Why This Document Exists

Shocklab is an AI research group, which puts us in an unusual position: the tools we study are also the tools we use. Large language models, code assistants, image generators, and other AI systems are now woven into how we write, code, analyse data, and think. That is not going to change, and we should not pretend otherwise.

But the fact that these tools are useful does not mean they are neutral. They hallucinate. They plagiarise. They leak data. They encode biases. They make confident-sounding claims that are wrong. And because they are convenient, they can quietly erode the skills, judgement, and intellectual ownership that make research valuable in the first place.

This document exists to make our shared expectations about AI use explicit — not to restrict it, but to ensure that when we use these tools, we do so honestly, thoughtfully, and in a way that we can defend publicly. It sits alongside the main Shocklab Code of Conduct which remains the primary document governing how we treat each other. This one governs how we work with machines.

1. Core Principles

Three commitments from the main Code of Conduct — Care, Safety, Honesty — apply directly here, with an additional fourth:

• **Honesty.** We are transparent about when and how we use AI tools. We do not present AI-generated work as solely our own. We do not hide AI involvement to make ourselves look more productive or capable.

• **Accountability.** If your name is on a piece of work, you are responsible for everything in it — including the parts an AI helped produce. "The model generated it" is not a defence for errors, plagiarism, or misleading claims.

• **Competence.** AI tools should augment your skills, not replace them. You should be able to understand, explain, and defend any AI-assisted output that appears in your work. If you cannot, you have not used the tool responsibly.

• **Humility.** These tools are powerful but unreliable, and the landscape is changing fast. None of us fully understands what they can and cannot do. Acknowledge uncertainty. Ask questions. Share what you learn.

2. Disclosure and Transparency

The single most important norm in this document: **be honest about your AI use.**

What must always be disclosed

• Any use of AI tools in drafting, editing, or substantially restructuring text that appears in a publication, thesis, report, or grant application.

• Any use of AI tools to generate, debug, or substantially modify code that forms part of a research output.

• Any use of AI tools in data analysis, including feature engineering, model selection, or interpretation of results.

• The specific tool and, where possible, the model version used (e.g. "Claude 3.5 Sonnet, via the API" or "GitHub Copilot with GPT-4"). Model behaviour varies across versions, so this matters for reproducibility.

How to disclose

• In papers: follow the venue's policy. Where no policy exists, include a statement in the Methods or Acknowledgements section describing what the tool did and what you did.

• In theses: UCT's policy on AI disclosure applies. Include a declaration as part of your plagiarism statement, specifying which tools were used and for what purpose.

• In code: include a comment or README entry noting which components were AI-assisted. Commit messages like "refactored with Copilot" or "initial draft from Claude, manually verified" are good practice.

• In lab communications: no formal disclosure needed for routine use (drafting emails, brainstorming), but be honest if asked.

What does not need formal disclosure

• Using a spell checker, grammar checker, or standard autocomplete.

• Using AI to help you understand a concept (i.e. as a tutor), provided the final work is your own.

• Casual use in brainstorming or planning that does not directly produce research output.

The boundary is not always clean. When in doubt, disclose. Nobody will be penalised for over-disclosing.

3. Verification and Quality Control

AI systems produce plausible-sounding outputs that may be incorrect, fabricated, biased, or subtly misleading. The burden of verification is always on the human.

• **Never trust AI-generated citations without checking them.** LLMs routinely fabricate references — complete with plausible authors, titles, journals, and DOIs that do not exist. Every citation must be verified against the actual source.

• **Never trust AI-generated code without testing it.** AI-written code can contain subtle bugs, security vulnerabilities, or incorrect logic that passes superficial inspection. Test it the same way you would test your own code — more carefully, in fact, because you did not write it and may not fully understand its assumptions.

• **Never trust AI-generated numerical results or statistical claims.** LLMs are not calculators. They can and do make arithmetic errors, misapply statistical tests, and invent data points. Verify computationally.

• **Be sceptical of AI-generated summaries of papers or documents.** The model may omit key caveats, conflate distinct arguments, or misrepresent the source. Read the original.

• **Watch for bias in AI outputs.** These models are trained on internet-scale data and reflect its biases. Be particularly alert when working on topics involving people, demographics, health, or policy.

If an AI tool saves you time generating a first draft, invest some of that saved time in verification. The net should be better work, not just faster work.

4. Data Security and Confidentiality

AI tools — particularly cloud-hosted ones — process your inputs on remote servers. What you type into them may be logged, stored, or used for training.

• **Never enter personal data, participant data, or any data covered by ethics approval into a cloud-hosted AI tool** unless the tool's data processing agreement explicitly permits it and your ethics approval covers it. This includes names, ID numbers, health data, survey responses, and any data that could identify a person.

• **Never enter unpublished research results, draft papers, or confidential grant material into a public AI tool** without considering the intellectual property implications. If in doubt, ask Jonathan.

• **Be cautious with proprietary code, API keys, credentials, and access tokens.** Do not paste these into AI assistants.

• **Prefer local or self-hosted models** when working with sensitive data. If you need help setting this up, ask in the lab.

• **Understand your tool's data policy.** Know whether the tool you are using retains your inputs, uses them for training, or shares them with third parties. OpenAI, Anthropic, Google, and others have different policies depending on the product tier (free vs. paid vs. API). Read them.

