



How and why ownership matters for human-AI collaboration.



THE UNIVERSITY OF
SYDNEY

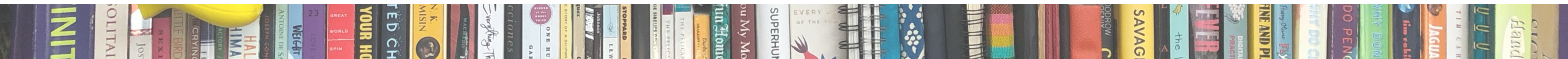
Who Owns the Future* of AI?

equitable access

legal rights

cognitive engagement

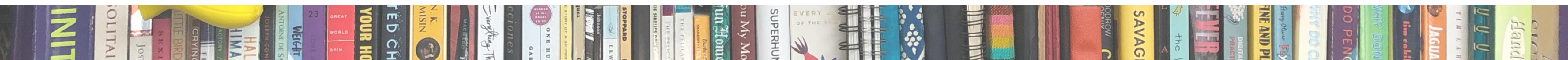
accountability

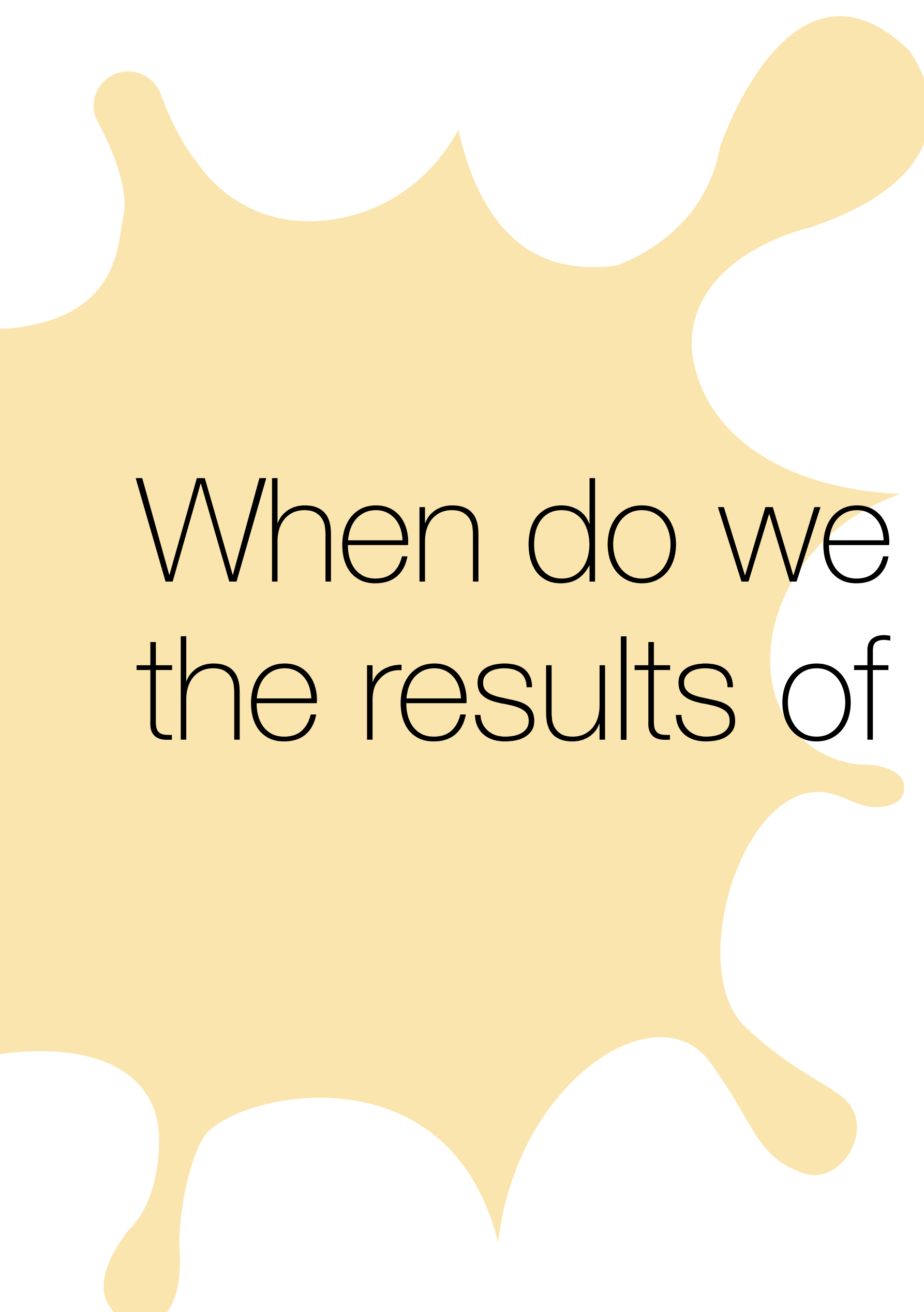


Who Owns the Future* of AI?

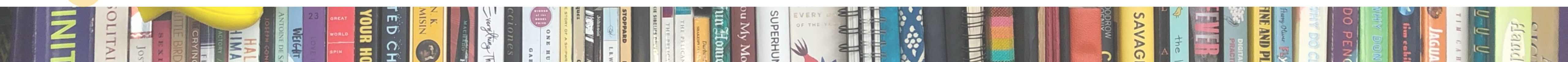
What does it mean to own AI writing?

