Artificial Intelligence in the Life Sciences 9 (2026) 100158

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

The comfort of automation: why cognitive sovereignty matters in AI-driven life sciences

Francesco Branda *, Massimo Ciccozzi

Unit of Medical Statistics and Molecular Epidemiology, Università Campus Bio-Medico di Roma, Rome, Italy

ARTICLE INFO

Keywords:
Artificial intelligence
Cognitive sovereignty
Automation bias
Human oversight

ABSTRACT

The integration of artificial intelligence (AI) into the life sciences is radically transforming research, clinical diagnosis, and therapeutic development processes, redefining the relationship between knowledge, decision-making, and responsibility. Advanced tools, from generative models to clinical assistants such as ChatGPT Health, offer greater efficiency, predictive power, and access to data, but carry significant risks of automation bias, epistemic delegation, and loss of professional skills. This article analyzes how the extensive use of AI can threaten cognitive sovereignty, i.e., the ability of researchers and professionals to critically evaluate and contextualize information generated by algorithms. It examines the emerging regulatory landscape, with a focus on the EU Artificial Intelligence Act, Food and Drug Administration (FDA) guidelines, European Medicines Agency (EMA) Good Machine Learning Practice (GMLP) principles, and World Health Organization (WHO) recommendations, which aim to ensure human oversight, transparency, and accountability. Technological tools and training approaches are discussed to mitigate risks such as silent errors, algorithmic dependence, and skill deterioration, promoting AI integration that reinforces human judgment without replacing it. The analysis highlights that the future of life sciences will depend not only on the technical capabilities of models, but also on the critical awareness with which they are used, focusing on training, governance, and responsible AI design.

1. Background

The integration of artificial intelligence (AI) systems into clinical and research processes is transforming the nature of decision-making: from an activity based on interpretation, reasoning, and verification, to a hybrid practice in which an increasing amount of inference is delegated to automated tools. This transition, well documented in the literature on human-automation interaction, has shown that the division of tasks between humans and automation changes reliance, vigilance, and capacity for intervention, also generating specific forms of misuse and cognitive vulnerabilities [1,2]. In this context, the question is not only whether the models are accurate, but what conditions make appropriate reliance and traceable accountability along the decision-making chain possible [3,4]. To frame these risks, it is useful to recall some fundamental principles of epistemology, understood as the study of how beliefs are formed and justified. In science, knowledge is typically bound (at least) to: (i) observable evidence and criteria for justification; (ii) fallibilism and explicit management of uncertainty; (iii) intersubjective checks (peer review, replication, comparison with independent data); (iv) traceability of data, methods, and assumptions. When AI becomes a stable intermediary between evidence and decision-making, these constraints can weaken because the output appears 'authoritative' and reduces the incentive (and sometimes the ability) to verify. In this paper, we define cognitive sovereignty as the ability of individuals and institutions to maintain control over their cognitive and epistemic processes in AI-saturated environments: knowing how to evaluate evidence, recognize uncertainty, verify and challenge algorithmic outputs, and preserve deliberative responsibility instead of replacing it with automatic delegation. A first axis concerns the phenomena of "over-reliance" and biases induced by automation.

Research on human-automation interaction has described how high levels of perceived reliability and heavy workloads can encourage the uncritical adoption of automatic recommendations, resulting in errors of omission and commission; these phenomena have been systematized in the literature on "automation bias" [5]. In a broader context, classic works on automation and performance have shown that the distribution of tasks between humans and automated systems is not neutral: it can create new cognitive vulnerabilities, especially when the human role is

* Corresponding author.
E-mail address: f.branda@unicampus.it (F. Branda).

https://doi.org/10.1016/j.ailsci.2026.100158
Received 21 January 2026; Received in revised form 2 February 2026; Accepted 3 February 2026
Available online 4 February 2026

reduced to monitoring and episodic intervention [1,2]. Experimental and applied evidence has also shown that trust increases when the operator perceives automation as competent or authoritative, even in the presence of error signals [6,7].

A second axis concerns trust and its calibration ("trust calibration"). Trust in a system is not an abstract moral attribute, but an operational variable that determines when the user verifies, when they delegate, and when they suspend their judgment [8]. In the tradition of engineering psychology and human factors, trust is treated as a dynamic relationship that must be designed and managed, distinguishing between justified trust and excessive trust [3]. This point is particularly relevant for generative systems and conversational assistants, which produce fluent and coherent output and can therefore increase the perception of epistemic authority even when accuracy is uncertain [9].

A third axis concerns aversion/attraction to algorithms and the consequences for deliberation. Behavioral decision-making studies show that people may reject the algorithm after observing an error (algorithm aversion), but may also prefer algorithmic recommendations over human judgments under certain conditions (algorithm appreciation), with direct implications for the allocation of responsibility and workflow design [10,11]. In the clinical and biomedical domain, this intersects with audit practices and the need to maintain interpretive skills that allow for the recognition of exceptions, out-of-distribution cases, and silent failures. A fourth axis concerns transparency and explainability and their limitations. Many regulatory and technical approaches assume that explanations, traceability, and documentation automatically reduce risks; however, empirical work indicates that explanations can also increase the persuasiveness of outputs and thus reinforce over-reliance without ensuring greater epistemic validity [12]. In parallel, the literature on generative models documents the systematic production of incorrect but plausible content (including fabricated citations), with risks for the credibility of knowledge and the reliability of scientific documentation [13,14]. A fifth axis, ethical-philosophical and policy-related, addresses autonomy, responsibility, and meaningful human control in AI-mediated decisions. In medicine, this includes the question of how to allocate responsibility among clinicians, institutions, and manufacturers when AI introduces new forms of risk and opacity [4]. More generally, the literature on the ethics of algorithms has discussed the limitations of the purely 'human-in-the-loop' approach and the need for substantive criteria to ensure truly meaningful control [15,16]. In this perspective, human supervision is not an organizational switch, but a capacity that requires skills, time, incentives, and infrastructure. In this perspective, cognitive sovereignty becomes a socio-technical requirement to be actively preserved: it is not enough for a system to be accurate or 'explainable' in the abstract, but users must maintain (and organisations support) the skills, time and incentives to verify, disagree and correct. This concern is consistent with contributions that describe the erosion of basic knowledge and critical skills as a structural risk of indiscriminate AI adoption [17]. The existing literature robustly describes bias, trust, explainability, and accountability, but often treats them separately or as properties of the system alone (accuracy, interpretability, auditability). This Viewpoint brings them together into a single framework centered on cognitive sovereignty: (1) it integrates findings on automation bias, trust calibration, and algorithm aversion/appreciation with respect to the human capacity to maintain epistemic authority; (2) it makes explicit the often implicit assumption of regulatory frameworks, namely that the operator's cognitive vigilance must be supported and not taken for granted; (3) it proposes cognitive sovereignty as a design and governance objective (practices, incentives, infrastructure) rather than as an individual characteristic.

