OpenCódice Working Paper OC-WP-2026-001

The Agentic R&D Decalogue: Ten Principles for Human-Agent Collaboration in Scientific Research

Francisco-Javier Rodrigo-Ginés
fran@opencodice.org

March 21, 2026 • **DOI:** 10.5281/zenodo.19151516

Abstract

Artificial intelligence agents are becoming active participants in scientific research. They draft papers, run experiments, review literature, and manage submissions. Yet the research community lacks a shared set of principles for how this collaboration should work. This paper proposes a decalogue for agentic R&D: ten principles for human-agent collaboration that preserve scientific integrity, accelerate discovery, and distribute the benefits of AI-assisted research equitably. The principles are grounded in an empirical analysis of research workflows, a framework for trustworthy AI oversight, and practical experience structuring projects for agent collaboration.

Keywords: Agentic AI, Scientific Research, Human-Agent Collaboration, Research Integrity, Open Science, Decalogue

Highlights

- A decalogue: ten actionable principles for human-agent collaboration in research
- A vision for agentic R&D that prioritises integrity over speed
- Grounded in empirical analysis of research workflows across 68 tasks
- Actionable commitments for researchers, institutions, and publishers

1 Preamble

Science is humanity's most reliable method for understanding the world. It works not because individual scientists are infallible, but because the process, peer review, replication, critical debate, cumulative evidence, corrects for individual error over time. Every element of the scientific process exists to serve this self-correcting function. As Popper [1959] argued, science progresses through conjectures and refutations; as Merton [1973] showed, its norms of universalism, communalism, disinterestedness, and organised scepticism are what distinguish it from other forms of knowledge production.

Throughout history, transformative technologies have reshaped how science is practised without changing what science is. The printing press enabled the dissemination of scientific knowledge beyond monastic walls, making peer scrutiny possible at scale. Statistical methods in the twentieth century allowed researchers to draw inferences from noisy data, giving rise to evidence-based medicine and the social sciences as we know them. The internet and digital repositories gave rise to the open science movement [Fecher and Friesike, 2014], democratising access to publications, data, and code. Each transition brought both genuine acceleration and legitimate concerns about quality and integrity. AI agents represent the next such transition: tools that do not merely store or transmit scientific knowledge but actively participate in producing it.

AI agents are entering the research process at an accelerating pace. Language models draft manuscripts. Autonomous systems design and run experiments. Multi-agent workflows coordinate entire research pipelines from literature review to submission. An analysis of the scientific lifecycle reveals at least 65 distinct research tasks across nine phases, and early assessments suggest that 57% have high or very high AI automation potential. Recent surveys confirm that researchers across disciplines are already experimenting with these tools, often without institutional guidance or shared norms [Liang et al., 2024].

This is an opportunity and a risk.

The opportunity is liberation: freeing researchers from the mechanical, repetitive, and administrative tasks that consume their time, so that they can invest in the creative, critical, and interpersonal work that drives discovery. A researcher who spends less time formatting papers and more time thinking about hypotheses produces better science. Early evidence suggests that AI tools can reduce the time spent on routine tasks by 30–50%, potentially redirecting hundreds of hours per year toward substantive intellectual work [Noy and Zhang, 2023].

The risk is erosion: the gradual hollowing out of the human judgment, critical thinking, and epistemic responsibility that make science trustworthy. A research pipeline that produces papers faster is not better if those papers are shallower, less rigorous, or less honest. The *reproducibility crisis* [Baker, 2016] has already shown that quantity without rigour undermines public trust in science; adding AI acceleration without safeguards could deepen this problem. There is also the subtler risk of epistemic homogenisation: if all researchers use similar AI tools trained on similar data, the diversity of perspectives and approaches that drives scientific creativity may narrow. Science benefits from disagreement, heterodox thinking, and the occasional contrarian; a research ecosystem in which AI agents steer every literature review toward the same consensus risks losing precisely the intellectual friction that produces breakthroughs.

This paper proposes a third path: **agentic R&D**, a model of human-agent collaboration that captures the opportunity while guarding against the risk. We distil it into a decalogue: ten principles grounded in an empirical analysis of research workflows, a framework for trustworthy AI oversight, and practical experience structuring projects for agent collaboration.

What does agentic R&D look like in practice? Consider three scenarios. First, a doctoral student uses an AI agent to conduct a systematic literature search across five databases, de-duplicate results, and produce structured summaries of 200 papers; the student then reads the summaries, identifies the research gap, and formulates the hypothesis. Second, a research team

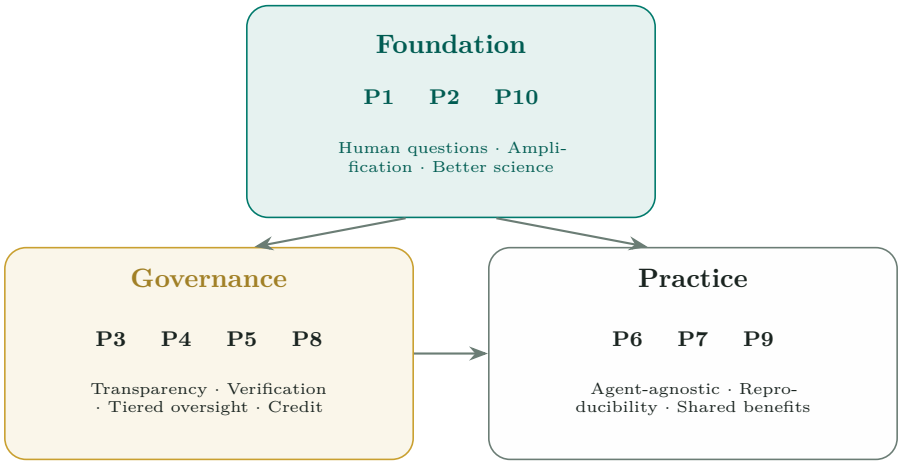

Figure 1: Relationship between the ten principles of agentic R&D. Foundation principles (top) inform both Governance and Practice clusters. The arrow from Governance to Practice indicates that governance mechanisms constrain and enable practical implementation. Principles within each cluster are mutually reinforcing rather than hierarchical.

