2.1 Background and motivation 5
2.2 Research problem and purpose 6
2.3 Scope of the study 6

3 Literature Review 7
3.1 AI agents and the scientific research process 7
3.2 Scientific validity, reproducibility and evaluation 8
3.3 Responsibility, security and human oversight 8
3.4 Evidence from the six research domains 9
3.4.1 Educational computer games 9
3.4.2 Digital forensics 9
3.4.3 Biometrics 9
3.4.4 Information security 9
3.4.5 Social choice theory 9
3.4.6 Queueing theory 10
3.5 Synthesis and research gap 10

4 Methodology 11
4.1 Research design 11

4.2 Problem identification and requirements elicitation 11
4.3 Artefact construction and iterative refinement 12
4.4 Evaluation logic 13
4.5 Methodological and ethical safeguards 13

5 Research Problems and Objectives 15
5.1 Cross-cutting problem structure 15
5.2 The agent as a scientific instrument 16
5.3 Research questions 16
5.4 Scientific research objectives 17
5.5 Requirements of the proposed framework 17
5.6 Domain-specific research and evaluation designs 18
5.7 Testable design propositions 19

6 Discussion 21
6.1 Scientific contribution and its evidentiary requirements 21
6.2 Position relative to existing approaches 21
6.3 Limitations and threats to validity 22
6.4 Implications for research practice 22

7 Conclusion 25

Bibliography 27

1. Project Description

1.1 Abstract

The AI-ARENA project focuses on developing and evaluating a methodological framework for applying artificial intelligence (AI)-based agents in multidisciplinary research within the digital society. The core scientific approach will be Design Science Research (DSR), within which an AI framework will first be designed as an artifact, followed by a series of case studies in key domains. Each case study includes the development and evaluation of specific AI agents in different contexts: biometrics, digital forensics, information security, computer games, education, queueing theory, social choice theory, and the ethics and bias of AI systems. The agents will serve as tools for analysis, automation, and experimentation within specific problems, and the project will also address the technical, ethical, and regulatory challenges of their use. The goal is to create a flexible and applicable framework that connects AI technology with domain knowledge and scientific methodology, delivering concrete scientific and educational outcomes and contributing to open and responsible science.

1.2 Introduction

In recent years, advances in generative artificial intelligence (GenAI), particularly large language models (LLMs), have led to the emergence of so-called AI agents – systems designed not only to generate content but also to plan and autonomously execute tasks in complex digital environments [14]. AI agents represent the next stage in the evolution of intelligent systems, with a particular emphasis on autonomous action, multi-step reasoning, tool use, and inter-agent collaboration [21, 28].

Existing research has demonstrated significant progress in agent architectures, particularly in integrating planning, reflection, and tool execution. However, numerous challenges remain, including high inference costs, difficulties in evaluation and standardization, and fragmented approaches to agent development and testing [16]. Although AI agents have shown promising results in domains such as web navigation, business process automation, and education [25], a systematic, methodologically grounded framework to guide their application in broader research

contexts is still lacking.

Furthermore, the security and ethical implications of agent-based systems are becoming increasingly prominent. AI agents operate in high-risk contexts (e.g., digital forensics and cybersecurity), where their autonomy and behavioral complexity raise important questions concerning accountability, transparency, and governance [3, 18]. Challenges related to unpredictable user inputs, the complexity of internal execution processes, and interactions with untrusted external entities have also been identified [8].

Despite the growing body of literature addressing agent architectures, evaluation methodologies, domain-specific applications, and security challenges, an integrated scientific framework that connects the development, application, and evaluation of AI agents across different research domains within a coherent approach has yet to be proposed. Such a framework is essential to enable:

- systematic investigation and comparison of different approaches to agent development,
- validation of their impact in specific contexts (e.g., biometrics, education, digital forensics, computer games, cybersecurity, and digital society),
- and the development of guidelines for the responsible and ethically acceptable use of AI agents in digital society.

The AI-ARENA project aims to address this gap by proposing the development and empirical evaluation of a framework for the use of AI agents across multiple interconnected research domains, employing Design Science Research (DSR) as its foundational scientific methodology.

1.3 Team Members



Markus Schatten (Project leader)
Project leader; Full Professor
Artificial Intelligence Laboratory
Faculty of Organization and Informatics, University of Zagreb



Petra Grd
Associate Professor
Center for Forensics, Biometrics and Privacy
Faculty of Organization and Informatics, University of Zagreb



Marcel Maretić
Assistant Professor
Department of Quantitative Methods
Faculty of Organization and Informatics, University of Zagreb



Bogdan Okreša Đurić
Assistant Professor
Artificial Intelligence Laboratory
Faculty of Organization and Informatics, University of Zagreb



Tomislav Peharda
Teaching Assistant
Artificial Intelligence Laboratory
Faculty of Organization and Informatics, University of Zagreb



Igor Tomičić
Associate Professor
Artificial Intelligence Laboratory / Center for Forensics, Biometrics and Privacy
Faculty of Organization and Informatics, University of Zagreb



Domagoj Tuličić
Teaching Assistant
Center for Forensics, Biometrics and Privacy
Faculty of Organization and Informatics, University of Zagreb



Petra Žugec
Assistant Professor
Department of Quantitative Methods
Faculty of Organization and Informatics, University of Zagreb

2. Introduction

2.1 Background and motivation

Artificial intelligence (AI) agents are increasingly used to coordinate activities that extend beyond the generation of text. By combining language models with planning, memory, software tools and access to external information, an agent may search the literature, prepare executable analyses, run simulations, inspect outputs and propose subsequent actions. Recent accounts of scientific agents describe both the breadth of these opportunities and the substantial methodological work required to integrate them into research practice [11, 12, 19]. Such systems differ from conventional software instruments because the sequence of actions, the interpretation of instructions and even the intermediate computational procedures may vary across executions.

This flexibility creates a scientific problem that cannot be reduced to software performance. Successful execution of a requested task is not equivalent to a valid inference about the phenomenon under investigation. An executable script may implement an inappropriate statistical procedure; a correct-looking summary may misrepresent its cited sources; and a plausible numerical estimate may contradict the assumptions of the mathematical model. Evaluations of scientific agents have documented failures in coding, experimental reasoning and the recovery of established scientific results, even when agents are given substantial computational resources [4, 34]. The problem is amplified when agents act with access to sensitive data, research infrastructure or consequential decision processes.

