# Tutorial & Assignment Week 4B — Claude Science

Goals: Use Claude Science to make progress on a project you care about, practice directing an AI agent, and evaluate the evidence and work behind its result.

Due: Submit the indicated portions on Gradescope by Sunday Oct 18th at 11:59pm, and discuss with your TA during tutorial in Week 5.

Hello everyone! This week, most of the tutorial will be yours to work on something you want to understand, decide, or build. We will start with a short conversation about the guest speakers and a technical deep dive into verifiability, and then you will use the tool on a project of your choice.

In Week 2, you practiced describing a task clearly. In Week 3, you chose the sources an AI system could use. This week, you will give an AI a goal and let it take several steps toward that goal, while you decide whether its approach makes sense and what to trust.

## Assignment outline:

This document is intended to be started in tutorial and completed for homework if necessary. Work through it in the order given.

- Discuss the guest speakers at your table, followed by a few brief share-outs.
- Get a better understanding of the tech behind these tools
- Use the rest of the tutorial for your own project/own purposes while TAs conduct individual Assignment 3 checks.
- Submit a short account of your exploration, a check of the resulting work, and a trust reflection.

## Part 1: Tutorial group work

### Group Activity 1 — Guest Speaker Reactions!

At your table, share something that struck you in what the guest speakers said. What did you find useful, surprising, or worth questioning? Did anything change how you think about using AI in your own work?

**Deliverable:** You don't need to submit anything for this portion of the assignment.

## Group Activity 2 — Deep Research

We will show how to start a research task, give it a useful scope, and inspect the resulting sources. Watch for the choices the person directing the research still needs to make: what question to ask, what evidence matters, and whether the findings actually answer the question.

> Ex: An engineer is beginning a new role in a research lab and is asked to contribute to a project developing technology for military applications. They aren't sure they morally agree with this work. Rather than relying on assumptions, they decide to investigate what the lab actually does and how its outputs are used before deciding whether to participate.
>
> They use an AI research tool (such as Deep Research) with the following prompt:
>
> - **Goal:** I'm deciding whether to join a research project developing military technology, and I want to decide based on evidence, not assumptions.
> - **Research:** What this type of lab does, who funds it, and how its work is used in practice, including civilian spin-offs and any documented harmful or military uses.
> - **My context:** I'm an early-career AI/engineering researcher, uncommitted, and open to persuasion either way. I care about downstream consequences, not just stated intent.
> - **Sources:** Prioritize primary sources (lab publications, funding disclosures, grant and contract records), peer-reviewed work, and credible reporting from the last five years. Include both critics and defenders.
> - **Output:** A brief of about 800 words covering (1) what the lab does, (2) funding and stakeholders, (3) documented uses, (4) the strongest arguments for and against participating, and (5) questions to ask the PI.
> - **Rules:** Cite sources for key claims, separate evidence from your inferences, and flag disagreements or gaps that could affect my decision.

At this link, Report, you'll find the report that Claude gave about this question. Note: it is excessively long and you should really only skim through it. But, keep in mind that this is the type of "normal" length that a deep research report provides.

Now, within your group, choose one or more of the questions to discuss.

1. **How much should we trust claims that rest on anonymous or contested sourcing?**
   The Minab findings come from anonymous officials describing an unreleased review, and the Gaza claims rest on a few intelligence sources the IDF disputes. Which claims in the brief would you be comfortable repeating, which would you hedge, and what would you want to see before upgrading your confidence?
2. **What would you check yourself before relying on an AI-generated research brief?**
   The brief flagged some of its own weak spots, like the $54.6B vs. $1B DAWG budget figure and the varying Google letter signature counts. Pick two or three claims and trace them to their original sources. Did the citations support the claims as stated? What did you find that the brief couldn't tell you, and what does that suggest about when to use deep research and when to go to primary sources?
3. **Where do credible sources disagree, and is it about facts or values?**
   Andrew Ng and Palmer Luckey argue that capable defense AI is a moral obligation, while the ICRC and the Minab reporting point to harms from the same technology. Sort the disagreements

in the brief into factual disputes (could be settled with better evidence) and value disagreements (would persist even with perfect evidence). Which type is the brief better at helping you resolve?