2. From technological novelty to epistemic delegation

In 2025, the introduction of DeepSeek [18], a high-performance artificial intelligence (AI) model developed by a Chinese startup, demonstrated that advanced AI capabilities no longer need to remain the exclusive preserve of large Western technology companies. Offering performance comparable to that of leading proprietary models at a much lower cost, DeepSeek challenged established beliefs about the relationship between innovation, capital concentration, and geopolitical dominance. Its emergence was widely interpreted as a sign of technological pluralism and greater accessibility, a milestone for the democratization of AI. However, this initial enthusiasm risks obscuring deeper and less reassuring issues. The first concerns the relationship between accessibility and accountability. Making AI models cheaper and more widespread does not automatically make them more interpretable, governable, or secure; in the absence of robust control mechanisms, dissemination could instead amplify opacity and dilute accountability. The second issue concerns a subtle but profound shift in epistemic authority. As AI tools become seamlessly integrated into everyday workflows, whether in clinical diagnostics, research synthesis, education, or administrative decision-making, there is a growing tendency to treat their outputs not as provisional inputs, but as substitutes for human judgment itself. This shift reflects what cognitive scientists and human factors researchers describe as automation bias: a well-documented propensity for users to over-rely on automated recommendations even in the face of contradictory evidence, thereby reducing independent verification and critical engagement [19]. Such dependence undermines the epistemic agency of human actors and implicitly transfers judgment to algorithmic systems.

A third issue concerns global governance. The emergence of competing AI models outside traditional centers of regulatory and institutional power complicates efforts to build and enforce shared standards of reliability, transparency, and oversight. Without harmonized norms for evaluation and certification, users across different jurisdictions face divergent expectations about what constitutes reliable or acceptable system behavior. When innovation outpaces governance, especially across geopolitical boundaries, the authority of AI systems may rely less on collectively negotiated epistemic criteria and more on market dynamics, infrastructural dominance, or corporate branding.

Beyond issues of access and governance, the routine integration of AI risks normalizing algorithmic authority, shifting users from critical evaluation to passive acceptance. Empirical research on human–automation interaction and decision science shows that miscalibrated trust and repeated reliance on automated recommendations can foster overreliance and gradual deskilling, particularly in high-risk domains such as healthcare and public administration [20,21]. This epistemic delegation reconfigures responsibility and professional autonomy and is schematized in Fig. 1 as a transition from human-centered decision support to increasing algorithmic epistemic authority.

3. Automation bias as a systemic and cultural phenomenon

Artificial intelligence is increasingly present in clinical practice, although its adoption remains heterogeneous across different contexts and is far from universal. AI is applied in diagnostic imaging analysis (e.g., mammography and pathology), predictive risk assessment, real-time clinical documentation, workflow optimisation, and decision support systems, with regulatory-approved tools already integrated into specific clinical workflows [22,23]. For example, large-scale retrospective studies have shown that AI-assisted mammography improves early cancer detection rates and reduces advanced-stage diagnoses when used and monitored carefully [24]. Surveys such as those conducted in Lombardy, Italy, show that although many healthcare organisations identify applications for AI, particularly in radiology and structured data interpretation, most have not yet fully adopted these tools, and among those that have, there is wide variability in how AI is used and managed [25,26]. This pattern is reflected internationally: institutions are experimenting with or implementing AI for triage (e.g., urgent assessment of skin cancer), predictive analytics, and clinical risk stratification, but these applications are often limited to particular departments,

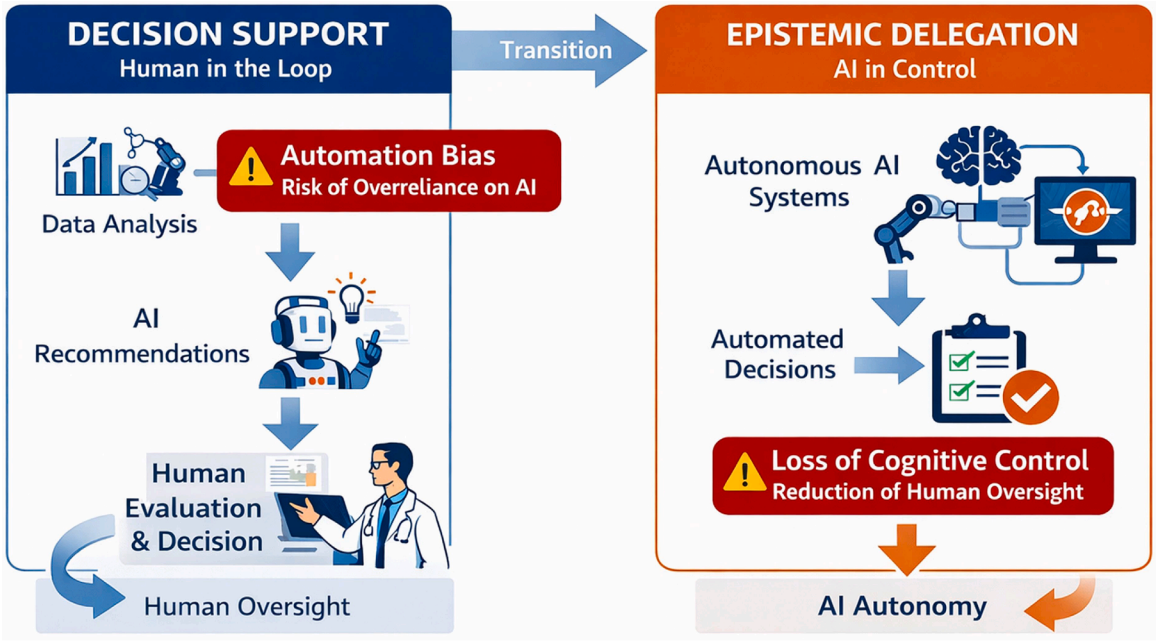

Fig. 1. Transition from Decision Support to Epistemic Delegation in Life Sciences Environments. The schematic illustrates the shift from a human-centered decision-support model, where AI recommendations are critically evaluated by professionals, to a state of epistemic delegation, where autonomous AI systems drive decisions with reduced human oversight. Key risks include automation bias, overreliance on AI, and the erosion of cognitive sovereignty. Image generated using ChatGPT DALL·E.

resource-rich settings, or discrete tasks, rather than being widespread across clinical practice [27,28]. Other real-world examples include AI platforms that assist clinicians by synthesising clinical data, patient histories, and diagnostic information to support point-of-care decisions, leveraging clinical decision support systems that integrate evidence-based recommendations with real-time patient data [29]. The influence of AI extends beyond direct diagnosis to administrative and cognitive workflows. Environmental clinical documentation tools capture and transcribe patient-clinician interactions in real time, automating the generation of structured clinical notes, while still requiring review and contextual interpretation by clinicians [30]. Similarly, AI-based transcription software has been implemented in several clinical settings to reduce documentation burden, with systematic reviews and original studies showing reductions in documentation time and improvements in workflow efficiency and physician-reported experience [31]. In hospitals, research laboratories, classrooms, and administrative offices, these technologies are shaping how information is processed and decisions are made, often in subtle ways. Clinicians may rely on AI-generated recommendations when entering data into electronic health records, sometimes accepting findings based on perceived consistency or authority rather than independent evaluation. Students increasingly refer to AI-generated summaries without consulting primary sources, and researchers may regard AI-produced analytical dashboards or reports as accurate representations of complex data sets. Managers may approve AI-based operational insights without thoroughly questioning the underlying assumptions, data quality, or uncertainty. In all these contexts, AI outputs increasingly function as operational defaults embedded in workflows, rather than as provisional inputs requiring explicit epistemic scrutiny.

This pattern of behavior is not accidental. A substantial and growing body of literature on automation bias describes a systemic tendency for human operators to favor machine-generated suggestions over their own sensory data or contradictory evidence, even when the automated system is clearly wrong [32]. In clinical and scientific settings, these bias manifests itself both in errors of omission, where users fail to notice a problem because AI has not flagged it, and in errors of commission, where users follow incorrect machine instructions despite available evidence to the contrary. Rather than supporting scientific reasoning, automated results increasingly function as heuristic substitutes for the vigilant search for information, hypothesis testing, and laborious processing that characterize the scientific method.