Scientific research therefore requires a distinction between the *functional success of an agent* and the *evidential validity of a research outcome*. The former concerns whether a computational task is completed; the latter concerns whether the task, its assumptions, data, methods and interpretation support the stated scientific claim. Messeri and Crockett [22] further argue that increasingly capable AI tools can foster an illusion of understanding, particularly when readily available explanations displace critical examination of scientific assumptions. The implication is not that agent assistance is necessarily unsuitable for science, but that methodological safeguards must be explicit and commensurate with the role assigned to the agent.

2.2 Research problem and purpose

A growing literature discusses agent architectures, benchmark performance, scientific workflows, provenance, responsible AI and discipline-specific validation. Nevertheless, combining these strands into a procedure that guides the *scientific use* of agents across heterogeneous disciplines remains a non-trivial design challenge. Approaches developed for digital forensics, for example, address evidence integrity and legal defensibility; methods for queueing models emphasise analytical and numerical validity; and educational studies require valid measures of learning and appropriate participant protection [2, 5, 17, 30, 32]. A common procedural framework must preserve these differences rather than replace them with a generic agent score.

The research problem addressed here is the design of a methodological framework that connects the role and autonomy of an AI agent with scientific method selection, evidence provenance, independent validation, responsible operation and an evaluable research process. The central research question is:

> *How can a modular methodological framework be designed and evaluated to support scientifically valid, reproducible, transparent and responsible use of AI agents across heterogeneous research domains?*

The intended research artefact is a *methodological framework*, not any particular agent or language model. Its potential scientific contribution lies in an explicit conceptualisation of agents as research instruments, an integrated set of procedures and design principles, and evidence that following those procedures offers value beyond ad hoc agent use. Domain-specific agents provide opportunities to instantiate, challenge and refine the framework; their technical performance alone cannot validate it.

2.3 Scope of the study

Six domains provide deliberately heterogeneous settings for developing and examining the approach: educational computer games, digital forensics, biometrics, information security, social choice theory and queueing theory. They differ in their objects of inquiry, acceptable forms of evidence, risks, evaluation measures and possibilities for independent verification. This heterogeneity is analytically useful because a framework that remains meaningful across such settings must distinguish a small set of shared methodological obligations from domain-dependent validation methods.

The contribution of the present study is *design-oriented*: it synthesises relevant literature, establishes a structured problem definition, derives research objectives and specifies a defensible methodology for the subsequent construction and evaluation of the artefact.

3. Literature Review

3.1 AI agents and the scientific research process

Agentic systems are often characterised by their ability to decompose goals into tasks, select and call tools, use feedback to revise actions and maintain information across multiple steps. A recent methodological primer situates these capacities within scientific activities that include literature analysis, simulation and experimental orchestration [11]. A lifecycle survey similarly illustrates the range of emerging agents for hypothesis generation, experimentation, analysis and scientific communication, while documenting limitations in benchmarks and operational assurance [12]. The breadth of application means that an agent may operate as a literature assistant, an executable analysis component, a simulation controller or an autonomous experiment coordinator. These roles entail different obligations for validation and oversight.

It is important to distinguish an *agentic workflow* from ordinary use of a language model. A model that suggests code during interactive use is not necessarily a component on which the execution of a scientific analysis depends. Conversely, an agent with authority to change datasets, execute software, select experimental conditions or publish outputs becomes part of the evidential chain. Its inputs, tool actions, intermediate products and human approvals may be necessary to reconstruct how a claim was produced. The development of retrieval-augmented scientific assistants, such as the system studied by Asai and colleagues [1], illustrates why source grounding and verifiable citations are central even for apparently modest literature-related tasks.

The scientific status of an output consequently depends on its role. A proposed hypothesis, an extracted observation, a computed estimate and an independently corroborated finding are not interchangeable. A generative explanation may be useful for exploration but cannot, by itself, establish the correctness of an analytical result. A defensible framework should make these distinctions operational by requiring researchers to specify the intended epistemic status of each important output and the procedure by which that status may be upgraded to a validated claim.

3.2 Scientific validity, reproducibility and evaluation

Empirical studies of agents expose the limitations of evaluating scientific systems solely through task-completion rates. ScienceAgentBench contains 102 tasks derived from peer-reviewed research across several scientific disciplines. In the assessed configuration, its strongest agent independently completed 32.4% of the tasks within three attempts [4]. This figure refers to a particular benchmark and set of evaluated systems; it should not be extrapolated to every type of scientific work. FIRE-Bench, which evaluates rediscovery of documented research findings, reports substantial run-to-run variation and recurring weaknesses in the design, execution and interpretation of experiments [34]. Together, these studies establish a reason to separate code execution, task success and scientific inferential quality.

Attempts to automate larger parts of the research process have produced useful demonstrations but also document failures in experimental implementation, manuscript accuracy and methodological rigour [19]. Evaluation therefore needs an independent reference wherever one is available: analytical solutions, verified test datasets, adjudicated expert findings, established experimental procedures or reproducible simulations. Without a valid reference, measures such as perceived coherence, self-evaluation by the same model, or the production of a manuscript are weak substitutes for scientific corroboration.

Reproducibility introduces further complications. Model updates, non-deterministic inference, changing retrieval corpora, tool dependencies and underspecified prompts can prevent the exact reconstruction of an agent-assisted analysis. This does not imply that byte-identical results are always achievable. Rather, the objective may be to reproduce a specified scientific conclusion within predefined tolerances while retaining an auditable record of inputs, configurations, procedures, outputs and deviations. Research on provenance and evidence ontologies offers useful representational foundations for linking claims to the operations and evidence from which they were derived [6]. Evaluations of agent systems should additionally report resource cost and evaluation design, rather than interpreting isolated performance scores as sufficient evidence of practical value [16].

3.3 Responsibility, security and human oversight

Agent autonomy creates risks beyond ordinary analytical error. An agent may encounter adversarial content, disclose restricted information, execute a harmful tool action or continue a workflow after an assumption has become invalid. The NIST Generative AI Profile provides a risk-management reference for identifying and documenting potentially harmful outcomes [23]; the OWASP agentic security guidance addresses threats associated with access, tool use and untrusted instructions [26]. Controlled agent-security benchmarks such as AgentDojo and WASP provide empirical evidence that malicious instructions embedded in untrusted content can redirect tool-using agents, although attack success depends on the task and threat model [7, 10]. These sources support risk identification and control design but do not constitute evidence that a particular implementation is safe.