configures an agent to run a battery of statistical analyses on experimental data, producing tables and visualisations; the team interprets the results, identifies anomalies, and decides which findings merit further investigation. Third, a principal investigator uses an agent to draft an initial manuscript from structured notes and experimental logs; the PI rewrites the argument, sharpens the claims, and takes full responsibility for the final text. In each case, the agent handles volume and mechanics while the human provides direction, judgment, and accountability.

We present this decalogue not as dogma but as a practical framework: ten principles that we believe should guide the integration of AI agents into scientific research, offered for debate, refinement, and adoption by the research community. We recognise that these principles will need to evolve as the technology matures and as the research community develops collective experience with agentic workflows. What we offer here is a starting point, informed by both principle and practice, for a conversation that the scientific community must have before default practices calcify into norms.

2 Ten Principles for Agentic R&D

The principles that follow are not arbitrary; they emerge from the intersection of three bodies of knowledge: the philosophy and sociology of science, the emerging field of trustworthy AI governance, and the practical experience of research teams already working with AI agents. They are intended to be actionable rather than aspirational, specific enough to guide daily decisions while general enough to apply across disciplines and institutional contexts.

The ten principles are organised into three clusters (Figure 1). The *Foundation* principles (1, 2, 10) establish the core philosophy: humans lead, agents amplify, and the ultimate goal is better science. The *Governance* principles (3, 4, 5, 8) specify the accountability mechanisms: transparency, verification, proportional oversight, and honest credit. The *Practice* principles (6, 7, 9) address the infrastructure: agent-agnostic design, reproducibility, and equitable access. Foundation informs both Governance and Practice; Governance constrains and enables Practice. The arrows in the figure represent dependencies: without a clear foundation, governance mechanisms lack purpose; without governance, practical infrastructure lacks accountability.

Foundation Principles

The foundation principles define the philosophical bedrock of agentic R&D. They answer the most basic questions: who leads the research process (Principle 1), what role agents play (Principle 2), and what the ultimate measure of success is (Principle 10). These three principles are non-negotiable; if any of them is violated, the resulting practice cannot be called responsible agentic R&D regardless of how well the other principles are followed.

2.1 Principle 1: Humans Set the Questions

> *The formulation of research questions, the identification of gaps in knowledge, and the creative leaps that define scientific progress must remain human responsibilities.*

AI agents can suggest questions, identify patterns in the literature, and propose hypotheses. But the decision of what is worth investigating, what matters, and what constitutes a meaningful contribution to human knowledge is a fundamentally human judgment. Agents assist in exploration; humans decide the direction.

Science is not simply the accumulation of answers; it is the art of asking the right questions. A research question encodes values, priorities, and a vision of what knowledge is missing from the world. These are decisions that require lived experience, disciplinary intuition, and an understanding of the social context in which research operates. An agent can identify that a topic is under-studied, but only a human can judge whether it *should* be studied, and why it matters.

This is not merely an assertion of human primacy for sentimental reasons. It reflects a practical reality: the most consequential decisions in research are those that determine what questions get asked. A field that allows AI agents to set its research agenda risks optimising for what is computationally tractable or what generates the most citations rather than what is most important. The history of science is full of transformative questions that emerged not from data analysis but from human curiosity, ethical concern, or practical need: Darwin's question about the origin of species, Curie's investigation of radioactivity, or Turing's inquiry into machine intelligence. These questions were shaped by their authors' experiences, observations, and intellectual courage, qualities that no current AI system possesses.

In practice, this means a researcher defines the research question and hypothesis before invoking any agent. The agent may suggest related questions, identify gaps in the literature, or highlight contradictions in existing findings, but the decision to pursue a direction is always human. A project begins with a human-authored research brief, not with a prompt asking an agent what to study. For example, a researcher investigating media bias in political coverage would first articulate the specific aspect of bias they wish to examine, the theoretical framework they intend to use, and why this question matters to society. Only then would they ask an agent to survey existing literature on the topic, identify datasets, or suggest methodological approaches. The research brief serves as both a starting point and a reference document against which subsequent agent suggestions can be evaluated.

2.2 Principle 2: Agents Amplify, They Do Not Replace

> *The role of AI agents in research is to amplify human capability, not to replace human judgment. Every agentic contribution should make a human researcher more effective, not less necessary.*

This principle guards against the perverse incentive to use agents to increase publication volume at the expense of quality. The goal is not more papers but better research. An agent that helps a researcher think more deeply about their data is more valuable than one that produces three drafts in the time it takes to write one.

The amplification metaphor is deliberate. A microscope amplifies human vision; it does not replace the biologist's understanding of what they are looking at. Similarly, an AI agent that summarises 500 papers amplifies a researcher's ability to survey a field; it does not replace their capacity to evaluate which findings are methodologically sound, which are surprising, and which change the direction of inquiry. The moment an agent's output is accepted without critical evaluation, amplification has become replacement.