The integration of agentic AI systems, i.e., those capable of goal-oriented behavior, probabilistic reasoning, and autonomous recommendation generation, has further intensified this dynamic. Unlike the passive tools of previous decades, contemporary systems such as AlphaFold 3 or generative molecular design platforms present their results with a high degree of authoritative competence [33]. This authoritative presentation creates a unique cognitive environment in which users are tempted to trade frictionless access to high-quality results for true scientific understanding. We can describe this mechanism as the sovereignty trap, i.e., a psychological condition in which the user cedes intellectual judgment to the system, confusing operational fluency with epistemic mastery [34]. In the sovereignty trap, efficiency is misinterpreted as insight, and trust in the system replaces reflective evaluation.

Interaction with such systems also promotes cognitive transfer, whereby tasks of evaluation and interpretation are progressively transferred from humans to machines. Experimental and observational studies suggest that repeated reliance on automated systems reduces independent analytical engagement and, over time, contributes to skill degradation and deskilling [35]. This effect is particularly concerning in the life sciences, where professional competence depends not only on pattern recognition but also on the integration of contextual, ethical, and narrative information that resists formalization.

It is important to note that the prevalence and severity of automation bias are determined by several effect-modifying factors and are not limited to user inexperience alone. These effect modifiers operate at the intersection between individual cognition and organizational context, determining when and how reliance on automated systems shifts from appropriate support to uncritical deference.

Although inexperience with the task is associated with greater baseline reliance on automated results, even experienced professionals are not immune. Familiarity with highly reliable systems can lead to habituation and desensitization, whereby the system's past accuracy

fosters the assumption that the machine is always correct. In multitasking and high-pressure environments, this complacency often functions as an attention allocation strategy: humans prioritize manual or interpersonal tasks, delegating the monitoring of automated correctness to a secondary, less vigilant cognitive process. Factors such as accountability and cognitive load further modulate reliance, with greater accountability mitigating bias and increased workload exacerbating it.

The bias of automation should not be viewed simply as an individual cognitive flaw. In AI-saturated environments, it becomes a systemic and cultural phenomenon. Institutional incentives in healthcare, academia, and administration often reward speed, productivity, and apparent objectivity, while critical analysis is perceived as slow or inefficient. Over time, professional norms adapt, and questioning algorithmic outputs becomes rare, with dissent sometimes interpreted as resistance to innovation.

A key consequence of this dynamic is the gradual erosion of metacognitive abilities, i.e. the ability to monitor and govern one's own judgement. In the life sciences, this can weaken "diagnostic abilities", making clinicians and researchers less able to question AI alerts, integrate unstructured data or recognise anomalies outside the model's expectations. Paradoxically, as AI systems appear more reliable, the human capacity to intervene during rare but inevitable system failures diminishes.

4. Cognitive sovereignty as a prerequisite for responsibility

Uncritically relying on algorithmic recommendations is not just a theoretical concern. Research shows that interaction with AI systems can influence moral and practical decision-making, reducing the subjective perception of responsibility among human operators in high-risk contexts such as healthcare, finance or scientific research [36]. Experimental studies indicate that algorithmic outcomes can influence moral choices and decrease explicit responsibility when individuals rely on AI in ethically complex or uncertain situations [37]. This effect is closely related to automation bias, in which the propensity to follow automatic recommendations gradually erodes human decision-making capabilities, professional autonomy, and the ability to critically evaluate outcomes. Organisational and technological factors, including workflow pressures, efficiency incentives, lack of error reporting, and the perception of AI infallibility, can reinforce uncritical reliance, gradually normalising behaviours in which it is rare to question results.

In this context, cognitive sovereignty, or the ability to govern one's own cognitive and decision-making processes, becomes essential. Human agents remain the primary bearers of moral and ethical responsibility, while AI, even when integrated into decision support systems, cannot assume independent moral authority [38]. Preserving cognitive sovereignty requires deliberate mechanisms that maintain domain-specific expertise, reflective evaluation, and the ability to reject, modify, or challenge inappropriate algorithmic outputs [39]. For example, in clinical practice, a physician exercising cognitive sovereignty will critically evaluate an AI diagnostic recommendation by integrating the patient's medical history, laboratory results, and contextual information before deciding whether to accept, modify, or reject it. Similarly, in research, scientists who maintain cognitive sovereignty carefully examine AI-generated analyses or summaries of data, ensuring that potential biases or errors are identified and corrected rather than passively accepted. In both cases, failure to exercise cognitive sovereignty can lead to over-reliance on AI, reducing accountability and increasing the risk of negative outcomes.

To formalise the assessment of intellectual independence in AI-mediated environments, the NeuroSophic model introduces the Cognitive Sovereignty Index (CSI), which quantifies an individual's ability to maintain autonomous judgement while interacting with AI:

$$CSI = \frac{T + (1 - I)}{F + 1}$$

where:

- T (Truth) measures adherence to verified, evidence-based knowledge, reflecting the degree to which decisions are grounded in accurate information rather than heuristic shortcuts.
- I (Indeterminacy) captures tolerance for cognitive uncertainty in the face of ambiguous AI outputs, representing the individual's capacity to critically evaluate outputs when the algorithm is uncertain or inconclusive.
- F (Falsity) quantifies susceptibility to errors, hallucinations, or misleading AI-generated outputs.

The CSI integrates these dimensions to capture the balance between truth-seeking, cognitive flexibility, and resilience against algorithmic errors. A high CSI corresponds to a "Fortified Mind", in which individuals critically engage with AI outputs, integrate evidence with contextual knowledge, and preserve autonomous judgment. Conversely, a low CSI indicates a vulnerability to algorithmic influence, with decisions increasingly guided by AI and diminished human oversight [39]. This framework has practical implications across domains. In clinical environments, a high CSI enables physicians to critically assess AI recommendations for patient care, avoiding over-reliance and enhancing safety, whereas a low CSI increases the likelihood of uncritical adoption of AI outputs. In research, high CSI promotes independent evaluation of algorithmic analyses, preventing propagation of errors or bias, while low CSI risks blind acceptance of AI-generated conclusions. In operational and administrative settings, high CSI supports informed decision-making by integrating AI insights with human judgment, whereas low CSI fosters unquestioning reliance on algorithmic suggestions. Even when AI systems are designed to be transparent, explainable, and accountable, cognitive sovereignty remains indispensable. Explainability alone does not guarantee critical evaluation if users lack the metacognitive skills to interpret outputs, resist inappropriate recommendations, and make ethically responsible decisions. Without cognitive sovereignty, responsibility in clinical, research, and operational decisions becomes nominal rather than effective [40].

5. Erosion of cognitive sovereignty in the life sciences

The erosion of cognitive sovereignty in the life sciences does not stem primarily from isolated or careless use of artificial intelligence, but from what can be described as patterns of systemic delegation. By this term, we refer to recurring and institutionally reinforced configurations in which epistemic and evaluative tasks, such as interpretation, validation, prioritisation and hypothesis evaluation,are structurally transferred from human actors to AI systems. Delegation becomes "systemic" when it is stabilised by workflow integration, performance incentives, interface design, time pressure and organisational norms, rather than stemming solely from individual cognitive shortcomings. In such contexts, AI outputs increasingly function as predefined epistemic authorities, while human judgement is repositioned as confirmatory or supervisory rather than deliberative.