**Deliverable:** You don't need to submit anything for this portion of the assignment.

# Part 2: Individual AI tool work

## Explore what Claude created for you

Follow the Claude Science Setup Guide. During setup, Claude will suggest starter tasks based on the context you give it. Choose one that interests you, or describe a task of your own. If you already started one, pick up where you left off. You do not need to invent a second project for this assignment.

Use the rest of the tutorial to practice with the project and materials Claude creates. Ask questions, try changes, follow something surprising, or take the work in a direction that would be useful to you. You can also use this time to make progress toward your final project, due in about six weeks.

Your goal is to understand something about the problem and about how the agent works on it. You do not need a polished final product. An unfinished analysis, a failed attempt, or a result that changes your mind can be useful if you can explain what you learned.

## Connect your exploration to a technical question

Use the technical ideas from the opening discussion to guide your curiosity. What is the model generating? What information comes from a tool or a source? What code actually runs? How does the agent use earlier results to choose its next step? Pick a question that matters for your project and explore it through the work.

Here are some possible directions, depending on what Claude creates for you. These are suggestions, not a checklist.

- **If you are working with models or data:** Investigate how the result was produced and evaluated. What is being measured? Would a different dataset, baseline, or modeling choice change the conclusion? For example, a model that performs well on its training data may perform differently on data it has not seen.
- **If you are working with research or literature:** Explore how the agent finds and compares evidence. What comes from retrieved sources, and what is an inference? Does changing the question or scope change the evidence it uses? What can a summary leave out?
- **If you are working with a prototype or simulation:** Try it on a case you care about. What does it actually do? What assumptions are built into it, and where does it break? Ask Claude to help you investigate or extend it.

Customize, Files, and Compute are there to explore whenever they become useful. You might look at the instructions shaping the work, open the files behind a result, or inspect the environment where code runs. You do not need to visit every menu or configure additional computing resources.

Talk with your tablemates or TA if you get stuck or want to compare approaches. Keep enough of the conversation and resulting work to show what you tried and what happened.

**Deliverable, all copy and pasted into Gradescope:**

1. Briefly describe the starter task or project you explored, what you tried, and what you learned about a technical question. A short paragraph is enough.

# Part 3: Non-AI tool work

## Check something that matters

Step away from the AI and look at the work yourself. Choose one important claim, result, or behavior from your project and check whether the evidence supports it. Use a check that makes sense for what you were doing.

For a research claim, open the original source and see whether it supports the conclusion as stated. For an analysis, inspect the data or calculation behind a result. For a prototype, try a case where you can judge the behavior yourself. You can use ordinary tools to perform the check; asking Claude whether it is correct is not an independent check.

Claude Science can retain the code, environment, and conversation behind artifacts. Its reviewer can check claims against the execution record, but it does not re-run analyses or establish that the method is appropriate for your question. These records can help you investigate; you still need to decide what the result means.

**Deliverable, all copy and pasted into Gradescope:**

1. Write a short paragraph explaining what you checked, how you checked it, and what you found. What does this tell you about the result, and what remains uncertain?

# Part 4: Reflection — Trust

Write one short paragraph, approximately 4–6 sentences, responding to at least one of the questions:

- What did Claude Science let you accomplish or understand? Give one concrete example of a choice you still had to make yourself.
- Do you trust the result? How did exploring the work or checking it yourself affect your confidence?

- What would you do next before relying on or sharing the result? You may also connect your experience to something a guest speaker said.

**Deliverable:** Copy and paste your reflection into Gradescope.

# Submission

You should have three items to submit to Gradescope. Be ready to discuss these with your TA next week:
1. A short account of your exploration, with one or two prompts or exchanges
2. A check of one important claim, result, or behavior
3. Trust reflection