A defensible methodological treatment of autonomy requires explicit boundaries: which actions the agent may perform, which require approval, which data it may access, and which events require interruption or escalation. Oversight should be proportional to the consequences and reversibility of actions. For instance, allowing an agent to suggest alternative simulation settings entails a different risk than allowing it to modify original forensic evidence or autonomously respond to a production security incident. Human review is useful only if reviewers receive the information needed to evaluate the action, not merely a fluent post hoc explanation.

Regulatory and ethical requirements depend on the purpose, setting and data of each investigation. The European AI Act identifies obligations and exclusions whose applicability cannot be

decided from the label “research” alone [9]. Separate rules on personal data, research participation, forensic evidence and institutional ethics may apply. Hence, privacy, consent, data minimisation, scientific integrity and safety assessment must be incorporated into the research design before data collection or agent execution where relevant.

3.4 Evidence from the six research domains

3.4.1 Educational computer games

AI-supported game-based learning may enable adaptive feedback and personalised interactions, but perceived engagement is not a substitute for measured learning. A systematic review by Kaşarcı and Yurt [17] considers empirical evidence on AI-supported game-based learning and notes methodological challenges relating to outcomes and implementation. Agent evaluation should therefore separate learning gain, retention, motivation, time on task and safety, ideally using pre-specified comparisons with appropriate controls. When human participants are involved, assessment instruments and ethical safeguards are essential components of validity.

3.4.2 Digital forensics

Digital forensics requires preservation of original evidence, transparent analytical procedures and defensible attribution of findings. A recent review of large language models in forensics identifies possible benefits alongside unresolved issues of methodological validation and forensic integrity [5]. Practitioner-oriented research also demonstrates that actual adoption depends on issues of traceability, explanation, practical feasibility and legal presentation, not only model capability [32]. This domain provides a stringent test of whether agents can assist analytical work without altering evidential material or obscuring the relation between a claim and its source.

3.4.3 Biometrics

Biometric accuracy varies with demographic characteristics, acquisition conditions and task definition. Public evaluations from the US National Institute of Standards and Technology document demographic effects in face-recognition performance and the need for application-appropriate testing [24]. A methodologically adequate agent-assisted biometric study must distinguish verification from identification, quantify errors at defined operating thresholds and evaluate relevant subgroups and distribution shifts. An agent that summarises these results is not automatically qualified to make consequential identity decisions.

3.4.4 Information security

Security research faces a dual problem: agents may assist with threat detection and incident analysis while introducing additional attack surfaces through model inputs, tools and permissions. Risk-oriented guidance emphasises prompt injection, unsafe action execution and the dangers of excessive privilege [23, 26]. AgentDojo and WASP additionally demonstrate why evaluations must examine realistic multistep interactions rather than isolated model responses [7, 10]. Scientific evaluation consequently needs controlled environments, appropriate ground-truth incidents, realistic false-positive costs and adversarial testing. Detection quality and operational safety must be treated as separate dependent variables.

3.4.5 Social choice theory

Social choice research offers formal properties against which algorithmic behaviour can be tested. Holliday, Kristoffersen and Pacuit [15] study manipulation under limited information using learned strategies, demonstrating how strategic behaviour can be analysed computationally. Majumdar and

colleagues [20] investigate inconsistencies and biases in AI-generated voting preferences and their consequences for collective decisions. These findings motivate formal validation of voting rules, synthetic-profile experiments and sensitivity analyses. Preferences generated by a language model should not be assumed to represent observed human preferences without independent evidence.

3.4.6 Queueing theory

Machine learning has been applied to distributions and performance measures in queueing systems for which exact analysis may be difficult. Relevant examples include steady-state estimation for GI/GI/1 queues and tandem networks, as well as transient approximation for time-varying arrivals [2, 29, 30]. These studies illustrate the need for carefully chosen analytical, numerical and simulation baselines. In particular, stationary distributions must not be inferred where stability conditions are not met, and numerical accuracy needs to be assessed over the relevant parameter range rather than only on representative examples.

3.5 Synthesis and research gap

The reviewed literature already offers substantial methodological resources: evaluations of scientific agents, lifecycle descriptions, provenance concepts, domain-specific adoption frameworks, risk-management guidance and specialised validation methods. Consequently, an assertion that no relevant framework exists would be unjustified. The defensible research opportunity lies in examining whether these resources can be integrated into a single *research-methodological procedure* that (i) assigns agents explicit epistemic and operational roles, (ii) connects those roles to domain-appropriate scientific methods, (iii) records an auditable evidential chain, (iv) governs risk and autonomy, and (v) evaluates the incremental benefit of the procedure itself across heterogeneous settings.

The integration claim is a research proposition, not a demonstrated finding. It requires comparison with existing approaches and empirical evidence that the integration improves identifiable aspects of scientific practice. Because a framework may increase documentation cost or impose unsuitable requirements on certain disciplines, the limits of its usefulness are as important as successful demonstrations.

4. Methodology

4.1 Research design

Design Science Research (DSR) is adopted because the central question concerns the construction and evaluation of an artefact intended to address a recurring methodological problem. The process follows the six activities described by Peffers and colleagues: problem identification, definition of solution objectives, design and development, demonstration, evaluation and communication [27]. These activities are iterative rather than a guarantee that a first design will succeed. Gregor and Hevner [13] further distinguish the construction of a useful artefact from the articulation of transferable design knowledge; both are necessary for a substantive scientific contribution.

The unit of design is a modular methodological framework consisting of a process model, role and autonomy rules, scientific-method selection guidance, evidence and provenance requirements, risk controls, domain profiles and evaluation procedures. The unit of empirical evaluation is a *research workflow enacted under specified conditions*, not an isolated prompt or a collection of agent-generated texts. This distinction makes it possible to examine whether framework use changes scientific outcomes, their auditability or the resources required to produce them.

The present analysis represents the conceptual and problem-formulation stage of a DSR programme. Its literature synthesis is focused and selective rather than a completed systematic mapping or scoping review. No database-wide screening counts, stakeholder observations or measured effects are inferred. Subsequent evidence generation is specified below so that initial design assumptions can be corroborated, revised or rejected.

4.2 Problem identification and requirements elicitation

Problem identification aims to establish an evidence-based understanding of the scientific, methodological, technical and ethical challenges associated with the use of AI agents in multidisciplinary research. Given the diversity of research contexts and the evolving nature of agentic AI technologies, this process may benefit from combining complementary forms of inquiry. The selection of specific research methods should depend on the nature of the problems under investigation, the availability and quality of existing evidence, and the characteristics of the participating research domains.