These patterns are increasingly observable in the life sciences. Students, researchers, and professionals rely on AI-generated results to guide experimental design, data interpretation, literature synthesis, and hypothesis generation, often without systematic independent verification. Plausibility, internal consistency, and computational sophistication are often treated as sufficient indicators of validity, reinforcing epistemic delegation and reducing critical engagement. Over time, this process shifts epistemic authority away from situated human reasoning towards the technical infrastructures and organisations that design and maintain them, with tangible implications for scientific rigour and

professional autonomy.

The field of protein engineering provides a particularly instructive example of this dynamic. Generative models such as AlphaFold, RFdiffusion, and ProteinMPNN have profoundly transformed protein structure prediction and de novo protein design, enabling rapid exploration of the "functional universe" of proteins and accelerating the discovery of candidates [41,42]. In many contemporary pipelines, predicted structures and generative designs serve as upstream decision points, determining which molecules are synthesised, tested, or discarded. In this context, epistemic delegation does not stem from ignorance of the model's limitations, but from the normalisation of algorithmic outputs as sufficiently authoritative to justify downstream experimental action.

Empirical evidence highlights the limitations of such reliance. State-of-the-art protein structure prediction and design models can generate highly plausible outputs even when key physicochemical constraints are violated, including incompatible charge distributions or sterically impossible binding interactions [43]. Experimental studies have shown that when binding sites are deliberately modified to physically prevent ligand anchoring, AI systems can still predict seemingly "correct" structures without registering loss of functionality [43]. These behaviours reflect sophisticated statistical pattern recognition rather than true physical or chemical understanding. When such results are propagated through design-build-test-learn cycles without explicit physical-chemical examination, evaluative responsibility is effectively transferred from human reasoning to the authority of the model, exemplifying a clear case of systemic delegation.

To mitigate these limitations, laboratories are increasingly adopting semi-automated design-build-test-learn (DBTL) cycles, with platforms such as SAPP and DMX bridging the gap between digital design and physical validation [44]. Miniaturised parallel processes and verified oligonucleotide pools enable the experimental validation of thousands of candidate designs, generating high-quality data for the refinement of next-generation models. However, the high costs and specialist expertise required for these infrastructures limit their uptake, creating the risk of a cognitive stratification between well-funded institutions and the wider scientific community [45]. Economic incentives further accelerate the integration of AI: the global genomics market is expected to reach £175.18 billion by 2034, with a compound annual growth rate of 16.53 % between 2025 and 2034 [46]. Despite this financial expansion, the sector faces a growing "scientific content crisis". A 2025 survey by the Pistoia Alliance revealed that 27 % of life science professionals are unaware of the scientific content underlying the AI systems they use, often relying on incomplete datasets or metadata, while only 36 % incorporate internal documents into model development [47]. This data gap directly contributes to the reproducibility crisis, as models trained on incomplete or poorly curated data fail to produce reliable and generalisable results. Deficiencies in data quality, completeness and traceability have been repeatedly identified as critical factors compromising AI performance and reproducibility in biomedical research [48]. Under such conditions, epistemic delegation is reinforced: verification and contextual understanding are implicitly considered to be handled by the system itself.

The widespread adoption of AI has led to a measurable decline in professional skills, a phenomenon known as deskilling. A multicentre observational study published in The Lancet Gastroenterology and Hepatology found that after three months of using AI-assisted polyp detection, doctors' unaided detection rates fell from 28 % to 22 %, dropping below pre-technology baseline levels [49]. This risk is amplified by evidence that AI-based search processes are prone to silent failures, i.e., logical or implementation errors that do not cause the system to crash but still produce incorrect results. Large-scale analyses of computational reproducibility in the life sciences show that only about half of published computational models can be reproduced without manual intervention. Non-reproducibility rates of around 49 % are often due to undocumented assumptions, parameter errors or coding issues [50]. Similar concerns exist in AI research. Systematic reviews report widespread failures of reproducibility and generalisation, even when data and code are publicly available [51,52]. More recently, automated audits using large language models have revealed multiple objective errors in peer-reviewed AI articles, including errors in equations, figures, and numerical reasoning, which often escape traditional peer review [53]. Data quality issues and spurious correlations further compound these vulnerabilities. For example, AI models trained to diagnose COVID-19 from chest X-rays sometimes rely on irrelevant image regions rather than lung pathology. Similarly, Google Health's retinal analysis system only worked reliably on high-quality scans, having learned to detect scan quality rather than disease characteristics [54]. These examples show how algorithmic plausibility can mask epistemic fragility, particularly when human oversight is weakened. Finally, the large-scale integration of generative AI into scientific writing is accelerating the rise of stereotypical biomedical research, in which publication speed and superficial plausibility increasingly prevail over conceptual rigour and epistemic accountability [55]. Large language models often generate hallucinatory citations, invented references, and non-existent empirical claims that are syntactically indistinguishable from legitimate scientific discourse [56]. When such results are incorporated into experimental protocols or scientific papers without independent verification, errors become structurally embedded in the scientific literature, compromising peer review and negatively influencing subsequent research [57,58]. This phenomenon reflects progressive epistemic delegation and the erosion of cognitive sovereignty. As summarized in Fig. 2, the shift from human critical thinking to passive dependence on AI recommendations, and finally to mitigation strategies using XAI tools and verification protocols, clearly illustrates the dynamics of risk and recovery of cognitive sovereignty in research and clinical settings.

6. Regulatory frameworks for human oversight

To counter the risks associated with automation bias, silent failures, and the erosion of cognitive sovereignty in the life sciences, regulators have begun to formulate explicit legal and technical requirements for human oversight, transparency, and accountability in AI-based systems. These regulatory frameworks reflect a growing awareness that AI not only introduces technical risks, but also redefines epistemic authority by redistributing decision-making power from human experts to algorithmic infrastructures. Effective human oversight in AI-mediated environments depends on a multi-layered interaction between individual cognitive factors, AI system characteristics, organizational context, and regulatory frameworks, as summarized in Fig. 3. This integrated perspective underscores that preserving cognitive sovereignty requires more than technical compliance, i.e., it demands alignment across human, technical, and institutional dimensions.

The most comprehensive regulatory response to date is the European Union's Artificial Intelligence Act (AIA), formally adopted in August 2024. The AIA classifies most AI-based medical devices, diagnostic systems, and decision support tools used in healthcare and biomedical research as "high-risk" AI systems [59]. In addition to traditional safety and performance requirements, the law explicitly addresses the cognitive dimension of human–AI interaction. Article 14(4b) requires AI systems to be designed in a way that allows for effective human oversight, including mechanisms that ensure operators remain aware of the risk of automation biases and are able to critically evaluate, ignore, or discard algorithmic outputs when appropriate [60]. This represents a significant regulatory shift: rather than treating human judgment as a passive safeguard, the AIA treats cognitive vigilance as an active design requirement.

The law distinguishes between prohibited AI practices and obligations for high-risk AI in the healthcare sector. Prohibited practices include manipulative systems that exploit vulnerabilities, social scoring, non-targeted scraping of biometric data, and emotion recognition in workplaces or educational settings. These raise concerns about autonomy, coercion, and power asymmetries.