The distinction matters because replacement is insidious. It rarely happens as a conscious decision; it happens gradually, as researchers become accustomed to accepting agent outputs without scrutiny. A researcher who initially reads every agent-generated summary carefully may, after months of reliable outputs, begin to skim them. Over time, the human's role drifts from active collaborator to passive approver, and the quality of the research degrades without anyone noticing. Guarding against this drift requires deliberate practices: periodic deep-reading of agent outputs, explicit verification routines, and a culture that values critical engagement over efficiency.

In practice, a researcher who uses an agent to draft a literature review reads every cited paper (or at minimum, every paper that supports a key claim) and rewrites the synthesis in their own analytical voice. The agent provides breadth; the human provides depth and judgment. If a researcher cannot explain and defend every claim in a manuscript, the agent has replaced rather than amplified. Concrete warning signs include: accepting agent-generated abstracts without revision, submitting papers that the authors have not read in full, or using agent-produced experimental interpretations verbatim. Research teams can guard against replacement by instituting a "can you defend it?" test: before submission, each author must be able to explain, without notes, the reasoning behind every major claim in the paper.

Governance Principles

The governance principles translate the foundation's philosophy into concrete accountability mechanisms. They specify how transparency is achieved (Principle 3), how trust is built (Principle 4), how oversight is calibrated (Principle 5), and how credit is assigned (Principle 8). Together, they form the institutional infrastructure that makes responsible agentic R&D verifiable rather than merely claimed.

2.3 Principle 3: Transparency Is Non-Negotiable

> *Every publication that involves AI agent contributions must transparently report what the agent did, how it was supervised, and what was verified. The Transparency Card should become standard practice.*

Binary disclosure ("AI was used") is insufficient. The research community needs structured, detailed reporting of AI contributions at the task level. We propose the Transparency Card as the standard format: a structured appendix documenting, for each research phase, whether an agent was involved, what tool was used, what autonomy tier was applied, and what human verification was performed. We commit to including one in every OpenCódice publication.

Major publishers have already recognised that simple disclosure is inadequate. *Nature* [Nature Editorial, 2023] and *Science* [Thorp, 2023] have published editorial policies requiring disclosure

of AI tool usage, but these policies stop short of specifying what form that disclosure should take. The Transparency Card fills this gap by providing a structured, machine-readable format that enables systematic analysis of AI involvement across the literature.

The value of structured transparency extends beyond accountability. It enables the research community to build an empirical understanding of which agentic workflows produce reliable results and which do not. If thousands of papers include Transparency Cards documenting the specific tools, configurations, and oversight mechanisms used, meta-researchers can analyse these data to identify best practices, common failure modes, and the relationship between oversight intensity and output quality. Transparency thus serves not only as a check on individual papers but as a data source for improving agentic R&D practices across the community.

In practice, every publication produced with agent assistance includes an Agentic R&D Card documenting which phases involved AI, what tools were used, and what autonomy tier was applied. For example: "Literature search: Agent (Claude, Tier 3: supervised autonomy); data analysis: Agent (GPT-4, Tier 2: human-on-the-loop); writing: Human with agent drafting support (Tier 1: human-in-the-loop); all citations independently verified by human". The card also records verification outcomes: how many agent-generated citations were checked, how many errors were found and corrected, and what percentage of the final text differs from the agent's initial draft. This level of detail allows reviewers to assess the depth of human engagement and readers to calibrate their confidence in specific sections.

2.4 Principle 4: Trust Is Earned Through Verification

> *No agent output should be published without verification. Trust in agentic research is built through audit trails, reproducible workflows, and human validation at critical checkpoints.*

The trustworthiness of agent-assisted research cannot be assumed; it must be demonstrated. This means verifying citations, checking data, validating results, and maintaining audit trails that enable independent verification. The cost of verification is the price of trustworthy agentic research.

Verification is not merely a technical requirement; it is an epistemic one. The history of science is punctuated by cases where trusted processes produced unreliable results, from the Schon fabrication scandal in physics to the replication crisis in psychology [Baker, 2016]. AI agents introduce a new category of error: plausible-sounding but incorrect outputs ("hallucinations") that can be difficult to detect precisely because they are fluently expressed. Building trust requires not just verifying individual outputs but maintaining systems that make verification routine and auditable.

The challenge of verification scales with the volume of agent output. An agent can produce in minutes what would take a researcher days. If verification takes as long as doing the work manually, the efficiency gains disappear. Effective verification strategies therefore rely on structured sampling, cross-validation, and automated checks rather than exhaustive manual review. For example, a research team might automatically verify all URLs and DOIs in agent-generated bibliographies, spot-check a random 25% of factual claims against primary sources, and use a second, independent agent to cross-check statistical calculations. The goal is a verification pipeline that is rigorous enough to catch errors but efficient enough to preserve the productivity benefits of agentic workflows.

In practice, every agent-generated citation is checked against the original source. Every statistical result is independently reproduced using the raw data. Every factual claim is traced to its evidence. Projects maintain an audit log recording each agent interaction, its output, and the verification decision. When an agent produces a literature summary, the researcher spot-checks at least 20% of the cited papers for accuracy of representation. The audit log is structured as a timestamped record that captures the agent query, the raw response, the verification action taken, and the final disposition (accepted, revised, or rejected). This log serves as both an internal quality control mechanism and evidence for the Transparency Card.