1. Who owns the outputs of AI-assisted writing
2. Who owns the model that assists you?





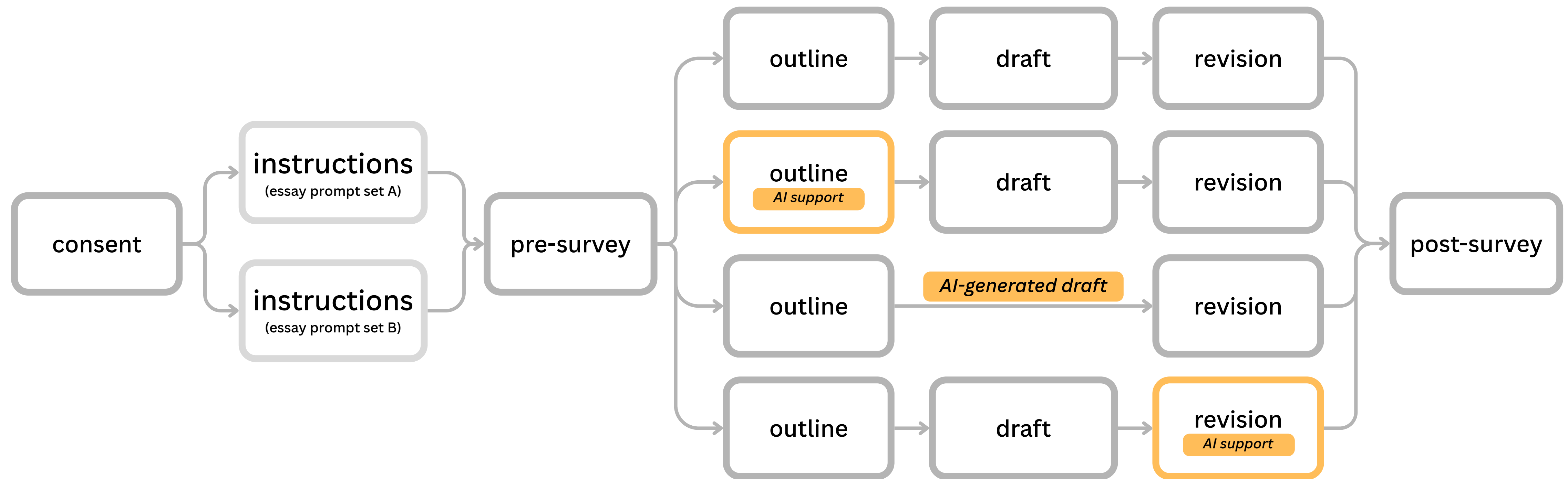
When do we
the results of



AI-assisted writing decreases ownership

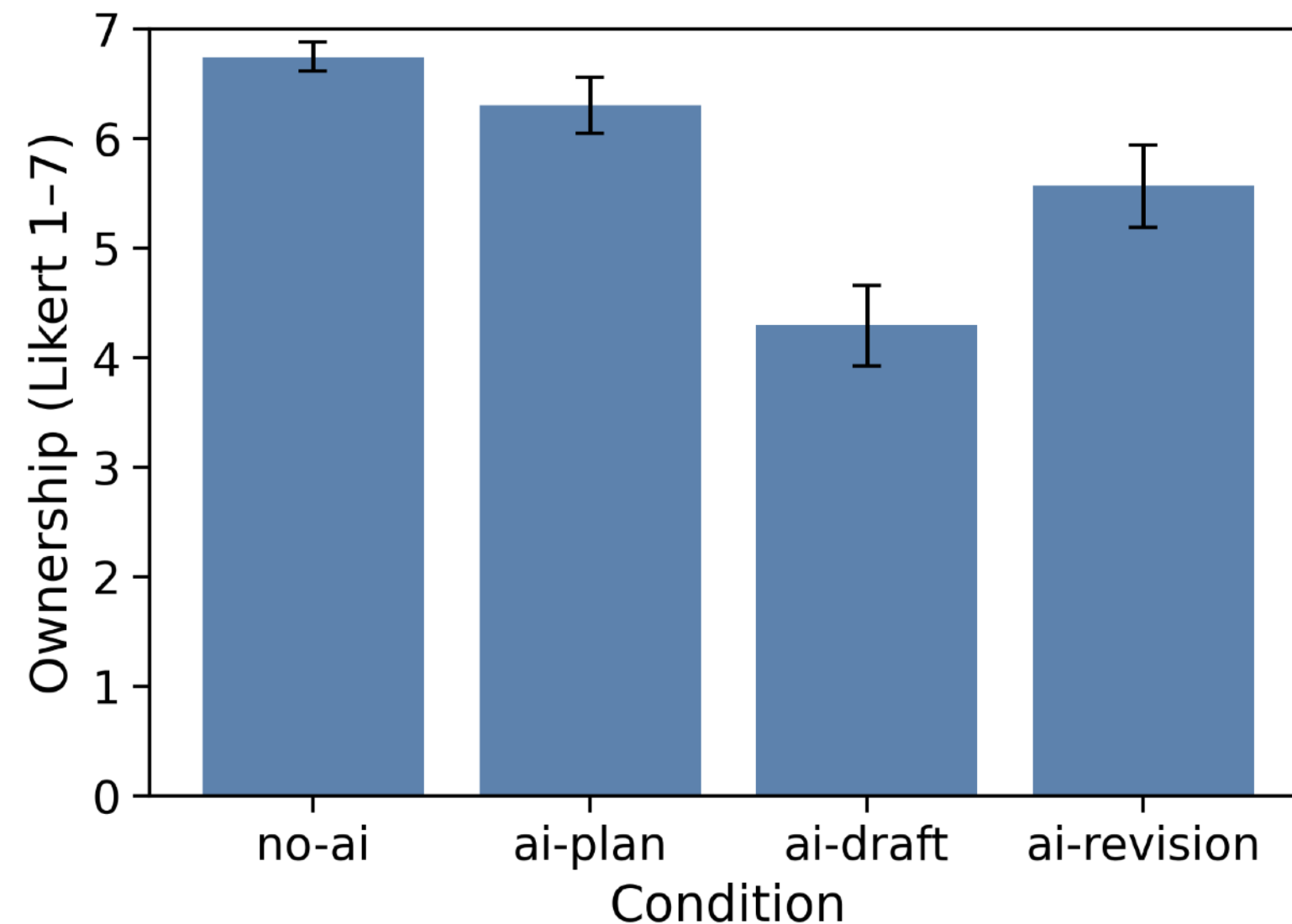
(wait, but why?)





1. How does the stage of AI support change **ownership**?
2. How does the stage of AI support change **text and idea attribution**?