One possible approach involves reviewing and synthesising relevant scientific literature to identify existing applications of AI agents in research workflows, methodological frameworks, evaluation practices and documented limitations. Depending on the scope and maturity of the available literature, this could take the form of a scoping review, systematic mapping study or another appropriate form of structured literature synthesis. Such an investigation could draw on general scholarly databases and discipline-specific sources while considering differences in the evidentiary status of peer-reviewed studies, preprints and technical guidance. If a formal scoping review is undertaken, established reporting guidance, such as PRISMA-ScR [31], may support transparency and reproducibility.

Insights from the literature may be complemented by empirical investigation of researchers' experiences, practices and expectations. Potential approaches include semi-structured interviews, focus groups, surveys, observations of research workflows or analyses of concrete research scenarios. These methods could help identify practical challenges that are insufficiently addressed in published literature, including difficulties in validating agent-generated outputs, managing unexpected behaviour, maintaining human oversight and balancing methodological rigour with operational efficiency. The choice of participants, sampling strategy and analytical procedures would need to reflect the diversity of research domains and the specific questions being investigated. Qualitative or quantitative approaches, or their combination, may be appropriate depending on the type of evidence sought.

Another potentially valuable line of inquiry concerns the comparative analysis of existing methodological frameworks, agent-development approaches and AI governance practices. Such an analysis could examine how different approaches address scientific validity, provenance, reproducibility, agent autonomy, ethical considerations, evaluation and applicability across research domains. The comparison may help identify limitations, complementarities and opportunities for methodological improvement. Importantly, the absence of an explicitly documented mechanism should not automatically be interpreted as evidence that the mechanism does not exist. Likewise, distinctions should be maintained between empirically demonstrated problems, reported practitioner experiences, conceptual concerns and anticipated challenges.

The findings of these investigations would provide a basis for deriving and refining the requirements of the proposed methodological framework. Requirements elicitation may involve established requirements engineering techniques, expert consultations, collaborative workshops, structured prioritisation or other suitable methods. For example, expert walkthroughs could help assess the relevance and feasibility of candidate requirements across research domains, while approaches such as the Delphi method might be considered where structured expert consensus would be beneficial. The appropriateness of these methods would depend on the characteristics of the requirements, the availability of relevant expertise and the degree of agreement needed.

An important consideration throughout this process is the relationship between identified problems and the requirements intended to address them. Establishing traceability between supporting evidence, formulated requirements and potential evaluation criteria could strengthen the methodological transparency of subsequent framework development. At the same time, requirements should remain open to refinement as new insights emerge, particularly where different research domains reveal conflicting needs or context-dependent constraints. The overall objective is therefore not to prescribe a fixed set of methodological procedures in advance, but to establish a sufficiently rigorous and adaptable basis for defining the functional, methodological and ethical requirements of the framework.

4.3 Artefact construction and iterative refinement

Artefact development should begin with a process model that specifies research tasks, scientific inference criteria, permitted agent actions, human checkpoints and required evidence records.

Participatory design sessions can test whether the model is comprehensible to domain experts. Early prototypes may instantiate individual mechanisms, such as evidence logging, an autonomy policy or a method-selection checklist, before these are integrated into a complete workflow.

Each design decision should be associated with a rationale, a problem or requirement it addresses, and a testable expected effect. Formative evaluation can employ scenario walkthroughs, expert inspection, test-case execution and failure injection. The framework should support revision when a mechanism proves unworkable in a domain, while preserving a trace of the changes made. Software implementations may support the framework, but the artefact is not defined by a particular programming language, model provider or tool stack.

4.4 Evaluation logic

Evaluation must operate on two levels. *Agent-level evaluation* establishes whether a particular agent, under defined conditions, produces scientifically acceptable results and respects its operational constraints. *Framework-level evaluation* establishes whether following the methodological framework improves the production, evaluation and documentation of those results compared with a credible alternative. This two-level logic follows the DSR requirement to show artefact utility and the FEDS distinction between formative/summative and artificial/naturalistic evaluation [13, 33].

Where feasible, three experimental conditions should be compared: (A) a conventional research workflow without an AI agent; (B) an ad hoc workflow using an AI agent without the methodological framework; and (C) an otherwise comparable agent-assisted workflow using the framework. The contrast between A and B addresses the effect of agent assistance, whereas the contrast between B and C addresses the *incremental utility of the framework*. The scientific question, task difficulty, reference datasets, models and computational resources should be kept comparable or modelled as covariates. In settings where all three conditions are not feasible, a paired, counterbalanced or carefully matched design may be used, with limitations made explicit.

Primary outcomes must be specified before testing. They may include independently verified scientific correctness, completeness of claim-to-evidence links, reproducibility of key findings, frequency of critical violations and time or cost per *valid* result. Secondary outcomes can include usability, researcher workload and perceived trust, but perceived trust must never substitute for correctness. Replication units should be defined at the appropriate level (research task, dataset, researcher or experimental run); repeated stochastic runs should not be misreported as independent participants. Analysis should prioritise effect sizes, confidence intervals, sensitivity to models and prompts, and the documentation of adverse or null outcomes. Sample sizes and practical acceptance thresholds should be justified by pilot results, precision requirements and, where relevant, power analysis rather than invented numbers.

4.5 Methodological and ethical safeguards

Construct validity requires scientific quality to be measured through domain standards rather than a universal subjective quality score. Internal validity requires adequate baselines, control of order effects, task contamination and researcher expectations. External validity is tested through variation among cases and explicit description of boundary conditions, not by assuming that six cases constitute a representative sample of all sciences. Reliability requires a sufficiently complete record of environments, software versions, model access conditions, data transformations and analysis decisions to enable independent scrutiny.

Ethics and security are design constraints, not only post hoc evaluation dimensions. Data and tool permissions should be minimised; agents must not modify original forensic material or execute consequential security actions outside controlled settings. Studies involving identifiable biometric information, user behaviour or educational participants require context-specific data-protection and

ethics assessment before implementation. Risk logs, authorization records and stopping conditions should be treated as evidence artefacts alongside numerical results. Applicable legal obligations should be assessed for each use rather than assumed to be uniform across the six domains [9, 23, 26].

5. Research Problems and Objectives

5.1 Cross-cutting problem structure

The literature suggests that difficulties in agent-assisted research arise from several distinct sources. Some concern the reliability of generated content; others concern the scientific design within which agents operate; still others stem from governance, reproducibility or the burden of maintaining a complex workflow. Table 5.1 presents a candidate taxonomy. Its categories are analytically derived from the reviewed literature and require further empirical corroboration. They should not be interpreted as a frequency ranking or as proof that all domains experience every problem equally.