2.5 Principle 5: Oversight Scales With Risk

> *The level of human oversight should be proportional to the risk and creativity requirements of each task. Formatting needs no oversight. Hypothesis generation needs full human control.*

One-size-fits-all approaches to AI oversight are inefficient: requiring human approval for every agent action is as problematic as requiring none. Tiered autonomy models provide a framework for calibrating oversight to risk, from full agent autonomy for low-risk tasks to full human control for high-risk ones. The EU's Ethics Guidelines for Trustworthy AI [High-Level Expert Group on AI, 2019] similarly advocate for proportionate governance, and this principle applies that logic to the specific context of scientific research.

The principle recognises that research tasks vary enormously in their consequences for scientific integrity. Converting a table from CSV to LaTeX format carries minimal risk; an error is easily detected and corrected. Interpreting ambiguous experimental results carries high risk; an error could lead to false claims in the published literature. Calibrating oversight to these differences allows researchers to delegate freely where it is safe to do so, while maintaining tight control where it matters most.

Risk in research is not one-dimensional. It encompasses at least three factors: the potential for factual error (accuracy risk), the potential for misrepresentation of evidence (integrity risk), and the potential for irreversibility (consequence risk). A formatting error has low accuracy risk and is easily reversible. A misinterpreted dataset has high accuracy risk, high integrity risk, and potentially high consequence risk if it leads to published claims that influence policy or clinical practice. Research teams should assess tasks along all three dimensions when assigning autonomy tiers, recognising that a task with low accuracy risk but high consequence risk (such as generating a patient-facing summary of clinical trial results) may still require tight oversight.

In practice, a research team assigns each task an autonomy tier before delegating it to an agent. Tier 0 (full automation) applies to formatting, reference formatting, and file management. Tier 1 (human-in-the-loop) applies to literature summarisation and data visualisation. Tier 2 (human-on-the-loop) applies to experimental execution with pre-approved protocols. Tier 3 (human-in-control) applies to research question formulation, interpretation of results, and writing of discussion sections. These assignments are documented in the project's Transparency Card. Teams review and adjust tier assignments periodically as they gain experience with a given agent's reliability on specific tasks; an agent that consistently produces accurate literature summaries might be promoted from Tier 1 to Tier 0 for that task, while one that frequently misinterprets statistical results would be restricted to Tier 2 or higher oversight.

Practice Principles

The practice principles address the material conditions that make responsible agentic R&D possible at scale. They concern infrastructure (Principle 6), methodology (Principle 7), credit

(Principle 8), and equity (Principle 9). While the governance principles specify what must be done, the practice principles specify how to build systems that make compliance feasible and sustainable.

2.6 Principle 6: Agents Must Be Agent-Agnostic Ready

> *Research projects should be structured to support collaboration with any AI agent, not locked into a specific vendor or tool. Interoperability is a research infrastructure concern.*

Today's dominant agent may not be tomorrow's. Research projects that depend on a specific AI system create fragile infrastructure. Agent-ready project design ensures that the project structure, not the agent, is the durable asset. This means using standardised file formats, clear directory conventions, machine-readable metadata, and explicit instructions (such as an `AGENT.md` file) that any capable agent can follow.

The principle draws a parallel with data management best practices. Just as researchers are encouraged to store data in open, non-proprietary formats to ensure long-term accessibility, they should structure their projects so that any competent AI agent can contribute. A project that can only be worked on by one specific model is as fragile as a dataset stored in a proprietary binary format. The FAIR principles for data management (Findable, Accessible, Interoperable, Reusable) [Wilkinson et al., 2016] provide a useful analogy: research projects should be structured so that their components are findable by any agent, accessible through standard interfaces, interoperable across tools, and reusable in future work.

Vendor lock-in carries risks beyond mere inconvenience. If a research group builds its entire workflow around a proprietary agent platform that later changes its pricing, terms of service, or capabilities, years of methodological investment may be lost. Consider a laboratory that structures all its experimental protocols around a single AI system's proprietary format: when that system is deprecated or becomes prohibitively expensive, the laboratory must rebuild its workflows from scratch. Agent-agnostic design treats this as a preventable failure. By investing in clear, human-readable project structures that serve as the interface between researcher and agent, teams insulate themselves from the volatility of the AI tool market while preserving the ability to adopt superior tools as they emerge.

The concept of an `AGENT.md` file deserves elaboration. This is a plain-text file, placed at the root of a research project, that describes the project's structure, conventions, and expectations in language that both humans and AI agents can parse. It might specify the directory layout, naming conventions for data files, the citation style, formatting requirements, and task-specific instructions (for example, "when summarising papers, always include the sample size, methodology, and key limitations"). The `AGENT.md` functions as an onboarding document for any agent, new or replacement, and simultaneously as documentation for human collaborators.

In practice, a research project includes an `AGENT.md` file that describes the project structure, naming conventions, style guidelines, and task-specific instructions in plain language. The project uses standard file formats (LaTeX, BibTeX, CSV, JSON) rather than tool-specific formats. Configuration is separated from content. When switching from one AI agent to another, the project structure requires no modification; only the agent changes. Teams test agent-agnosticism periodically by having a different AI system perform a routine task (such as formatting a bibliography or generating a data summary) and verifying that the output integrates seamlessly with the existing project structure.

2.7 Principle 7: Reproducibility Extends to the Agent

> *Reproducibility requires documenting not just the data and methods but the agent: its model, version, configuration, and the prompts that guided its work.*

If an AI agent contributed to a research output, another researcher should be able to understand (and ideally reproduce) the agent's contribution. This means logging model versions, system prompts, tool configurations, and intermediate outputs.