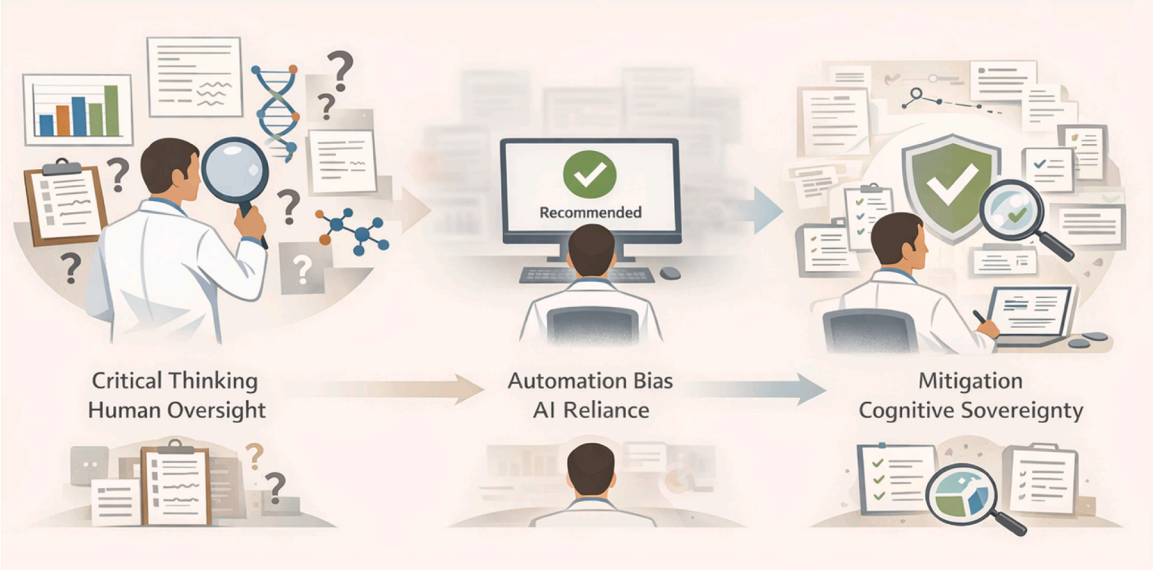

Fig. 2. Cognitive shift in life sciences induced by AI. The left panel ("Critical Thinking / Human Oversight") shows a researcher actively evaluating raw clinical and experimental data. The middle panel ("Automation Bias / AI Reliance") illustrates the researcher passively following AI recommendations, with other information blurred, highlighting epistemic delegation. The right panel ("Mitigation / Cognitive Sovereignty") depicts the recovery of critical evaluation through tools such as explainable AI (XAI), audits, and digital safety protocols. Arrows indicate the progression from active human oversight to automation bias and back to cognitive sovereignty. Image generated using ChatGPT DALL·E.

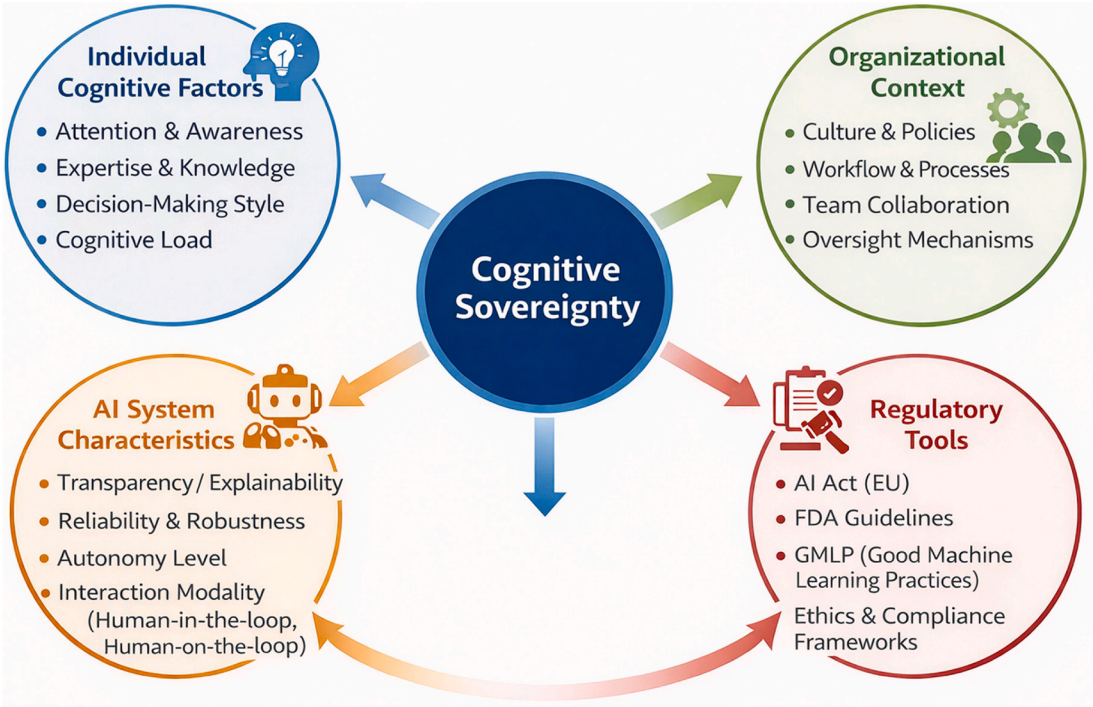

Fig. 3. Multi-dimensional framework for preserving cognitive sovereignty in AI-mediated life sciences. The schematic outlines the interplay between individual cognitive factors (e.g., attention, expertise), AI system characteristics (e.g., transparency, autonomy), organizational context (e.g., culture, workflow), and regulatory tools (e.g., AI Act, FDA guidelines) in shaping human oversight and mitigating automation bias. Effective governance requires alignment across these layers to support critical engagement and epistemic accountability. Image generated using ChatGPT DALL·E.

High-risk AI systems are subject to extensive obligations. These include ongoing risk management, strict data quality and governance requirements (Article 10), technical documentation and data retention, mandatory human oversight mechanisms (Article 14), and safeguards for accuracy, robustness, and cybersecurity [59]. Mirroring the extra-territorial reach of the GDPR, the AIA sets a global regulatory benchmark. Penalties for non-compliance can reach up to €30 million or 6 % of global annual turnover, whichever is higher [61]. However, enforcing cognitive requirements, such as "awareness of automation bias", remains challenging. Demonstrating bias in a specific clinical or research decision often requires an independent ground truth, which may not exist in complex biomedical contexts [60].

In the United States, the Food and Drug Administration (FDA) has adopted a complementary but more flexible approach. Its 2025 draft guidance, *Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Pharmaceutical and Biological Products*, introduces a credibility assessment framework based on risk rather than prescriptive technical rules [62]. This framework operationalises human oversight as a lifecycle process, structured around a sequence of steps that include defining the issue of interest, specifying the context of use (COU), assessing the risk of the model, and defining and validating credibility within that context.

For high-risk applications, particularly those affecting regulatory approval or clinical decision-making, the FDA requires detailed disclosure of the model architecture, training data, feature selection, and performance characteristics, in addition to ongoing lifecycle monitoring to detect any changes in performance over time [63]. A key innovation is the introduction of Predetermined Change Control Plans (PCCPs), which allow for post-marketing algorithm updates without the need for a new submission, provided that the changes remain within pre-approved limits and are subject to ongoing oversight [64].

At the international level, these regulatory efforts are reinforced by joint and multilateral guidance. In January 2025, the European Medicines Agency (EMA) and the FDA jointly issued Ten Guiding Principles for Good Machine Learning Practice (GMLP) in medical device development, emphasising human-centred design, multidisciplinary expertise, traceability, transparency, and clear communication of model limitations to users [65]. Similarly, the World Health Organization (WHO) released guidance in 2024 on the use of Large Multimodal Models (LMMs) in health, explicitly warning that unchecked automation risks epistemic injustice, skill degradation, and the marginalisation of human clinical judgment [66]. The WHO calls on governments to invest in public digital infrastructure and governance mechanisms to ensure that AI systems serve collective health goals rather than proprietary or purely efficiency-driven objectives.