Table 5.1: Candidate methodological problem taxonomy for AI-agent-supported research.

ID	Problem	Research consequence
P1	Epistemic reliability	Generated assertions, code or citations may be false despite persuasive presentation.
P2	Scientific validity	Successful execution does not establish that a hypothesis or inference is justified.
P3	Reproducibility	Changing models, prompts, datasets or software environments can prevent reconstruction of outcomes.
P4	Provenance and accountability	Data transformations, sources, tool actions and human approvals may be insufficiently documented.
P5	Evaluation inconsistency	Different baselines, outcome definitions and sampling units impede valid comparisons.
P6	Security, privacy and legal risk	Agents may expose sensitive information, exceed permissions or respond to adversarial inputs.
P7	Bias and distribution shift	Mean performance may obscure systematic errors across groups or conditions.
P8	Unbounded autonomy	Actions may exceed a defensible research role or proceed without suitable human review.
P9	Limited transferability	The validity criteria for evidence differ fundamentally across scientific domains.

ID	Problem	Research consequence
P10	Operational burden	Additional logging, orchestration and agent iterations may increase cost without improving scientific quality.

These ten categories generate two higher-order design problems. The first is *methodological integration*: individually useful controls may not form a coherent, practical procedure that connects the research question, agents, scientific methods, validation and governance. The second is *uncertain added value*: even a complete procedure is not justified unless its benefits outweigh its complexity and resource requirements. The first problem motivates artefact construction; the second demands an evaluation that measures the artefact itself.

5.2 The agent as a scientific instrument

A conceptual model is needed to avoid treating every AI output as the same kind of research object. An agent may function as an *advisor* suggesting literature or hypotheses, an *executor* performing a specified analysis, a *controller* coordinating tools or simulations, or an *evaluator* screening results against predetermined criteria. These roles may coexist, but each creates distinct questions about the origin and status of evidence. An agent's epistemic role is therefore determined by what it is authorised to contribute to the research process and how that contribution is independently assessed, not simply by its technical architecture.

The proposed scientific-instrument model contains seven related elements: (i) the research question and theoretical construct; (ii) the agent's assigned role; (iii) the evidence type and scientific method; (iv) the permitted actions and human decision points; (v) the provenance of inputs, transformations and outputs; (vi) the relevant uncertainty and validation procedure; and (vii) the criteria for accepting, rejecting or revising a proposed conclusion. This conceptualisation is a design conjecture whose completeness and practical value should be tested through expert analysis and workflow demonstrations.

For example, a language agent used to identify potential voting manipulations may propose a candidate strategy, but a mathematical solver or formal argument must establish the relevant property. An agent used in queueing research may orchestrate a discrete-event simulation, but the simulation design and reported uncertainty remain subject to standard statistical validation. In digital forensics, an agent-generated summary can help an examiner locate evidence, but the summary cannot replace preserved source material. These examples illustrate a common epistemic principle while preserving domain-specific standards of proof.

5.3 Research questions

The overarching question stated in Chapter 2 is developed into six subsidiary questions:

1. **RQ1:** Which methodological, epistemic, technical and ethical problems recur across AI-agent-supported research, and which remain domain-specific?
2. **RQ2:** How should the role of an AI agent as a scientific instrument be defined, and what requirements follow for method selection, evidence and accountability?
3. **RQ3:** Which modular process mechanisms can integrate scientific validation, bounded autonomy, provenance and ethical risk management without imposing inappropriate uniformity?
4. **RQ4:** Which domain-specific integrations or methods can be developed or improved, and how can their scientific validity and novelty be established?
5. **RQ5:** What incremental effects does framework-guided agent use have on scientific quality, reproducibility, safety, effort and cost compared with ad hoc agent use?

6. **RQ6:** Which design principles are transferable among the six domains, and under what conditions do their benefits or applicability fail?

5.4 Scientific research objectives

The research objectives follow the logic of DSR and link problem identification to artefact construction, empirical demonstration and generalisation. Their achievement requires the evidence described in Table 5.2; defining an objective is not equivalent to achieving it.

Table 5.2: Scientific objectives, methods and evidence of achievement.

ID	Objective	Method and required evidence
O1	Identify and classify cross-cutting and domain-specific problems of research with AI agents.	Protocolled literature mapping, researcher interviews and thematic coding; source-to-problem and domain-to-problem evidence matrix.
O2	Develop and critically assess a conceptual model of AI agents as scientific instruments and derive validated requirements.	Conceptual analysis, requirements engineering, expert walkthroughs and structured consensus; explicit acceptance criteria and reported disagreement.
O3	Design and iteratively refine a modular methodological framework for the responsible scientific integration of agents.	DSR design cycles, process modelling, prototyping and formative testing; documented design rationale and functional demonstrations.
O4	Develop or adapt domain-specific methods for integrating agents and establish their scientific validity.	Domain experiments, simulation, formal analysis and comparison with prior work; validated performance and justified claims of novelty or adaptation.
O5	Evaluate agent performance and the framework's independent contribution to research quality, reproducibility, safety and efficiency.	Controlled or matched workflow comparisons, FEDS-guided evaluation and mixed-method analysis; effect estimates, uncertainty and negative cases.
O6	Derive and assess transferable design principles, limits of applicability and reusable research resources.	Cross-case synthesis, external critique and independent reproduction where feasible; design principles with supporting evidence and boundary conditions.

The objectives distinguish *knowledge production* from *artefact capabilities*. O2 concerns the conceptual contribution; O3 concerns the methodological artefact; O4 concerns contributions that may emerge within individual domains; O5 tests usefulness rather than assuming it; and O6 formalises analytically transferable knowledge. The six objectives are complementary, not six independent claims of established novelty.

5.5 Requirements of the proposed framework

The operational requirements follow from the candidate problems and conceptual model. They are presented as an initial specification rather than a validated standard. A framework guided by these requirements would include: a research-task specification, a method-selection procedure, a role and autonomy policy, an evidence-provenance model, independent validation mechanisms, ethics/security controls, domain-specific profiles and a standardised evaluation dossier.

Table 5.3: Functional requirements and observable acceptance evidence.