Reproducibility is a cornerstone of the scientific method [Popper, 1959]. When a human researcher describes their methods, another researcher can attempt to replicate the study. The same standard must apply to agent contributions. If an agent was used to analyse data, the model name, version, temperature setting, system prompt, and input data must all be recorded. Without this information, the agent's contribution is a black box, and the research cannot be fully evaluated or reproduced.

AI agents introduce a distinctive challenge for reproducibility that goes beyond traditional computational reproducibility. Two runs of the same language model with identical prompts can produce different outputs due to stochastic sampling, and model behaviour may change between versions without public documentation of the changes. A paper that reports "we used GPT-4 to analyse our data" in 2024 cannot be meaningfully replicated in 2025 if the underlying model has been updated. This means that reproducibility in agentic R&D requires not only documenting the configuration but also archiving the actual outputs at each stage. The combination of configuration logging and output archiving creates a reproducibility chain that enables future researchers to evaluate the agent's contribution even if they cannot exactly replicate it.

We propose distinguishing between two levels of agentic reproducibility. *Configuration reproducibility* means that another researcher can see exactly what model, prompt, and parameters were used, enabling them to attempt replication with the same or equivalent tools. *Output reproducibility* means that the actual agent outputs are archived alongside the final publication, enabling evaluation of the agent's contribution without requiring access to the original model. Both levels are valuable; the first supports replication, while the second supports evaluation and audit. Together, they provide a robust basis for assessing the reliability of agent-assisted research.

In practice, a project's methods section (or supplementary materials) includes an "Agent Configuration" appendix listing: the model used (e.g., Claude 3.5 Sonnet, GPT-4o), the API version or access date, the system prompt or instruction file, key parameters (temperature, max tokens), and any tool integrations. Intermediate outputs (e.g., raw agent responses before human editing) are archived alongside the final manuscript. For computational experiments, the complete prompt sequence and agent responses are stored in a version-controlled repository. For writing tasks, the diff between the agent's draft and the final human-edited version is preserved, allowing reviewers to assess the extent and nature of human revision.

2.8 Principle 8: The Credit Follows the Contribution

> *Researchers who use AI agents retain full responsibility for their publications. Agents are tools, not authors. But the intellectual contribution of humans must be genuine, not nominal.*

Listing a human as author while an agent did the intellectual work is a form of dishonesty. Conversely, an agent that formatted a paper does not deserve authorship. The Transparency

Card makes the actual division of labour visible, enabling the community to assess whether human authorship claims are substantive.

This principle applies Merton's norm of intellectual honesty to the age of AI agents. The scientific community has long debated gift authorship and ghost authorship; AI agents introduce a new variant of the same problem. When a researcher claims authorship of a paper largely written by an agent, with minimal human intellectual contribution, this constitutes a form of ghost authorship that undermines the trust on which scientific credit depends.

The principle also has implications for the evaluation of researchers. If hiring committees, tenure boards, and funding agencies cannot distinguish between a researcher who produced ten papers through deep intellectual engagement and one who produced ten papers by prompting an agent and accepting the output, the incentive system breaks down. The Transparency Card addresses this by making the nature of human contribution visible, but the broader research community must also develop norms for interpreting this information. A researcher who transparently documents their collaboration with AI agents and demonstrates genuine intellectual contribution should be evaluated more favourably, not less, than one who conceals AI involvement.

In practice, each author on a publication can articulate their specific intellectual contribution beyond what the agent provided. If an agent drafted the entire manuscript, the human authors must demonstrate that they substantively revised the argument, verified the evidence, and can defend the claims. The Transparency Card provides the evidence base for this assessment, and co-authors review it before submission to ensure that credited contributions are genuine. A useful test is the "intellectual ownership" criterion: for each section of a paper, at least one human author must be able to explain why the argument takes the form it does, what alternatives were considered and rejected, and how the section connects to the paper's overall contribution. If no author can do this for a given section, that section has not received sufficient human intellectual engagement.

2.9 Principle 9: Benefits Must Be Shared

> *Agentic R&D should reduce barriers to research, not raise them. Tools, frameworks, and best practices for agent-assisted research should be openly shared.*

If agentic R&D benefits only well-funded labs with access to expensive AI systems, it will widen the gap between research haves and have-nots. We commit to open-sourcing all frameworks, templates, and tools produced by this project. This commitment aligns with the principles of open science [Fecher and Friesike, 2014], which hold that the outputs of publicly funded research should be accessible to all.

The risk of inequitable access is real and immediate. State-of-the-art language models require significant computational resources, and API access carries ongoing costs. If agentic R&D practices are developed exclusively by well-resourced institutions and shared only through proprietary tools, the result will be a two-tier research system in which some researchers benefit from AI acceleration while others are left further behind. This dynamic mirrors earlier inequalities in computational research: when high-performance computing was available only to a few institutions, entire fields of inquiry were effectively closed to researchers at smaller universities or in lower-income countries. Agentic R&D must not repeat this pattern.

Sharing benefits also means sharing knowledge about failures and limitations. The research community learns as much from documented failures as from successes. When a team discovers that an agent-assisted workflow produces unreliable results for a particular task, publishing

that finding openly prevents other teams from investing time and resources in the same dead end. A culture of openness about what does not work is as important as sharing what does. This includes publishing negative results from agentic experiments, documenting cases where human-only workflows outperformed agent-assisted ones, and contributing to shared repositories of best practices and known pitfalls.