1. How does the stage of AI support change **ownership**?



Participants talked about emotional resonance, self-efficacy, and their sense of voice or expression:

- I wrote “*from the heart*”, it was a topic I “*hold dear*” or am “*extremely passionate about*”
- “*I feel I am very capable of discussing this topic since I have experienced it over the past few years*”
- “*the outline was mine, but the voice, the writing style, was not*”

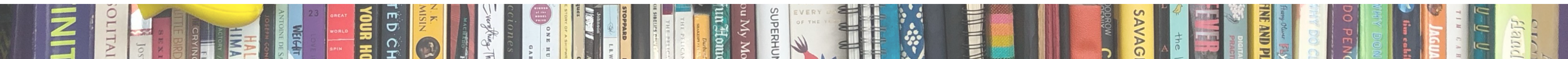
** *no AI v. avg AI*

** *AI draft v. (AI plan, AI revision)*

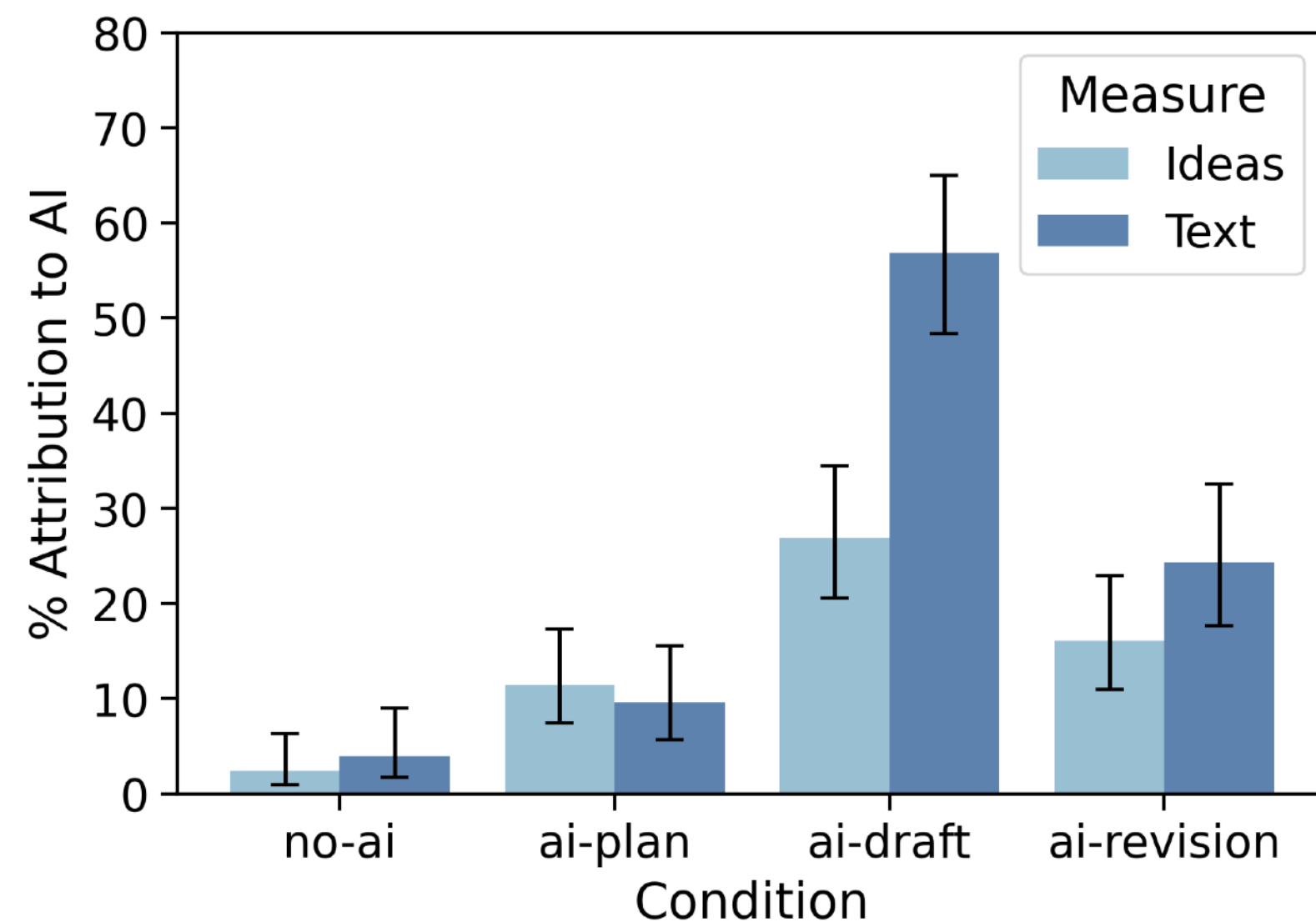
* *AI plan v. AI revision*

** $p < 0.001$

* $p < 0.005$



2. How does the stage of AI support change **ideas & text attribution**?



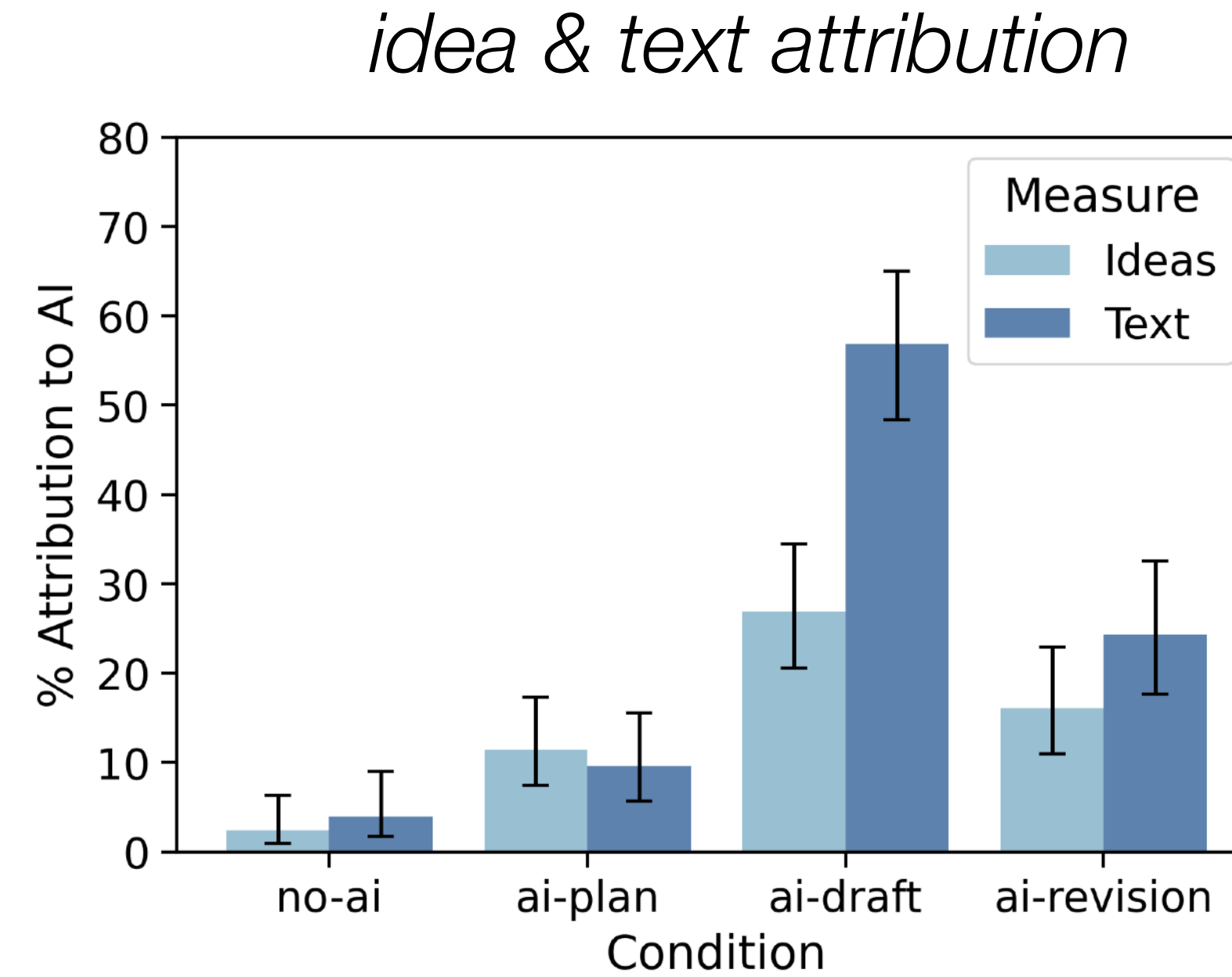
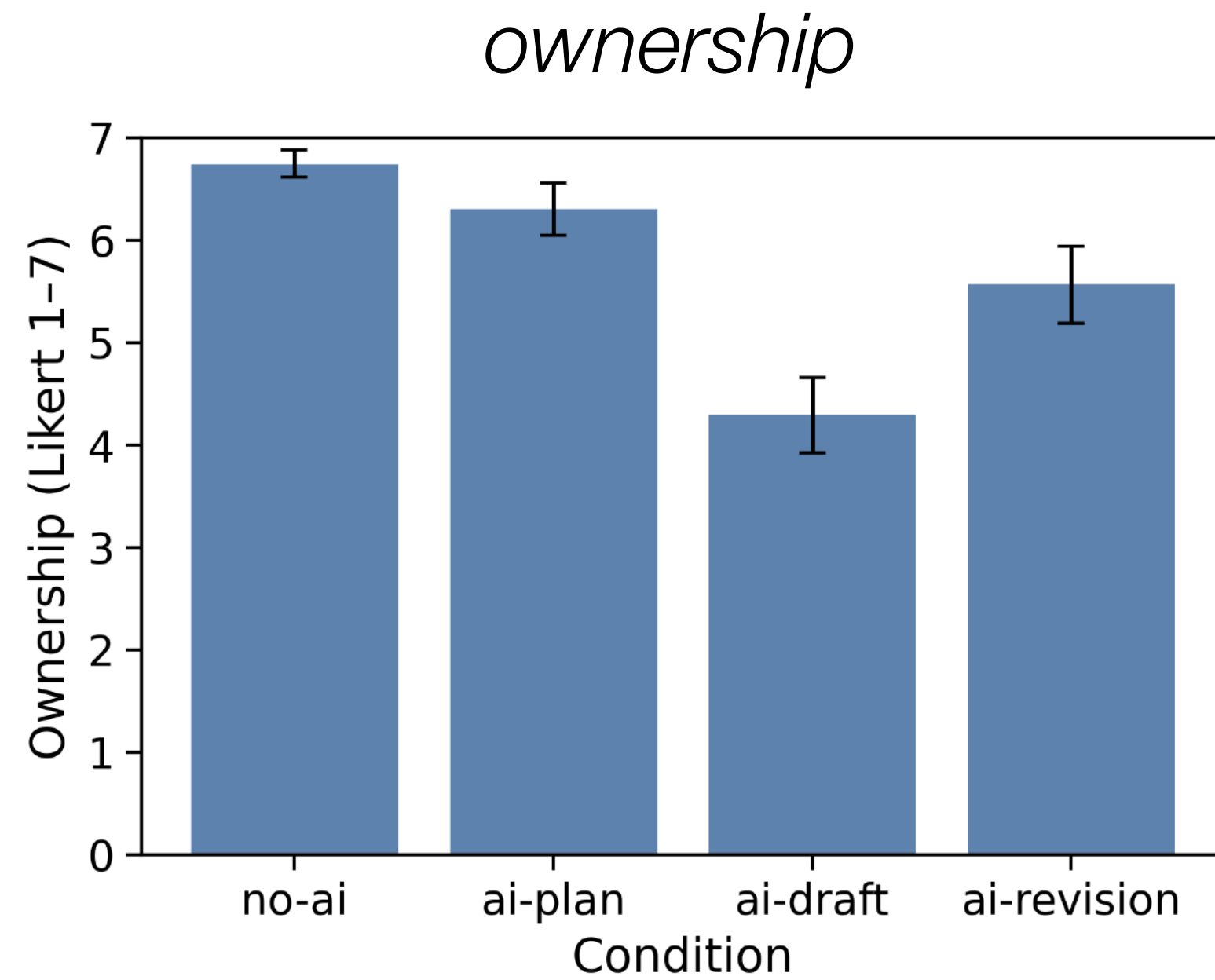
Participants wrote about the myriad ways in which ideas where accepted, ignored, integrated, changed, or added to by either themselves or the AI assistance:

- *"the AI did come up with outline ideas, but they matched/confirmed what I had already thought"*
- *"I seeded the ideas and they were fleshed out, but not changed"*
- *"The AI offered several points that I was simply not aware of. While I chose whether or not to add them, adding them seemed like a no-brainer - but they were also not my ideas, borne of my own research."*

** *no AI v. avg AI*

** *AI draft v. (AI plan, AI revision)*
AI plan v. AI revision

** $p < 0.001$



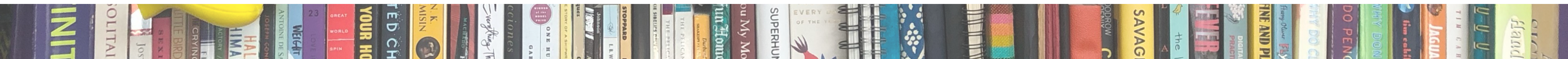
1. AI plan introduced *less* ideas than expected; AI draft introduced *more*.
2. AI draft is dwarfed by both AI text *and* ideas.