Requirement	Expected capability	Acceptance evidence
R1: task specification	Define the research question, scientific claim, evidence type, permissible data and success criteria.	A case can be independently interpreted from its protocol.
R2: method selection	Relate the proposed claim to a scientifically justified experiment, proof, simulation or analysis.	Domain reviewers can evaluate and challenge the rationale.
R3: autonomy governance	Set agent roles, access rights, approval gates, stop conditions and escalation routes.	Scenario tests show permitted and blocked actions.
R4: provenance	Record the relevant data, versions, tool calls, transformations, outputs and decisions.	Independent auditors can reconstruct the evidence path.
R5: validation	Provide domain reference procedures, error measures and review requirements.	Key agent-produced claims receive independent checks.
R6: ethics and security	Document risk, data use, consent/approval needs and residual limitations.	Reviewable risk decisions and completed adversarial tests where applicable.
R7: domain adaptation	Maintain a common procedural core with specialist validation profiles.	Each domain uses the core without sacrificing its scientific standards.
R8: evaluation and openness	Report valid outcomes, comparators, uncertainty, cost, failures and reproducibility artefacts.	Comparable dossiers and lawful independent reuse.

A minimal workflow implementing these requirements would proceed from *question and evidence specification* to *method selection*, *agent-role and risk definition*, *controlled execution*, *independent validation*, and *reporting and reuse*. The framework should allow iteration when findings require a revised hypothesis or method; each revision should be recorded to avoid retroactively presenting exploratory decisions as preregistered tests. Requirements concerning traceability and safety may be mandatory across cases, while detailed measures and validation procedures must be specialised.

5.6 Domain-specific research and evaluation designs

The common framework must be instantiated using scientifically appropriate methods. Table 5.4 summarises defensible starting designs and the primary quantities that should be measured. These designs are propositions for empirical testing, not reported experiments.

Table 5.4: Candidate domain-specific validation designs.

Domain	Primary scientific method	Principal validation outcomes
Educational games	Randomised or quasi-experimental comparison with pre-test, post-test and delayed retention; appropriate validated scales.	Learning and retention gains, response safety, workload and subgroup differences.
Digital forensics	Controlled forensic scenarios; preserved source images/logs; blinded expert adjudication and independent reconstruction.	Correct findings, false attribution, source-to-claim traceability and evidence integrity.
Biometrics	Stratified evaluation by relevant population and acquisition conditions; attack/shift stress tests where appropriate.	FMR/FNMR or identification errors, subgroup gaps, uncertainty and attack robustness.

Domain	Primary scientific method	Principal validation outcomes
Information security	Isolated cyber-range experiments with held-out threats, reference detectors and adversarial scenarios.	Detection quality, alert burden, response delay and unauthorised agent actions.
Social choice theory	Formal examination of voting-rule properties, exact algorithms and controlled preference-profile simulations.	Manipulation success, optimality gap, theoretical consistency and model sensitivity.
Queueing theory	Analytical benchmark cases, replicated discrete-event simulation and numerical/ML comparison.	Distributional and parameter error, stability compliance, interval coverage and cost.

Several domain-specific precautions are essential. In educational studies, participant allocation and delayed testing should be chosen to distinguish novelty effects from learning. In forensics, original evidence must remain immutable; the agent's claim should link to an independently verifiable trace. In biometrics, thresholds and demographic subgroup analyses must be specified for the task actually studied. In security, high recall cannot compensate for unacceptable unauthorised actions. In social choice, a formally validated result is distinct from the behavioural realism of synthetic preferences. In queueing theory, stationary analysis requires stability, while transient estimates must be examined against time-dependent reference procedures. Such differences define the appropriate boundaries of the common framework rather than exceptions to be ignored.

5.7 Testable design propositions

The methodological framework is expected to be evaluated against a small number of propositions that make its intended effects observable. **DP1 (epistemic alignment)** proposes that explicitly linking each agent role to a scientific claim and a validation method improves the proportion of independently substantiated findings. **DP2 (bounded autonomy)** proposes that risk-proportionate permissions and human approval points reduce consequential unauthorised actions without rendering workflows unusable. **DP3 (evidence traceability)** proposes that structured provenance and claim–evidence links increase the rate at which independent researchers can audit or reconstruct results. **DP4 (modular validity)** proposes that a common methodological core combined with domain profiles improves consistency of research documentation while preserving domain-specific standards. **DP5 (added utility)** proposes that framework-guided workflows can improve validated research outcomes relative to ad hoc agent workflows at an acceptable marginal cost.

These propositions are conditional design hypotheses, not universal laws. DP1–DP3 can be examined through controlled workflow tasks, DP4 through cross-case comparison and expert assessment, and DP5 through comparative outcome and cost analysis. Each should admit counterexamples and specify relevant boundaries, including the possibility that conventional methods or simpler controls perform as well as the integrated framework.

6. Discussion

6.1 Scientific contribution and its evidentiary requirements

The proposed approach aims to contribute at three connected levels. The *conceptual* level defines the status of AI agents as scientific instruments within an accountable process of knowledge production. The *methodological* level develops a modular artefact that relates research questions, agent roles, scientific methods, provenance, autonomy and validation. The *empirical* level tests these arrangements through distinct domains and, where warranted by evidence, derives reusable design principles. Within DSR, an artefact becomes a stronger scientific contribution when its design rationale, utility and generalisable implications are explicit [13, 27].

The significance of the work therefore cannot be judged by whether an agent performs tasks quickly or whether six prototypes can be implemented. A scientifically meaningful result would show *which* methodological mechanisms improve *which* research outcomes, under *which* conditions. For example, provenance may improve independent auditability while increasing execution and documentation time. Risk-based approval gates may reduce harmful actions while affecting latency. Such trade-offs are not methodological failures if they are measured and provide evidence for defensible design boundaries.

The scientific-instrument perspective also guards against a category error: an agent can produce a useful suggestion without producing verified knowledge. This distinction is particularly important where outputs are fluent and apparently authoritative. Anchoring the evidential status of agent outputs in independent validation makes possible a consistent procedural language across domains, without requiring that all disciplines share the same inferential methods.

6.2 Position relative to existing approaches

Several relevant lines of work already cover important portions of the problem. Scientific-agent primers and lifecycle reviews describe architectures and research tasks [11, 12]. ScienceAgentBench and FIRE-Bench develop increasingly rigorous evaluation environments [4, 34]. Provenance research provides established conceptual tools for representing evidence links [6]. In digital forensics, a practitioner-driven framework explicitly addresses responsible adoption [32]. NIST

and OWASP supply risk-management and security resources, while agent-security benchmarks supply controlled empirical tests [7, 10, 23, 26]. These contributions must be treated as existing knowledge to build upon, not as gaps to be ignored.