Equity in agentic R&D also requires attention to language and geography. Most AI research tools are optimised for English-language workflows and trained predominantly on English-language scientific literature. Researchers working in other languages, or studying phenomena specific to non-Anglophone contexts, face additional barriers. The agentic R&D community should actively support multilingual tools, document workflows for non-English research, and ensure that templates and best practices are accessible to researchers regardless of their working language or geographic location.

In practice, all project templates, LaTeX classes, agent instruction files, and workflow documentation are released under open licences (CC BY 4.0 for documents, MIT for code). Methodological papers describe workflows in sufficient detail for researchers with limited budgets to adapt them using open-source or lower-cost alternatives. Community contributions are welcomed and credited. When proprietary tools are used in a workflow, the documentation specifies open-source alternatives and explains what modifications are needed to achieve comparable results. Teams maintain public repositories of their agentic workflows, including the `AGENT.md` files, Transparency Card templates, and example audit logs, so that other research groups can adopt and adapt them without starting from scratch.

The Goal

The final principle stands apart from the others. While Principles 1–9 specify how agentic R&D should be conducted, Principle 10 specifies why. It is the criterion against which all other principles, and all practical decisions about human-agent collaboration, should ultimately be evaluated.

2.10 Principle 10: The Goal Is Better Science

> *The ultimate measure of agentic R&D is not efficiency, not publication count, and not cost savings. It is whether the science produced is more rigorous, more reproducible, more inclusive, and more impactful.*

Every decision about how to integrate AI agents into research should be evaluated against this standard. If a practice makes research faster but shallower, it fails. If it makes research slower but more rigorous, it succeeds.

This principle serves as the compass for all the others. When trade-offs arise (and they will), this principle resolves them. Should we use an agent to produce three papers in the time it takes to write one? Only if all three meet the same standard of rigour. Should we automate peer review to speed up publication cycles? Only if automated review is at least as thorough as human review. The measure is not speed or volume; it is the quality and integrity of the knowledge produced.

The principle also serves as a corrective to the prevailing incentive structures of contemporary science. The "publish or perish" culture [Rawat and Meena, 2014] already pressures researchers to prioritise quantity over quality; AI agents risk amplifying this pressure by making high-volume production easier. A research group that uses agents to publish twenty mediocre papers per year instead of five excellent ones has not advanced science; it has contributed to the noise that

makes genuine findings harder to identify. Principle 10 insists that agentic R&D should resist this temptation and instead channel the productivity gains from AI assistance toward deeper analysis, more thorough verification, and more careful communication.

In practice, research teams conduct periodic retrospectives asking: "Has our use of AI agents improved the quality of our research, or only its quantity?" Metrics of success include depth of analysis, robustness of conclusions, completeness of literature coverage, and reproducibility of results, not the number of papers published or the speed of submission. When an agent-assisted workflow produces lower-quality outputs than a human-only workflow would have, the team revises or abandons the agentic approach for that task. Teams track quality indicators over time (such as citation quality, reviewer feedback, and replication success rates) to build an evidence base for evaluating whether their agentic practices are achieving the goal of better science.

3 Commitments and Call to Action

The principles outlined in the previous section provide a conceptual foundation, but principles without action are merely aspirational. This section translates those principles into concrete commitments and specific calls to action addressed to the various stakeholders who shape the research ecosystem. We recognise that responsible agentic R&D cannot be achieved by researchers alone; it requires coordinated effort from publishers, institutions, funding agencies, tool developers, and the broader scientific community. Each stakeholder occupies a different position within the ecosystem and wields different levers of influence; the calls to action that follow are tailored to the specific role and capacity of each group.

3.1 Our Commitments

As the authors and practitioners behind the OpenCódice Agentic R&D initiative, we commit to:

1. **Transparency**: Including a Transparency Card in every OpenCódice publication that involves AI agent contributions. We will iterate on the card format based on community feedback, publish versioned specifications, and maintain backward compatibility so that older cards remain interpretable.

2. **Open tools**: Releasing all frameworks, templates, and tools (including the Transparency Card template, the `AGENT.md` specification, and the LaTeX document class) under open licences. We will maintain these tools actively, respond to community issues, and accept contributions through standard open-source workflows.

3. **Honest attribution**: Clearly distinguishing between human and agent contributions in all publications, resisting the temptation to obscure AI involvement. We will document our own failures and corrections openly, including cases where we initially over-relied on agent outputs and had to revise our practices.

4. **Continuous improvement**: Revisiting these principles annually as AI capabilities and research practices evolve. Each revision will document what changed, why, and what evidence motivated the change, creating a living record of the community's evolving understanding of responsible agentic R&D.

5. **Community engagement**: Inviting feedback, criticism, and collaboration from the research community. We will host open discussions, respond to published critiques, and actively seek perspectives from disciplines and institutions underrepresented in current AI research discourse.

3.2 Call to Researchers

Researchers are the primary agents of change in the scientific ecosystem. While institutions set policies and publishers define norms, it is individual researchers and research teams who decide, day by day, how AI agents are integrated into their workflows. The choices that researchers make now will establish the precedents and practices that define the next generation of scientific work. If the research community adopts transparent, principled approaches to agentic R&D from the outset, these practices will become the norm rather than the exception.

The adoption of responsible agentic R&D practices need not begin with institutional mandates; it can begin with individual researchers who choose to hold themselves to a higher standard. A single laboratory that consistently includes Transparency Cards in its publications creates a visible example for its field. A research group that openly documents its agentic workflows, including the failures, contributes to a shared knowledge base that benefits the entire community. Leadership in this area is available to anyone willing to invest the effort.