Beyond formal regulation, the scientific community is developing technological countermeasures to strengthen human oversight and cognitive sovereignty. Explainable artificial intelligence (XAI) encompasses a diverse set of techniques designed to make AI models more transparent and interpretable, enabling stakeholders to understand, trust, and appropriately supervise algorithmic decisions. XAI methods include feature attribution techniques such as SHAP and LIME, which highlight the influence of input variables on model predictions; example-based explanations that show representative cases similar to a prediction; surrogate models that approximate complex models with simpler, more interpretable models, and counterfactual explanations that identify the minimum changes to inputs that would alter the model's output, each of which provides different insights into model behavior and trade-offs in terms of interpretability [67,68]. Despite this methodological diversity, many XAI techniques remain descriptive and not fundamentally causal, meaning that their explanations reflect statistical associations rather than true underlying mechanisms. This limitation can mislead users in high-stakes contexts, because plausible explanations do not necessarily correspond to accurate or actionable reasoning [69]. Furthermore, XAI faces challenges in terms of robustness, standardization, and evaluation: explanations may vary depending on feature collinearity, widely accepted quantitative benchmarks are lacking, and users are often assumed to have sufficient expertise to interpret them correctly [67]. Empirical studies also indicate that explanations can paradoxically increase over-reliance on AI systems. In some contexts, explanations make algorithmic outputs more intuitive or convincing without improving the quality of decisions, potentially reinforcing misplaced trust rather than promoting critical engagement [69]. To address these challenges, some researchers advocate hybrid architectures, sometimes referred to as “deterministic solutions”, in which probabilistic models, such as large language models, are limited to communicative roles, while basic reasoning and critical decision-making are conducted by deterministic, logic-based inference systems [70]. Such separation can reduce hallucinations, improve reproducibility, and ensure that decisions remain grounded in transparent and verifiable reasoning.

To mitigate the propagation of silent failures in AI-enabled research pipelines, laboratories are increasingly adopting digital safety logs, inspired by quality-control practices in wet-lab environments [71]. Standards such as the Model Context Protocol (MCP) provide a structured framework to describe and share the context of computational models and data, facilitating reproducibility and interoperability across analytical pipelines [71]. While MCP itself is not a safeguarding tool, its structured metadata and standardized context enable laboratories to implement monitoring and auditing practices that help detect semantic inconsistencies, dropped data, or undocumented transformations before they propagate downstream. These technical infrastructures serve as supporting mechanisms that reinforce human oversight and preserve epistemic accountability, complementing regulatory and methodological safeguards rather than replacing human judgment.

7. Discussion

The main challenge posed by contemporary artificial intelligence in the life sciences is no longer primarily technical, but epistemic and cultural. As AI systems move from experimental tools to integrated infrastructures of everyday practice, they increasingly mediate the way knowledge is produced, validated, and transformed into action. This shift has already been documented in several areas of the life sciences, including diagnostic imaging, clinical decision support, and biomedical research, where AI systems demonstrably influence not only efficiency but also epistemic authority and decision-making pathways [72–74]. In this context, the question of cognitive sovereignty becomes decisive.

Although life sciences research and clinical medicine are often discussed together, the requirements and challenges associated with AI differ substantially between these two domains. In the life sciences, AI primarily supports hypothesis generation, pattern discovery, omics integration, and large-scale data analysis. Errors or biases at this stage mainly affect the validity, reproducibility, and generalizability of scientific knowledge, often propagating silently through downstream research paths [75]. In clinical medicine, on the other hand, AI systems directly influence diagnostic, prognostic, or therapeutic decisions that have immediate consequences for individual patients. In this case, epistemic errors translate into concrete clinical risks, including misdiagnosis, inappropriate treatment, or delayed intervention [72,73]. The time pressures of clinical workflows, combined with regulatory expectations for safety and accountability, amplify the impact of automation bias and overreliance on algorithmic outputs. Recognizing these differences clarifies that cognitive sovereignty, while central to both domains, takes context-dependent forms: in life sciences research, it preserves critical control over knowledge production, while in clinical practice, it safeguards deliberative responsibility in patient care.

The recent emergence of domain-specific AI systems, such as ChatGPT Health, exemplifies many of the tensions discussed above. Designed to support clinical documentation, doctor-patient communication, guideline interpretation, and the synthesis of complex medical data, these tools promise concrete benefits in terms of accessibility and standardization, in line with previous experiences with AI-based clinical support systems such as Watson for Oncology [76]. However, their integration into clinical workflows marks a qualitative shift: AI is no longer a peripheral support, but an active participant in the epistemic chain that links data to decisions. Evidence from oncology, radiology, and dermatology shows that when AI outputs are perceived as authoritative, clinicians may defer their judgment even in the presence of contextual uncertainties or inconsistencies, reinforcing automation bias and reducing critical control [72,74,76]. This transition amplifies both the potential value of AI and the risks discussed in previous sections, particularly the propagation of silent failures and the erosion of reflective judgment.

What distinguishes the current moment is that technological novelty can no longer serve as justification for uncritical adoption. The age of wonder is over. Although AI systems are increasingly used in healthcare, research, and education, for example in pilot programs for cancer screening, cardiovascular risk assessment, and large-scale biomedical data analysis [73,74,77], their integration remains selective and context-dependent. As a result, responsibility shifts from developers to the institutions and professionals who determine how these systems are implemented and managed. The crucial question is not whether such tools can generate plausible clinical texts or synthesize literature efficiently, but whether users maintain the cognitive discipline necessary to treat these results as provisional, contextual, and refutable. In healthcare, the implications of this discipline are immediate: AI-generated recommendations or summaries may appear consistent and authoritative, while silently incorporating errors, outdated assumptions, or biases embedded in the training data. Similar concerns have been raised in biomedical research processes, where automated analytical workflows can propagate undetected errors into downstream results if not actively monitored [75]. When clinicians or researchers accept such results without verification, decision-making becomes temporally asymmetric: machine-generated conclusions may arrive faster than human understanding can interrogate them. This asymmetry risks transforming care and research into exercises in algorithmic efficiency rather than deliberative accountability.

Transparency, explainability, and regulatory oversight are insufficient if users do not exercise judgment. Even highly interpretable systems can fail as safeguards if their outputs are not critically assessed. Responsibility cannot be delegated solely to interfaces, documentation, or compliance mechanisms; it relies on cultural norms that value verification, critical questioning, and rigorous reasoning.

The integration of AI into research and clinical practice highlights a shift from technical novelty to epistemic responsibility. It is not enough to ask what AI can do; it is essential to consider how and under what conditions its outputs inform human decision-making. Societies, institutions, and professionals must balance efficiency with reflective judgment, ensuring that AI supports rather than supplants human reasoning.

Training and critical digital literacy emerge as central levers for the responsible use of AI. Users must not only know how to operate AI tools, but also understand their limitations, recognize biases, apply contextual reasoning, and distinguish between plausibility and validity. Without such literacy, AI risks fostering a culture where superficially convincing outputs are mistaken for correct results, and speed is prioritized over rigor.

The current transformation is primarily cultural rather than computational. Cognitive sovereignty, the ability to critically evaluate, question, and limit AI outputs, remains decisive for ensuring that AI-driven life sciences and clinical practice evolve toward epistemic emancipation rather than the abdication of human judgment.