A credible claim of novelty must consequently be more specific than the assertion that a general methodological framework is missing. The proposed contribution is the *joint design and empirical evaluation* of research-method selection, epistemic roles, provenance, bounded autonomy and two-level evaluation in a modular cross-domain procedure. Whether this combination is genuinely new, or simply a useful adaptation, should be resolved through a documented comparison of primary sources and evaluation of comparable alternatives. A negative finding on novelty would require revising the contribution claim, not hiding related work.

6.3 Limitations and threats to validity

The current literature synthesis is selective and cannot demonstrate exhaustive coverage of the field. The problem taxonomy reflects an analytical organisation of identified concerns; its completeness and prevalence across disciplines remain open questions. Differences in evidence type also constrain synthesis: peer-reviewed empirical results, methodological essays, normative guidance and design conjectures do not have equivalent evidentiary force. A future systematic mapping exercise is required to establish the coverage of related frameworks and the strength of support for individual requirements.

A second limitation concerns causal attribution. If framework-guided researchers receive extra training, more time or a different agent model, observed differences may not be caused by the framework. Evaluation should control these factors wherever possible and document remaining confounding. Unblinded human assessment, selection of favourable cases and repeated tuning on evaluation tasks can also overstate utility. Independent scoring and held-out tasks are useful safeguards, but their feasibility will vary by domain.

Third, transferability cannot be inferred from case count alone. The six settings are purposefully heterogeneous but do not statistically represent every research discipline. Cross-case synthesis should therefore use analytical generalisation: specify a design principle, the mechanism through which it may operate, the evidence observed in each case and the conditions under which it should not be transferred. Models, tools and legal obligations will also evolve, making version control and periodic review essential for long-term applicability.

Finally, increased procedural control may burden researchers. Excessive logging, duplicated approvals or rigid templates can reduce practical adoption without improving scientific quality. Cost, complexity and researcher workload must therefore enter the evaluation alongside scientific correctness, safety and reproducibility. A framework that adds no measurable value for a class of low-risk tasks should permit a simpler procedure rather than assume that more governance is always better.

6.4 Implications for research practice

For research teams, the practical implication is to begin with a scientific question and evidentiary standard, then select an agent role and an appropriate method of verification. This reverses the common tendency to choose a powerful agent first and infer suitable uses afterwards. Human oversight should focus on decisions that influence scientific validity or produce consequential effects, while routine checks can be automated when their own reliability is established.

For future research, the framework offers a structure for reporting not only successes but also failed replications, model sensitivity, unanticipated agent actions and circumstances in which agent use produces no improvement. Publicly available protocols, code, environment descriptions and non-sensitive test resources can support independent scrutiny, while restricted data must be handled

through controlled-access alternatives. Transparent reporting of negative cases is a condition for credible methodological learning, not a defect to be eliminated from the record.

Financira
Europska unija
NextGenerationEU

7. Conclusion

AI agents create opportunities to extend scientific workflows, but their capacity to perform computational tasks does not, by itself, establish the validity of the resulting knowledge. The literature identifies recurring concerns involving evidential reliability, reproducibility, provenance, autonomy, security, bias, evaluation consistency and transfer across disciplines. These concerns motivate a methodological approach that treats the agent as a bounded scientific instrument operating within a traceable and independently verifiable research process.

This study has articulated a design-oriented research problem, proposed a cross-cutting problem taxonomy and specified six scientific objectives together with observable requirements for a modular methodological framework. It has also established an evaluation strategy that distinguishes the scientific performance of agents from the incremental value of the framework through domain-appropriate methods and, where feasible, controlled comparisons. The proposed artefact and its design propositions remain to be validated; no empirical effectiveness claim follows from their formulation alone.

The decisive research outcome will be evidence that a coherent combination of scientific method selection, provenance, bounded autonomy, independent validation and responsible governance improves research practice under identifiable conditions. Equally informative will be cases in which this integration offers limited benefit or imposes excessive cost. Establishing both its utility and its boundaries is necessary for a credible, transferable contribution to the methodology of AI-assisted scientific research.

Bibliography

[1] Akari Asai, Jacqueline He, Rulin Shao, et al. “Synthesizing Scientific Literature with Retrieval-Augmented Language Models”. In: *Nature* 650 (2026), pages 857–863. DOI: 10.1038/s41586-025-10072-4 (cited on page 7).

[2] Opher Baron et al. “Supervised ML for Solving the GI/GI/1 Queue”. In: *INFORMS Journal on Computing* 36.3 (2024), pages 766–786. DOI: 10.1287/ijoc.2022.0263 (cited on pages 6, 10).

[3] A. Chan, C. Ezell, M. Kaufmann, et al. “Visibility into AI Agents”. In: *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Rio de Janeiro, 2024 (cited on page 2).

[4] Ziru Chen, Shijie Chen, Yuting Ning, et al. “ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery”. In: *Proceedings of the International Conference on Learning Representations*. 2025. URL: https://proceedings.iclr.cc/paper_files/paper/2025/hash/f12b4df26344f3be803c06b555252efe-Abstract-Conference.html (cited on pages 5, 8, 21).

[5] Maxim Chernyshev et al. “Large Language Models in Digital Forensics: Capabilities, Challenges and Future Directions”. In: *Forensic Science International: Digital Investigation* 56, 302043 (2026). DOI: 10.1016/j.fsidi.2025.302043 (cited on pages 6, 9).

[6] Tek Raj Chhetri et al. “Bridging the Scientific Knowledge Gap and Reproducibility: A Survey of Provenance, Assertion and Evidence Ontologies”. In: *Companion Proceedings of the ACM Web Conference 2025*. 2025, pages 924–928. DOI: 10.1145/3701716.3715483 (cited on pages 8, 21).

[7] Edoardo Debenedetti et al. “AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents”. In: *Advances in Neural Information Processing Systems*. Volume 37. 2024. DOI: 10.52202/079017-2636 (cited on pages 8, 9, 22).

[8] Z. Deng, Y. Guo, C. Han, et al. “AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways”. In: *ACM Computing Surveys* (2025) (cited on page 2).

[9] European Union. *Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence*. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689 (visited on 11/06/2026) (cited on pages 9, 14).

[10] Ivan Evtimov et al. "WASP: Benchmarking Web Agent Security against Prompt Injection Attacks". In: *Advances in Neural Information Processing Systems*. Volume 38. 2025. DOI: 10.52202/085713-0666 (cited on pages 8, 9, 22).