We invite researchers to:

- Adopt the Transparency Card for AI-assisted publications, even when journals do not yet require it. Early adoption creates precedent and generates the evidence base that will eventually inform policy.
- Structure projects using agent-ready design patterns, including `AGENT.md` files, standardised directory layouts, and machine-readable metadata.
- Apply tiered autonomy models when delegating tasks to AI agents, and document the tier assignments in project records.
- Share experiences, best practices, and failures openly, including through blog posts, preprints, and community forums.
- Establish internal review processes for AI-assisted outputs that are at least as rigorous as those for human-only work.

3.3 Call to Students and Early-Career Researchers

The integration of AI agents into research raises particular concerns for those building their scholarly identity. Doctoral students may worry that using AI assistance diminishes the value of their dissertation. Early-career researchers may fear that transparent disclosure of AI contributions will lead reviewers or hiring committees to discount their work.

We believe the opposite is true. A researcher who can demonstrate thoughtful, transparent, and effective collaboration with AI agents is better prepared for the future of science than one who either avoids these tools entirely or uses them covertly. The key is that the intellectual contribution must be genuine: the student must formulate the questions, design the methodology, interpret the results, and construct the argument. The agent may help with execution, but the scholarship must be the student's own.

The current generation of doctoral students is uniquely positioned to shape the norms of agentic R&D. Unlike established researchers who may view AI tools as supplements to existing workflows, students are developing their research practices from scratch. The habits they form now, whether those habits involve thoughtful integration or uncritical delegation, will define their careers and influence their future students and collaborators. This is both a responsibility and an opportunity.

We encourage students and early-career researchers to:

- Discuss AI tool usage openly with supervisors and establish shared expectations early in the research process.

- Use the Transparency Card to document their own learning: which tasks they delegated, which they performed themselves, and how their judgment shaped the agent's contributions.
- Treat AI collaboration as a skill to be developed and demonstrated, not a shortcut to be hidden. Include descriptions of agentic workflows in research portfolios and job applications.
- Advocate for institutional policies that distinguish between responsible AI collaboration and academic dishonesty.
- Maintain a personal learning log that tracks how their use of AI tools evolves over time, including reflections on when delegation improved their work and when it did not.

3.4 Call to Publishers

Publishers occupy a unique position in the scientific ecosystem: they define the standards against which research is evaluated. The current patchwork of AI disclosure policies, ranging from outright bans to vague requests for acknowledgement, leaves researchers without clear guidance and reviewers without consistent criteria. Structured reporting is essential for reproducibility because it enables reviewers and readers to assess exactly how AI agents contributed to a given publication. Without such standards, the research community cannot distinguish between papers that used AI responsibly and those that did not.

The publishing community has successfully standardised other aspects of research reporting: the CONSORT guidelines for clinical trials [Schulz et al., 2010], the PRISMA framework for systematic reviews [Moher et al., 2009], and data availability statements are all examples of structured reporting requirements that initially met resistance but are now widely accepted as essential for scientific integrity. AI contribution reporting is the natural next step in this progression.

We invite journals and conferences to:

- Move beyond binary AI disclosure to structured reporting (the Transparency Card or an equivalent format), and provide clear templates that authors can use.
- Develop reviewer guidelines for evaluating AI-assisted papers, including criteria for assessing whether human intellectual contribution is genuine and sufficient.
- Explore mechanisms for verifying AI contribution claims, such as requiring submission of audit logs or agent configuration files as supplementary materials.
- Recognise and reward transparency: papers that include thorough AI contribution documentation should be viewed favourably, not penalised.
- Contribute to cross-publisher standardisation efforts so that researchers face consistent requirements rather than navigating a patchwork of incompatible policies.

3.5 Call to Funding Agencies

Funding agencies wield considerable influence over research practices through the conditions they attach to grants. They have the power to mandate transparency standards, require reproducibility documentation, and incentivise responsible innovation. Historically, funding agencies have been catalysts for systemic change in research practice: the NIH's data sharing policy, the European Commission's open access mandate, and the NSF's data management plan requirement all demonstrate that funding conditions can shift community norms more effectively than voluntary guidelines alone.

As AI agents become integral to research workflows, funding agencies must decide whether to treat this transformation as a matter for individual researchers to navigate or as a systemic challenge requiring coordinated policy. We believe the latter. The costs of irresponsible agentic R&D (unreliable results, wasted resources, eroded public trust) fall disproportionately on the public that funds scientific research. Agencies that invest in establishing clear standards now will avoid far greater costs in remediation later.

We invite funding agencies to:

- Require grant applicants to describe their intended use of AI agents in research workflows, including planned oversight mechanisms and transparency measures.
- Include AI transparency documentation (such as the Transparency Card) among the deliverables for funded projects that use AI tools, just as data management plans are now standard requirements.
- Fund research on agentic R&D practices themselves, including studies of how AI tools affect research quality, reproducibility, and equity across disciplines and institutions.
- Support the development of open infrastructure for responsible agentic R&D, including open-source tools for audit logging, transparency reporting, and reproducibility verification.
- Establish evaluation criteria that reward transparent and responsible AI use rather than penalising disclosure, ensuring that researchers are not disadvantaged by honesty about their workflows.

3.6 Call to Institutions

Universities and research institutions set the structural conditions under which research is conducted. Their policies on academic integrity, their investments in infrastructure, and their criteria for hiring and promotion collectively determine whether responsible agentic R&D is feasible in practice. An individual researcher who wishes to adopt transparent AI collaboration will struggle to do so if their institution treats all AI use as misconduct, or conversely, if it provides no guidance at all. Institutions must move beyond reactive positions and actively enable responsible integration of AI agents into the research process.