Writing as a mode of learning.

"I felt like the essay was completely written by AI but the ideas were mine. I feel as though if I were to write the essay it would take me much longer and would require me to have a deep understanding on the topic. When I write essays, the process of writing the essay helps me determine my final decision on the matter, the more I write the more I allow myself to think and reason with the evidence I found."

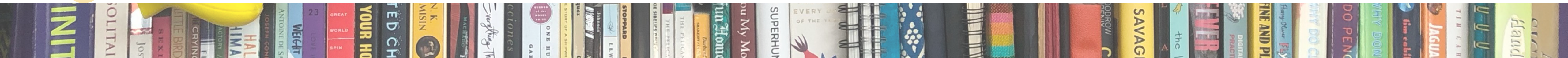
In the learning sciences, there is an idea of “desirable difficulties” or “productive struggle”.

- When is the work of writing worth the investment?
- When is the efficiency AI support worth the time saving?
- When should AI make writing harder, rather than easier?

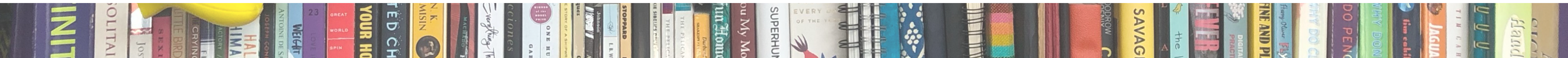




Who owns LLM training data?



What is Books3?



2.3 Books3

Books3 is a dataset of books derived of the contents of the Bibliotik project made available by Shawn Presser (Presser, 2020). Bibliotik consists of a mix of fiction and nonfiction books and is almost an order of magnitude larger than our next largest book dataset (BookCorpus2). We included Bibliotik because books are invaluable for long-range context modeling research and coherent storytelling.

The Pile: An 800GB
Dataset of Diverse Text
for Language Modeling

Shawn Presser. 2020. Books3. <https://twitter.com/theshawwn/status/1320282149329784833>.

 **Shawn Presser** ✓
@theshawwn

Suppose you wanted to train a world-class GPT model, just like OpenAI. How? You have no data.

Now you do. Now everyone does.

Presenting "books3", aka "all of bibliotik"

- 196,640 books
- in plain .txt
- reliable, direct download, for years: the-eye.eu/public/AI/pile...

thread 📌

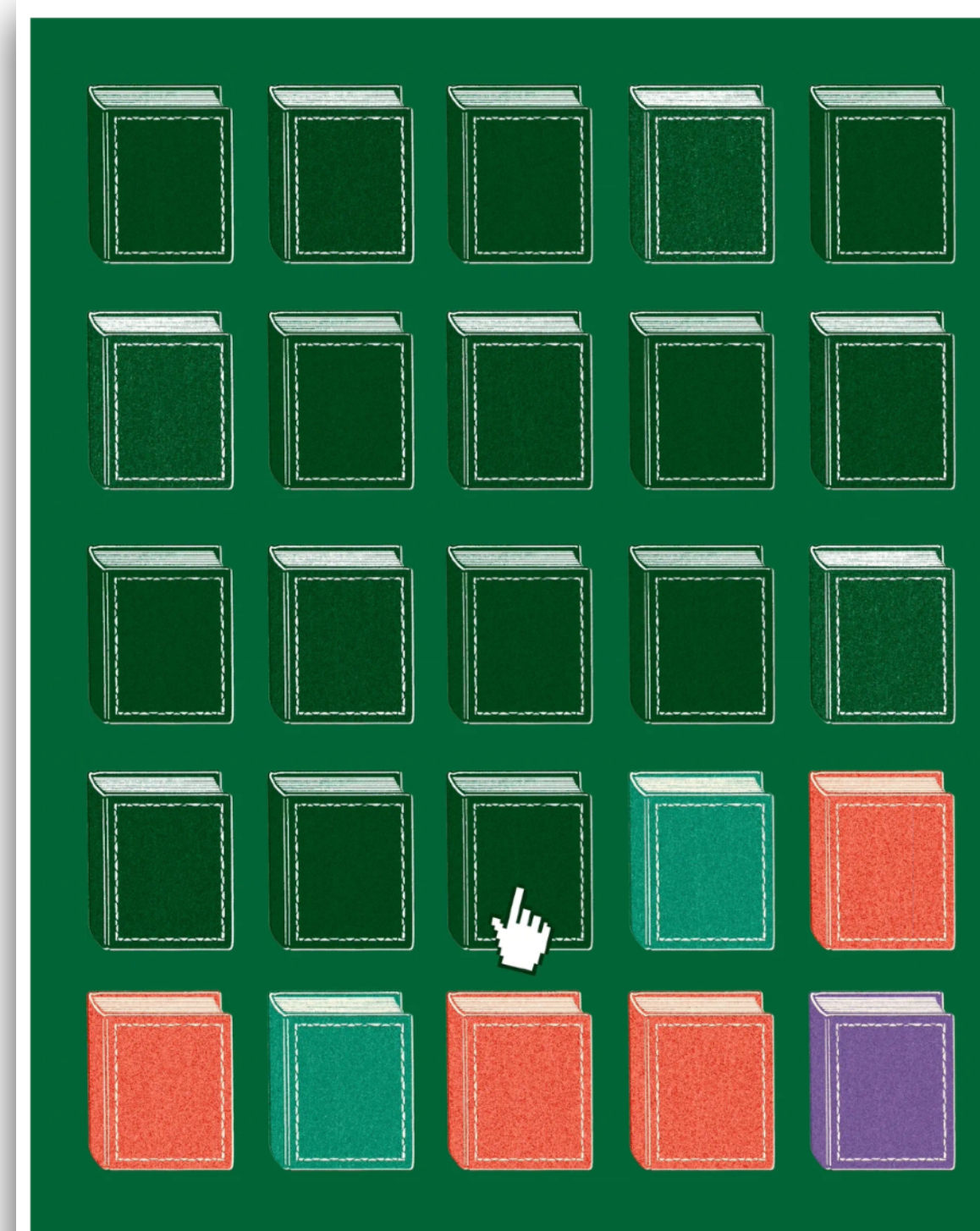


Illustration by Joanne Imperio / The Atlantic. Source: Getty.