[11] B. Feng, P. Walther, L. Zhang, et al. "Agentic AI Systems for Scientific Discovery". In: *Nature Reviews Methods Primers* 6, 81 (2026). DOI: 10.1038/s43586-026-00534-y (cited on pages 5, 7, 21).

[12] Yifan Gao, Leyi Yu, and Haoyue Li. "Large Language Model Agents for Autonomous Scientific Research: A Lifecycle Survey of Methods, Systems, Benchmarks, and Risks". In: *Journal of Superintelligence* 1.2, 8 (2026). DOI: 10.3390/superintelligence1020008 (cited on pages 5, 7, 21).

[13] Shirley Gregor and Alan R. Hevner. "Positioning and Presenting Design Science Research for Maximum Impact". In: *MIS Quarterly* 37.2 (2013), pages 337–355. DOI: 10.25300/MISQ/2013/37.2.01 (cited on pages 11, 13, 21).

[14] I. Hettiarachchi. "Exploring Generative AI Agents: Architecture, Applications, and Challenges". In: *JAIGS* 8.1 (2025) (cited on page 1).

[15] Wesley H. Holliday, Alexander Kristoffersen, and Eric Pacuit. "Learning to Manipulate under Limited Information". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 39. 13. 2025, pages 13915–13925. DOI: 10.1609/aaai.v39i13.33522 (cited on page 9).

[16] S. Kapoor, B. Stroebl, et al. *AI Agents That Matter*. 2024. arXiv: 2407.01502. URL: https://arxiv.org/abs/2407.01502 (cited on pages 1, 8).

[17] İsmail Kaşarcı and Eyüp Yurt. "Effectiveness of AI-Supported Game-Based Learning: A Systematic Review of Outcomes, Challenges, and Future Directions". In: *Behavioral Sciences* 16.7, 1050 (2026). DOI: 10.3390/bs16071050 (cited on pages 6, 9).

[18] N. Kolt. "Governing AI Agents". In: *Notre Dame Law Review* (2025). Forthcoming, according to provided project bibliography (cited on page 2).

[19] Chris Lu, Cong Lu, Robert Tjarko Lange, et al. "Towards End-to-End Automation of AI Research". In: *Nature* 651 (2026), pages 914–919. DOI: 10.1038/s41586-026-10265-5 (cited on pages 5, 8).

[20] Srijoni Majumdar, Edith Elkind, and Evangelos Pournaras. "Generative AI Voting: Fair Collective Choice Is Resilient to LLM Biases and Inconsistencies". In: *EPJ Data Science* 15, 24 (2026). DOI: 10.1140/epjds/s13688-025-00612-3 (cited on page 10).

[21] T. Masterman et al. *The Landscape of Emerging AI Agent Architectures for Reasoning, Planning, and Tool Calling: A Survey*. 2024. arXiv: 2404.11584. URL: https://arxiv.org/abs/2404.11584 (cited on page 1).

[22] Lisa Messeri and M. J. Crockett. "Artificial Intelligence and Illusions of Understanding in Scientific Research". In: *Nature* 627 (2024), pages 49–58. DOI: 10.1038/s41586-024-07146-0 (cited on page 5).

[23] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1. 2024. DOI: 10.6028/NIST.AI.600-1 (cited on pages 8, 9, 14, 22).

[24] National Institute of Standards and Technology. *Face Recognition Technology Evaluation (FRTE): Demographic Effects*. Continuously updated evaluation resource. URL: https://pages.nist.gov/frvt/html/frvt_demographics.html (visited on 11/06/2026) (cited on page 9).

[25] L. Ning, Z. Liang, Z. Jiang, et al. *A Survey of WebAgents: Towards Next-Generation AI Agents for Web Automation with Large Foundation Models*. 2025. arXiv: 2503.23350. URL: https://arxiv.org/abs/2503.23350 (cited on page 1).

[26] OWASP GenAI Security Project. *Top 10 for Agentic Applications*. 2025. URL: https://genai.owasp.org/2025/12/09/owasp-genai-security-project-releases-top-10-risks-and-mitigations-for-agentic-ai-security/ (visited on 11/06/2026) (cited on pages 8, 9, 14, 22).

[27] Ken Peffers et al. "A Design Science Research Methodology for Information Systems Research". In: *Journal of Management Information Systems* 24.3 (2007), pages 45–77. DOI: 10.2753/MIS0742-1222240302 (cited on pages 11, 21).

[28] P. Putta, E. Mills, N. Garg, et al. *Agent Q: Advanced Reasoning and Learning for Autonomous AI Agents*. 2024. arXiv: 2408.07199. URL: https://arxiv.org/abs/2408.07199 (cited on page 1).

[29] Eliran Sherzer. "Computing the Steady-State Probabilities of the Number of Customers in the System of a Tandem Queueing System: A Machine Learning Approach". In: *European Journal of Operational Research* 326.1 (2025), pages 141–156. DOI: 10.1016/j.ejor.2025.04.040 (cited on page 10).

[30] Eliran Sherzer et al. "Approximating $G(t)/GI/1$ Queues with Deep Learning". In: *European Journal of Operational Research* 322.3 (2025), pages 889–907. DOI: 10.1016/j.ejor.2024.12.030 (cited on pages 6, 10).

[31] Andrea C. Tricco et al. "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation". In: *Annals of Internal Medicine* 169.7 (2018), pages 467–473. DOI: 10.7326/M18-0850 (cited on page 12).

[32] Maryna Veksler, Kemal Akkaya, and Selcuk Uluagac. "Practitioner-Driven Framework for AI Adoption in Digital Forensics". In: *Forensic Science International: Digital Investigation* 55, 302030 (2025). DOI: 10.1016/j.fsidi.2025.302030 (cited on pages 6, 9, 21).

[33] John Venable, Jan Pries-Heje, and Richard Baskerville. "FEDS: A Framework for Evaluation in Design Science Research". In: *European Journal of Information Systems* 25.1 (2016), pages 77–89. DOI: 10.1057/ejis.2014.36 (cited on page 13).

[34] Zhen Wang et al. "FIRE-Bench: Evaluating AI Agents on the Rediscovery of Scientific Insights". In: *Proceedings of the 43rd International Conference on Machine Learning*. Volume 306. Proceedings of Machine Learning Research. 2026, pages 124896–124929. URL: https://proceedings.mlr.press/v306/wang26d.html (cited on pages 5, 8, 21).