The institutional response to AI in research has so far been characterised by extremes: some institutions have banned AI tools outright, while others have remained silent, leaving researchers to navigate ethical ambiguity on their own. Neither approach is adequate. Blanket bans ignore the legitimate productivity benefits of AI assistance and drive usage underground. Silence leaves researchers without the guidance they need to use these tools responsibly. What is needed is a middle path: clear policies that distinguish between responsible and irresponsible uses, supported by training, infrastructure, and evaluation criteria that reward transparency.

We invite universities and research institutions to:

- Develop policies for AI use in research that go beyond prohibition or blanket permission, specifying which uses are encouraged, which require disclosure, and which are prohibited.
- Invest in training researchers to collaborate effectively with AI agents, including workshops on prompt engineering, verification strategies, and ethical considerations.
- Ensure equitable access to AI tools across departments and career stages, so that agentic R&D does not become the exclusive province of well-funded laboratories.

- Update academic integrity policies to distinguish between responsible AI collaboration (transparent, verified, with genuine human contribution) and academic dishonesty (concealed, unverified, with nominal human contribution).
- Include AI collaboration skills in graduate training curricula, preparing the next generation of researchers for a world in which human-agent collaboration is the norm.

3.7 Call to Developers of AI Research Tools

The developers of AI systems used in research bear a particular responsibility. The design choices embedded in these tools shape how millions of researchers interact with AI, and those choices can either support or undermine the principles in this decalogue. A tool that logs its interactions by default makes transparency easy; one that does not makes transparency burdensome. A tool that exports its configuration in a standard format supports reproducibility; one that keeps its internals opaque undermines it. Developers have the power to make responsible agentic R&D the path of least resistance, and we urge them to exercise that power deliberately.

We invite developers to:

- Build transparency features directly into AI research tools: automatic logging of model versions, prompts, and configurations that can be exported for Transparency Cards.
- Implement audit trails by default, so that every agent interaction in a research workflow is recorded and retrievable.
- Support interoperability and open standards, enabling researchers to switch between tools without losing their project structure or workflow history.
- Design for proportional oversight, allowing researchers to configure autonomy levels for different task types rather than offering only "fully autonomous" or "fully manual" modes.
- Avoid lock-in practices that make research projects dependent on a single platform or model.
- Publish documentation about model versions, training data changes, and capability modifications that may affect reproducibility, so that researchers can accurately report the tools they used.

The promise of agentic R&D is not that AI will do science for us, but that it will free us to do science better. The ten principles in this decalogue are our attempt to ensure that this promise is fulfilled responsibly, transparently, and in service of the scientific enterprise that belongs to all of humanity. We offer them not as final answers but as a starting point for the conversation that the research community must have. We invite debate, criticism, refinement, and, above all, action.

References

Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016.

Benedikt Fecher and Sascha Friesike. Open science: One term, five schools of thought. *Opening Science*, pages 17–47, 2014.

High-Level Expert Group on AI. Ethics guidelines for trustworthy AI, 2019.

Weixin Liang, Yaohui Zhang, Zhengxuan Cao, Haley Zhao, Dianbo Shi, and James Zou. Mapping the increasing use of LLMs in scientific papers. *arXiv preprint arXiv:2404.01268*, 2024.

Robert K Merton. *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press, 1973.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, 6 (7):e1000097, 2009.

Nature Editorial. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature*, 613:612, 2023.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.

Karl Popper. *The Logic of Scientific Discovery*. Hutchinson, 1959.

Seema Rawat and Sanjay Meena. Publish or perish: Where are we heading? *Journal of Research in Medical Sciences*, 19(2):87–89, 2014.

Kenneth F Schulz, Douglas G Altman, and David Moher. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ*, 340:c332, 2010.

H Holden Thorp. Science journals: Editorial policies for ChatGPT. *Science*, 379(6634):775, 2023.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):1–9, 2016.

TRANSPARENCY CARD

The Agentic R&D Decalogue • OC-WP-2026-001

Francisco-Javier Rodrigo-Ginés • OpenCódice Research • 2026

Agent Identity

Claude (claude-opus-4-20250514) • Anthropic • CLI (Claude Code) • CLAUDE.md • Web search

Phase-Level AI Involvement

T1 Human Control **T2** Agent Drafts **T3** Agent Executes **T4** Full Autonomy

#	Research Phase	AI?	Tier	Tasks	Tools
1	Problem Identification	✔	T1	Proposed ten principles framing	Claude Code
2	Literature Engagement	✔	T1 T2	Searched philosophy of science; AI governance	Claude Code, web
3	Research Design	✔	T1	Proposed principle clustering; structure	Claude Code
4	Data Collection	✘	–	–	–
5	Experimentation	✘	–	–	–
6	Writing	✔	T1 T2	Drafted preamble; principles; commitments; TikZ	Claude Code
7	Peer Review	✘	–	–	–
8	Dissemination	✘	–	–	–
9	Funding & Admin	✘	–	–	–

Intellectual Contribution Statement

The ten principles, their Foundation/Governance/Practice clustering, calls to action, and all normative judgments were conceived by the author. The AI agent assisted with literature search, drafting, TikZ figures, and formatting. All agent-produced text was reviewed and revised. The philosophical framing and recommendations are the author's own work.

OpenCódice Agentic R&D Transparency Card v1.0 • opencodice.org

OpenCódice Working Paper OC-WP-2026-001 • 2026 • OpenCódice Research

opencodice.org • DOI: 10.5281/zenodo.19151516