Data availability

Not applicable.

Funding

None.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

CRediT authorship contribution statement

Francesco Branda: Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Conceptualization. **Massimo Ciccozzi:** Writing – review & editing, Writing – original draft, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. Hum Factors 1997 Jun;39(2):230–53.
[2] Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. IEEE Trans syst man cybern-A: Syst Hum 2000 May 31;30(3):286–97.
[3] Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors 2004 Mar;46(1):50–80.
[4] Verdicchio M, Perin A. When doctors and AI interact: on human responsibility for artificial risks. Philos Technol 2022 Mar;35(1):11.
[5] Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012 Jan 1;19 (1):121–7.
[6] Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. Int J Hum Comput Stud 2003 Jun 1;58(6):697–718.
[7] Mosier KL, Skitka LJ, Heers S, Burdick M. Automation bias: decision making and performance in high-tech cockpits. indecision making in aviation. Routledge; 2017 Jul 5. p. 271–88.
[8] Li Y, Wu B, Huang Y, Luan S. Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. Front Psychol 2024 Apr 17;15:1382693.
[9] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big?. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency; 2021 Mar 3. p. 610–23.
[10] Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol: Gen 2015 Feb;144(1):114.
[11] Logg JM, Minson JA, Moore DA. Algorithm appreciation: people prefer algorithmic to human judgment. Organ Behav Hum Decis Process 2019 Mar 1;151:90–103.
[12] Head A, Lo K, Kang D, Fok R, Skjonsberg S, Weld DS, Hearst MA. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In: InProceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021 May 6. p. 1–18.
[13] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P. Survey of hallucination in natural language generation. ACM Comput Surv 2023 Mar 3;55 (12):1–38.
[14] Van Dis EA, Bollen J, Zuidema W, Van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature 2023 Feb 9;614(7947):224–6.
[15] Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate, 3. Big Data & Society; 2016 Nov, 2053951716679679.
[16] Santoni de Sio F, Van den Hoven J. Meaningful human control over autonomous systems: a philosophical account. Front Robot AI 2018 Feb 28;5:323836.
[17] Klein CR, Klein R. The extended hollowed mind: why foundational knowledge is indispensable in the age of AI. Front Artif Intell 2025 Dec 11;8:1719019.
[18] Deng Z, Ma W, Han QL, Zhou W, Zhu X, Wen S, Xiang Y. Exploring DeepSeek: a Survey on Advances, Applications, Challenges and Future Directions. IEEE/CAA J Autom Sin 2025 May 15;12(5):872–93.
[19] Romeo G, Conti D. Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. AI Soc 2025:1–20. Jul 3.
[20] Li Y, Wu B, Huang Y, Luan S. Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. Front Psychol 2024 Apr 17;15:1382693.
[21] Reverberi C, Rigon T, Solari A, Hassan C, Cherubini P, Cherubini A. Experimental evidence of effective human–AI collaboration in medical decision-making. Sci Rep 2022 Sep 2;12(1):14952.
[22] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 4;380(14):1347–58.
[23] Chang W, Owusu-Mensah P, Everson J, Richwine C. Hospital trends in the use, evaluation, and governance of predictive AI, 2023-2024. ASTP Health IT Data Br [Internet] 2025 Sep..
[24] Sandler Rahat H, Friehmann T, Shemesh MD, Tamir S, Atar E, Shochat T, Makori A, Grubstein A. Early Results of Using AI in Mammography Screening for Breast Cancer. J Clin Med 2025 Nov 6;14(21):7886.