TECHNOLOGY

THESE 183,000 BOOKS ARE FUELING THE BIGGEST FIGHT IN PUBLISHING AND TECH

Use our new search tool to see which authors have been used to train the machines.

By Alex Reisner

Why might a writer be for or against the use of their work as training data for LLMs?

RQ1: How do writers reason about the real or hypothetical use of their work as training data for LLMs?

RQ2: Under what conditions, if any, would they consent to their work being used as training data for LLMs?

Creative Writers' Attitudes on Writing as Training Data for Large Language Models

Katy Ilonka Gero
Department of Computer Science
Harvard University
Cambridge, Massachusetts, USA
katy@g.harvard.edu

Nayun Eom
Department of Sociology
Harvard University
Cambridge, Massachusetts, USA
neom@fas.harvard.edu

Meera Desai
School of Information
University of Michigan
Ann Arbor, Michigan, USA
madesai@umich.edu

Jack Cushman
Harvard Library Innovation Lab
Harvard University
Cambridge, Massachusetts, USA
jcushman@law.harvard.edu

Carly Schnitzler
University Writing Program
Johns Hopkins University
Baltimore, Maryland, USA
cschnit1@jh.edu

Elena L. Glassman
SEAS
Harvard University
Allston, Massachusetts, USA
glassman@seas.harvard.edu

Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3706598>. 3713287

Abstract

The use of creative writing as training data for large language models (LLMs) is highly contentious and many writers have expressed outrage at the use of their work without consent or compensation. In this paper, we seek to understand how creative writers reason about the real or hypothetical use of their writing as training data. We interviewed 33 writers with variation across genre, method of publishing, degree of professionalization, and attitudes toward and engagement with LLMs. We report on core principles that writers express (support of the creative chain, respect for writers and writing, and the human element of creativity) and how these principles can be at odds with their realistic expectations of the world (a lack of control, industry-scale impacts, and interpretation of scale). Collectively these findings demonstrate that writers have a nuanced understanding of LLMs and are more concerned with power imbalances than the technology itself.

CCS Concepts

• Human-centered computing → Empirical studies in HCI; HCI theory, concepts and models; • Computing methodologies → Natural language generation.

Keywords

Large language models, natural language generation, creative writing, writing assistance, data, archival practices.

1 Introduction

As large language models have gained wide-spread usership, so have investigations into the text used to train them [10, 17, 27, 33, 37, 61, 65]. There has been a particularly swift and negative response to the use of creative writing, such as novels, memoirs, and longform journalistic books, as training data, as the authors of this writing express outrage at the use of their work without consent or compensation [9]. Such frustration has led to dozens of lawsuits against technology companies like OpenAI, Microsoft, and Google, alleging that this use of writing goes against U.S. copyright law [1].

Not everyone agrees that the use of creative writing for training data is illegal, or even immoral. In comments to the U.S. Copyright Office, The Author's Alliance, whose mission is to protect the interests of authors who write for the public, expressed concern about using writing as training data for LLMs.

Principles

Creative Chain

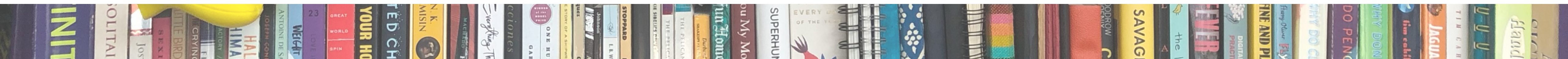
Respect

Human Element

"It's just being a link in the **creative chain**. I made something. Somebody else saw that, and they got a spark of an idea to do something else, and they want to build off of it. We're all building off of each other constantly." (P20)

"I think it just comes again to **respecting** the works and **respecting** the people. Because I think that's probably my main issue with a lot of AI stuff is just the complete lack of respect for the works that they have been using to profit off." (P18)

"Honestly, I think a lot of the reason that people like reading other people's stuff is because there's the **human element** there. You're looking at one guy's perspective of an insane situation ... It's the project of one person's mind." (P13)



Principles

Creative Chain

Respect

Human Element

Realistic Expectations

Lack of Control

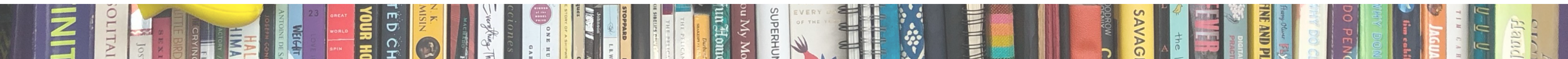
Industry Impacts

Interpretation of Scale

"One of the things that we're all really feeling right now is a **lack of any options** in this. When I found out that my work was in Books3 I emailed my agent right away. And I was like, I know there's like nothing you can do, I just want you to know that this is a thing, and that I don't like it." (P20)

"We could easily be **putting artists out of work**. And what does that mean for them as artists and humans, [who need] to put food on their table, but also like, what does that mean for culture and society?" (P9)

"Cause I'm trying to think, okay, upsides and downsides. It's hard to pinpoint one, because again, I'm a **little drop of water in a huge ocean**. I don't have a distinct style, so I don't... I don't really know." (P24)