UCT's Protection of Personal Information Act (POPIA) obligations apply to all data processing, including processing via AI tools.

5. Intellectual Ownership and Authorship

• AI tools cannot be authors. They do not take responsibility, they cannot consent to publication, and they cannot be held accountable. This is consistent with COPE, ICML, NeurIPS, and most major publishers' current positions.

• If an AI tool made a substantial contribution to a piece of work, acknowledge it — but in the methods or acknowledgements, not the author list.

• Be honest with yourself about the extent of your own contribution. If an AI wrote 90% of your literature review and you changed a few sentences, that is not your literature review. The goal is to use AI to help you produce better work, not to produce work you do not understand.

• Discuss authorship and AI-use norms at the start of any collaboration, not at submission time. Different collaborators, venues, and funders may have different expectations.

• Be aware that text generated by AI may inadvertently reproduce copyrighted material or closely paraphrase existing work. Treat AI-generated text with the same plagiarism vigilance you would apply to any other source.

6. Developing and Deploying AI Systems

As an AI research group, some of us are not just using AI tools but building them. Additional responsibilities apply.

Before you build

• Consider the potential misuse of what you are creating. This does not mean every project needs a formal risk assessment, but the question "how could this be misused?" should be part of your thinking from the start, not an afterthought bolted onto the broader impact statement.

• If your work involves human subjects, human data, or could affect real people, ensure appropriate ethics approval is in place.

• If your work involves dual-use capabilities — systems that could be used to deceive, surveil, manipulate, or cause harm — discuss this with Jonathan early, preferably before significant development effort.

During development

• Document your models, datasets, and training procedures thoroughly. This is good science, but it is also an ethical obligation: others need to be able to understand, scrutinise, and reproduce what you have done.

• Be transparent about the limitations of your system. Do not overstate capabilities in papers, demos, or communications.

• Test for bias and failure modes, not just accuracy on benchmark tasks.

• Use version control for everything, including model configurations, hyperparameters, and data preprocessing pipelines.

Before you release

• Think carefully about what you release publicly. Open-sourcing a model or dataset is generally a good thing, but not always. Consider whether your release needs access controls, usage guidelines, or a licence that restricts harmful applications.

• If you are releasing a model that generates text, images, or other media, consider whether it could be used to produce misinformation, non-consensual content, or other harmful outputs — and what mitigations you can put in place.

• Discuss release plans with Jonathan and, where relevant, with the broader lab.

7. Skill Development and Learning

One of the less obvious risks of AI tools is that they can undermine the learning process itself. This matters especially in a research group where many of us are at an early career stage.

• **Understand before you automate.** If you have not yet learned how to do something, doing it yourself first — even badly — builds understanding that you will not get from reading AI-generated output. Use AI to check your work or to learn from, not to skip the learning.

• **Be honest about your own capabilities.** If AI tools are masking gaps in your knowledge, those gaps will eventually surface — in a viva, a job interview, a debugging session, or a conversation with a collaborator. It is better to know what you do not know.

• **Teach each other.** If you discover a useful AI workflow, share it. If you discover a failure mode, share that too. The lab benefits from collective knowledge about these tools.

• **Develop critical evaluation skills for AI outputs.** The ability to quickly identify when an AI is wrong, biased, or hallucinating is itself a valuable skill — and one that is increasingly important in the field we work in.

8. Broader Impact and Societal Responsibility

We work on AI. This comes with a responsibility to think about the societal implications of our research, not just its technical contributions.

• Take broader impact sections seriously. They should not be perfunctory paragraphs added to meet a submission requirement. Think genuinely about who benefits from your work, who might be harmed, and what assumptions you are making.

• Engage with communities affected by your research where possible and appropriate. This is especially important in the South African and African context, where the impacts of AI may be distributed very differently from the Global North contexts in which most AI research is framed.

• Stay informed about AI policy and governance developments, particularly in South Africa and on the African continent. As AI researchers based in Africa, we have both a perspective and a responsibility that many labs in the Global North do not.

• Be willing to say "we should not build this" if that is the honest conclusion. Not every technically interesting project is a socially responsible one.

9. Keeping Up to Date

The AI landscape changes rapidly. Norms, capabilities, and risks that seem settled today may look very different in six months.

• This document will be reviewed at least once per year, alongside the main Code of Conduct.

• If you encounter a situation that this document does not cover, raise it. That is how living documents improve.

• If a venue, funder, or collaborator has AI-specific requirements that go beyond what is described here, follow the stricter standard.

• When in doubt about anything in this document, ask. Jonathan's door (physical or virtual) is always open for these conversations.

10. Acknowledgement

By joining Shocklab, you agree to uphold the principles in this document alongside those in the main Code of Conduct. Both documents may change, and you have a voice in those changes.

This document draws on the Montréal Declaration for Responsible AI, the G7 AI Code of Conduct, the EU Code of Practice for General-Purpose AI, Anthropic's Responsible Scaling Policy, the NeurIPS broader impact statement framework, the University of Rochester's Responsible Use of Generative AI in Research policy, and the Core Principles of Responsible Generative AI Usage in Research proposed by Sallam et al. (2025). Adapted for Shocklab at the University of Cape Town by Claude Cowork with Opus 4.6, Jonathan Shock and Shocklab students.