Reproducibility in the age of AI: Workflows with Division of Labor between Humans and Agents

Anshu Dubey
RIKEN Center for Computational Science

COLLABORATIVE AI / SCIENTIFIC SOFTWARE ENGINEERING

AI can write code. Who verifies the science?

Scientific software is a research instrument—not just a codebase.

AGENTS

Generate • translate • refactor
Run tools and workflows

THE CHALLENGE

Numerical correctness
Scientific understanding
Long-term maintainability

Completing a run is not scientific acceptance.

Humans specify. Agents implement. Humans accept.

A staged workflow with an explicit verification boundary.

DOMAIN EXPERTS

Specify science and acceptance tests

→

SOFTWARE ENGINEERS

Define tools, plan and execution policy

→

BOUNDED AGENTS

Implement and review code

→

VERIFICATION GATE

Numerical comparison + human approval

Only an approved stage unlocks the next.

The Lab Notebook makes agent work inspectable

Shared, versioned artifacts—not hidden state inside an agent session.

SPEC + PLAN

Human-owned intent

TOOLS + POLICY

Controlled execution

LOGS + PROVENANCE

What happened and why

APPROVALS

Human-owned acceptance

Same task and acceptance criteria → different orchestrators can be compared.