Principles

Creative Chain

Respect

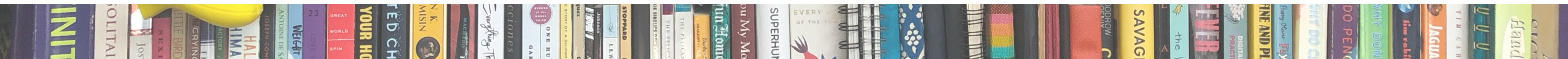
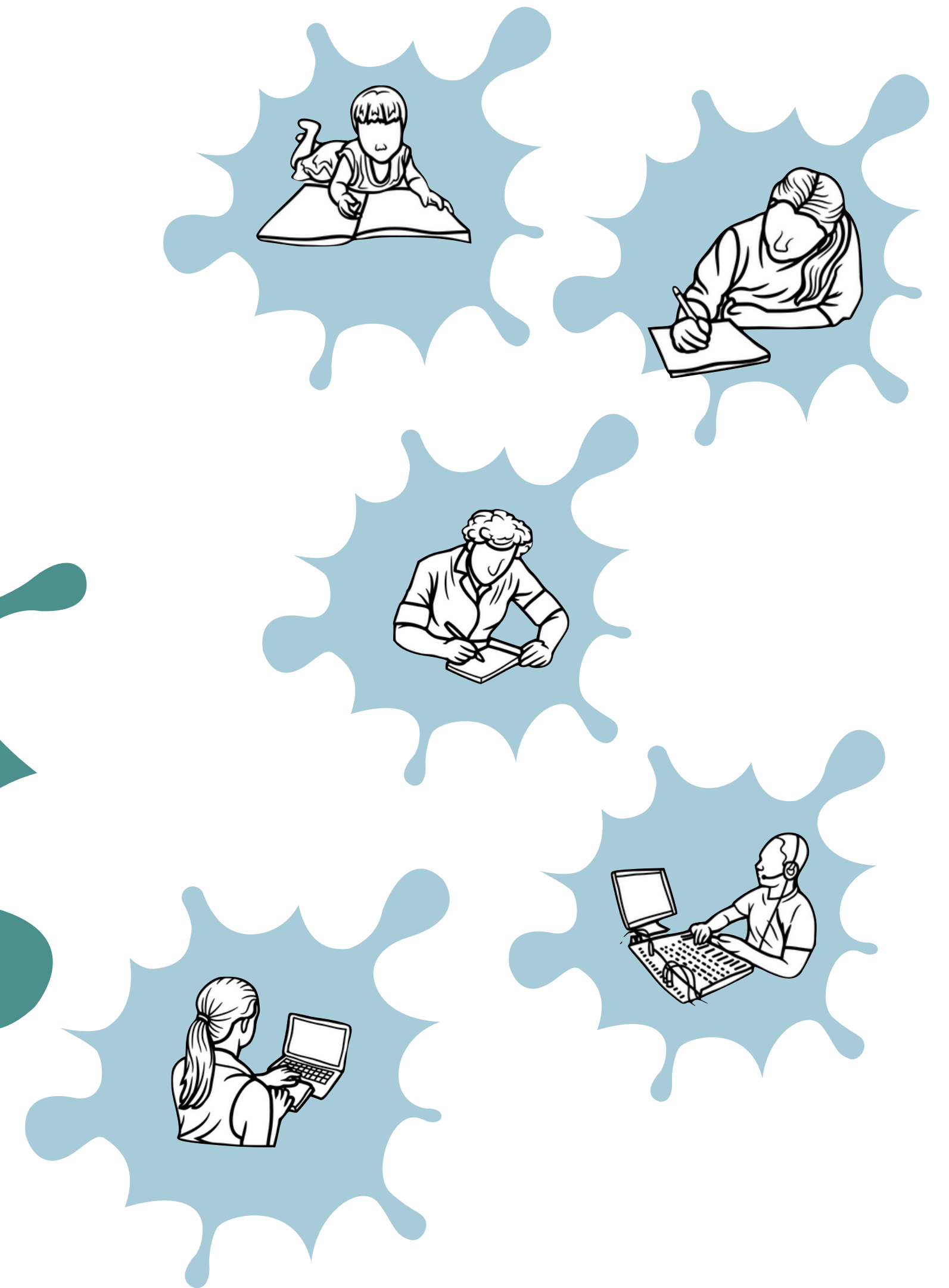
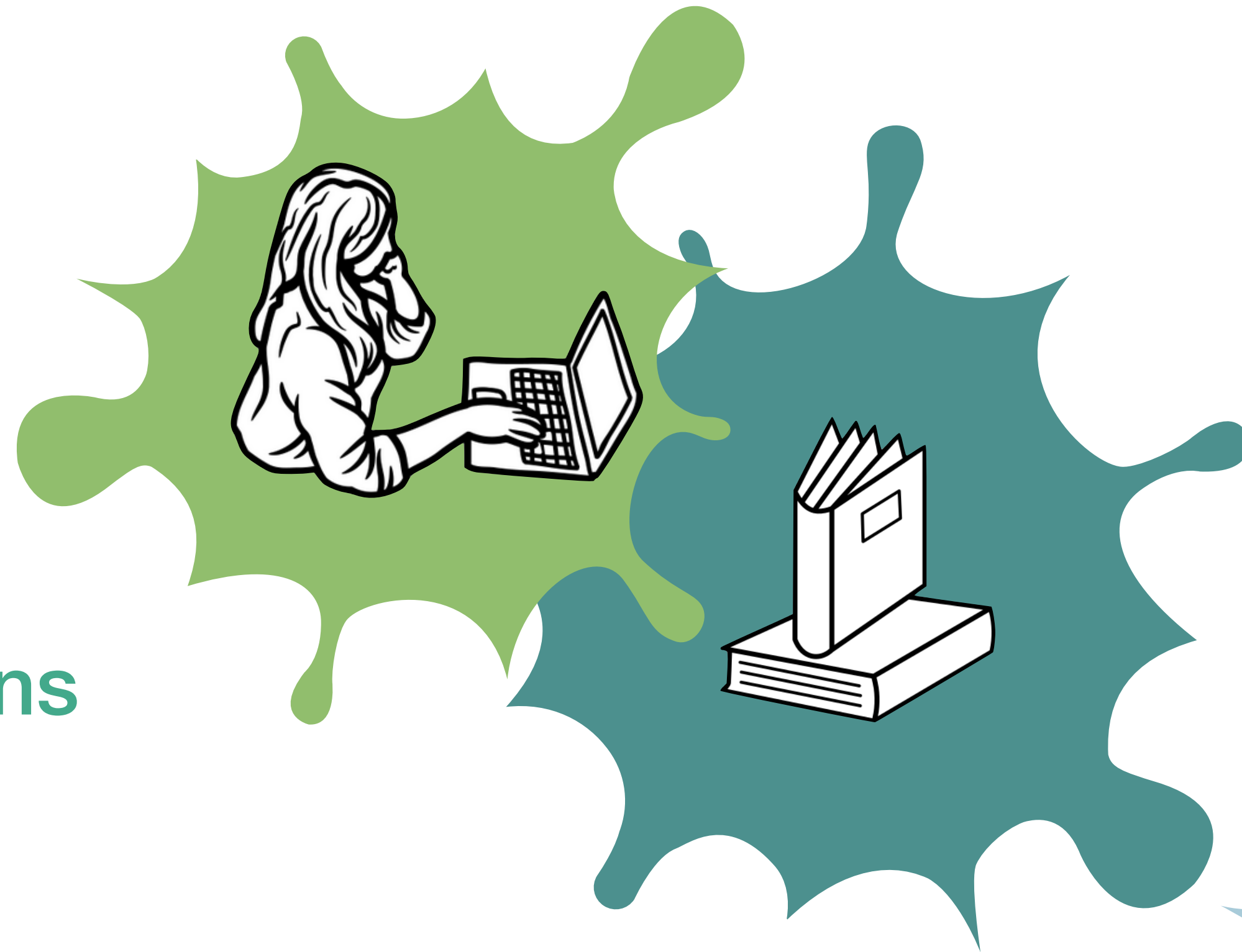
Human Element