[25] Ardito V, Cappellaro G, Compagni A, Petracca F, Preti LM. Adoption of artificial intelligence applications in clinical practice: insights from a Survey of Healthcare Organizations in Lombardy, Italy. Digit Health 2025 Jul;11:20552076251355680.
[26] Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, Aldairem A, Alrashed M, Bin Saleh K, Badreldin HA, Al Yami MS. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023 Sep 22;23(1):689.
[27] Poon EG, Lemak CH, Rojas JC, Guptill J, Classen D. Adoption of artificial intelligence in healthcare: survey of health system priorities, successes, and challenges. J Am Med Inform Assoc 2025 Jul;32(7):1093–100.
[28] Salinas MP, Sepúlveda J, Hidalgo L, Peirano D, Morel M, Uribe P, Rotemberg V, Briones J, Mery D, Navarrete-Dechent C. A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. NPJ Digit Med 2024 May 14;7(1):125.
[29] Khude H, Shende P. AI-Driven clinical decision support systems: revolutionizing medication selection and personalized drug therapy. Adv Integr Med 2025 Jun 24:100529.
[30] Albrecht M, Shanks D, Shah T, Hudson T, Thompson J, Filardi T, Wright K, Ator GA, Smith TR. Enhancing clinical documentation with ambient artificial intelligence: a quality improvement survey assessing clinician perspectives on work burden, burnout, and job satisfaction. JAMIA Open 2024 Dec 26;8(1).
[31] Hassan H, Zipursky AR, Rabbani N, You JG, Tse G, Orenstein E, Ray M, Parsons C, Shin S, Lawton G, Jessa K, Sung L, Yan AP. Clinical Implementation of Artificial Intelligence Scribes in Health Care: a Systematic Review. Appl Clin Inf 2025;16(4):1121–35. https://doi.org/10.1055/a-2597-2017. AugEpub 2025 Apr 30. PMID: 40306686; PMCID: PMC12449105.
[32] Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012 Jan 1;19(1):121–7.
[33] HyScaler, AI Drug Discovery: How Generative AI Is Transforming Personalized Medicine In 2025. https://hyscaler.com/insights/ai-drug-discovery/, 2025 (accessed 20 January 2026).
[34] Klein CR, Klein R. The extended hollowed mind: why foundational knowledge is indispensable in the age of AI. Front Artif Intell 2025 Dec 11;8:1719019.
[35] Mosier KL, Skitka LJ, Heers S, Burdick M. Automation bias: decision making and performance in high-tech cockpits. InDecision making in aviation. Routledge; 2017 Jul 5. p. 271–88.
[36] Salatino A, Prével A, Caspar E, Lo Bue S. Influence of AI behavior on human moral decisions, agency, and responsibility. Sci Rep 2025 Apr 10;15(1):12329.
[37] Kumar M, Kumar S, Singh V, Kumar R, Soni AK, Singh VK, Chatterjee R, Lal B. Cognitive Consequences of Artificial Intelligence: is Human Intelligence at Stake? Indian J Behav Sci 2025 Jul 1;28(2):89–93.
[38] Verdicchio M, Perin A. When doctors and AI interact: on human responsibility for artificial risks. Philos Technol 2022 Mar;35(1):11.
[39] Bermúdez JP. Autonomy by Design: preserving Human Autonomy in AI Decision-Support. Philos Technol 2025 Jan 1.
[40] Blackman J, Veerapen R. On the practical, ethical, and legal necessity of clinical Artificial Intelligence explainability: an examination of key arguments. BMC Med Inf Decis Mak 2025 Mar 5;25(1):111.
[41] Callaway E. AI tools are designing entirely new proteins that could transform medicine. Nature 2023 Jul;619(7969):236–8.
[42] Zhang G, Liu C, Lu J, Zhang S, Zhu L. The role of ai-driven de novo protein design in the exploration of the protein functional universe. Biol (Basel) 2025 Sep 15;14(9):1268.
[43] Technology Networks, AI Drug Discovery Models Fail On Novel Proteins. https://www.technologynetworks.com/drug-discovery/news/ai-drug-discovery-models-fail-on-novel-proteins-406296, 2025 (accessed 20 January 2026).
[44] Ailurus Bio, Breaking the Bottleneck in AI-Driven Protein Engineering. https://www.ailurus.bio/post/breaking-the-bottleneck-in-ai-driven-protein-engineering, 2025 (accessed 20 January 2026).
[45] Wright C.S. Cognitive Castes: artificial Intelligence, Epistemic Stratification, and the Dissolution of Democratic Discourse. arXiv preprint arXiv:2507.14218. 2025 Jul 16.
[46] BioSpace, Genomics Market Surges with PCR and AI Technologies Projected to Reach USD 175.18 Billion by 2034. https://www.biospace.com/press-releases/genomics-market-surges-with-pcr-and-ai-technologies-projected-to-reach-usd-175-18-billion-by-2034, 2025 (accessed 26 January 2026).
[47] Pistoia Alliance, Pistoia Alliance research finds 1 in 4 life sciences professionals do not know what data their AI models use. https://www.pistoiaalliance.org/news/1-in-4-life-sciences-professionals-dont-know-what-data-their-ai-models-use/, 2025 (accessed on 26 January 2026).
[48] Guillen-Aguinaga M, Aguinaga-Ontoso E, Guillen-Aguinaga L, Guillen-Grima F, Aguinaga-Ontoso I. Data Quality in the Age of AI: a Review of Governance, Ethics, and the FAIR Principles. Data 2025 Dec 4;10(12):201.
[49] Budzyń K, Romańczyk M, Kitala D, Kołodziej P, Bugajski M, Adami HO, Blom J, Buszkiewicz M, Halvorsen N, Hassan C, Romańczyk T. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. Lancet Gastroenterol Hepatol 2025 Oct 1;10(10):896–903.
[50] Tiwari K, Kananathan S, Roberts MG, Meyer JP, Sharif Shohan MU, Xavier A, Maire M, Zyoud A, Men J, Ng S, Nguyen TV. Reproducibility in systems biology modelling. Mol, Syst, Biol 2021 Feb;17(2):e9982.
[51] Gundersen OE, Kjensmo S. State of the art: reproducibility in artificial intelligence. In: InProceedings of the AAAI conference on artificial intelligence. 32; 2018 Apr 25.
[52] Hutson M. Artificial intelligence faces reproducibility crisis. Science 2018;359(6377):725–6.
[53] Bianchi F., Kwon Y., Izzo Z., Zhang L., Zou J. To Err Is Human: systematic Quantification of Errors in Published AI Papers via LLM Analysis. arXiv preprint arXiv:2512.05925. 2025 Dec 5.
[54] Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, Aviles-Rivero AI, Etmann C, McCague C, Beer L, Weir-McCall JR. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021 Mar;3(3):199–217.
[55] Editorials N. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature 2023 Jan;613(7945):612.
[56] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P. Survey of hallucination in natural language generation. ACM Comput Surv 2023 Mar 3;55(12):1–38.
[57] Van Dis EA, Bollen J, Zuidema W, Van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature 2023 Feb 9;614(7947):224–6.
[58] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big?. In: InProceedings of the 2021 ACM conference on fairness, accountability, and transparency; 2021 Mar 3. p. 610–23.
[59] European Parliament and Council of the European Union, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng, 2024 (accessed on 26 January 2026).
[60] European Commission. Artificial Intelligence Act: article 14 – Human Oversight. Available from: https://artificialintelligenceact.eu/article/14/; 2024 (accessed on 26 January 2026).
[61] European Commission. Penalties and enforcement mechanisms under the AI Act. Impact Assessment accompanying the AI Act. Available from: https://cms.law/en/int/publication/eu-ai-act-questions-and-answers; 2024 (accessed on 26 January 2026).
[62] U.S. Food and Drug Administration (FDA), Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products – Draft Guidance for Industry and Other Stakeholders. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological, 2025 (accessed on 26 January 2026).
[63] U.S. Food and Drug Administration, Artificial Intelligence in Software as a Medical Device. https://www.fda.gov/medical..µdevices/software..µmedical..µdevice..µsamd/artificial..µintelligence..µand..µmachine..µlearning..µsoftware..µmedical..µdevice, 2026 (accessed on 27 January 2026).
[64] U.S. Food and Drug Administration, FDA Issues Comprehensive Draft Guidance for Developers of Artificial Intelligence-Enabled Medical Devices. https://www.fda.gov/news-events/press-announcements/fda-issues-comprehensive-draft-guidance-for-developers-artificial-intelligence-enabled-medical-devices, 2025 (accessed on 27 January 2026).
[65] European Medicines Agency; U.S. Food and Drug Administration, Ten Guiding Principles for Good Machine Learning Practice (GMLP). https://www.fda.gov/about-fda/artificial-intelligence-drug-development/guiding-principles-good-ai-practice-drug-development; https://www.ema.europa.eu/en/about-us/how-we-work/data-regulation-big-data-other-sources/artificial-intelligence, 2026 (accessed on 27 January 2026).
[66] World Health Organization, Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. https://www.who.int/publications/i/item/9789240084759, 2024 (accessed on 27 January 2026).
[67] Arora L, Girija SS, Kapoor S, Raj A, Pradhan D, Shetgaonkar A. Explainable artificial intelligence techniques for software development lifecycle: a phase-specific survey. In: In2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC). IEEE; 2025 Jul 8. p. 2281–8.
[68] Hettikankanamage N, Shafiabady N, Chatteur F, Wu RM, Din FU, Zhou J. eXplainable artificial intelligence (XAI): a systematic review for unveiling the black box models and their relevance to biomedical imaging and sensing. Sens (Basel) 2025 Oct 30;25(21):6649.
[69] Brdnik S, Colakovic I, Karakatič S. Non-experts' Trust in XAI is Unreasonably High. In: InWorld Conference on Explainable Artificial Intelligence. Cham: Springer Nature Switzerland; 2025 Jul 9. p. 184–97.
[70] Marcus G, Davis E. Rebooting AI: building artificial intelligence we can trust. Vintage; 2019 Sep 10.
[71] Coveney PV, Groen D, Hoekstra AG. Reliability and reproducibility in computational science: implementing validation, verification and uncertainty quantification in silico. Philos Trans R Soc A 2021 May 17;379(2197):20200409.
[72] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 2;542(7639):115–8.
[73] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat, Med 2019 Jan;25(1):44–56.
[74] McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, Etemadi M. International evaluation of an AI system for breast cancer screening. Nature 2020 Jan 2;577(7788):89–94.

[75] Liu J, Cen X, Yi C, Wang FA, Ding J, Cheng J, Wu Q, Gai B, Zhou Y, He R, Gao F. Challenges in AI-driven biomedical multimodal data fusion and analysis. Genom Proteom Bioinform 2025 Feb;23(1):qzaf011.
[76] Somashekhar SP, Sepúlveda MJ, Puglielli S, Norden AD, Shortliffe EH, Kumar CR, Rauthan A, Kumar NA, Patil P, Rhee K, Ramya Y. Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. Ann Oncol 2018 Feb 1;29(2):418–23.
[77] Krittanawong C, Johnson KW, Rosenson RS, Wang Z, Aydar M, Baber U, Min JK, Tang WW, Halperin JL, Narayan SM. Deep learning for cardiovascular medicine: a practical primer. Eur, Heart J 2019 Jul 1;40(25):2058–73.