Realistic Expectations

Lack of Control

Industry Impacts

Interpretation of Scale



Principles

Creative Chain

Respect

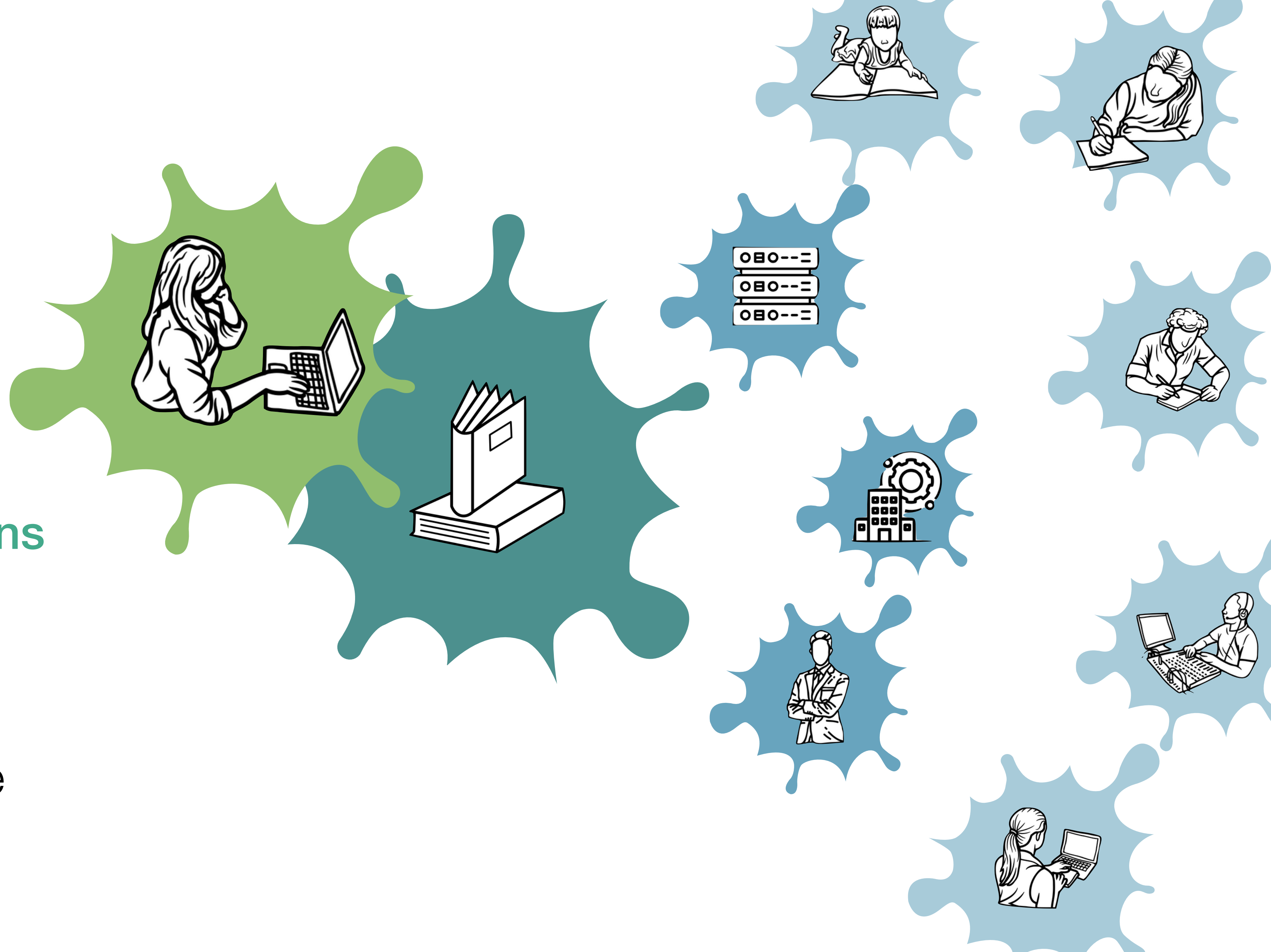
Human Element

Realistic Expectations

Lack of Control

Industry Impacts

Interpretation of Scale



Principles

Creative Chain

Respect

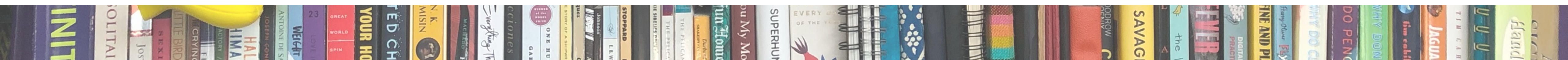
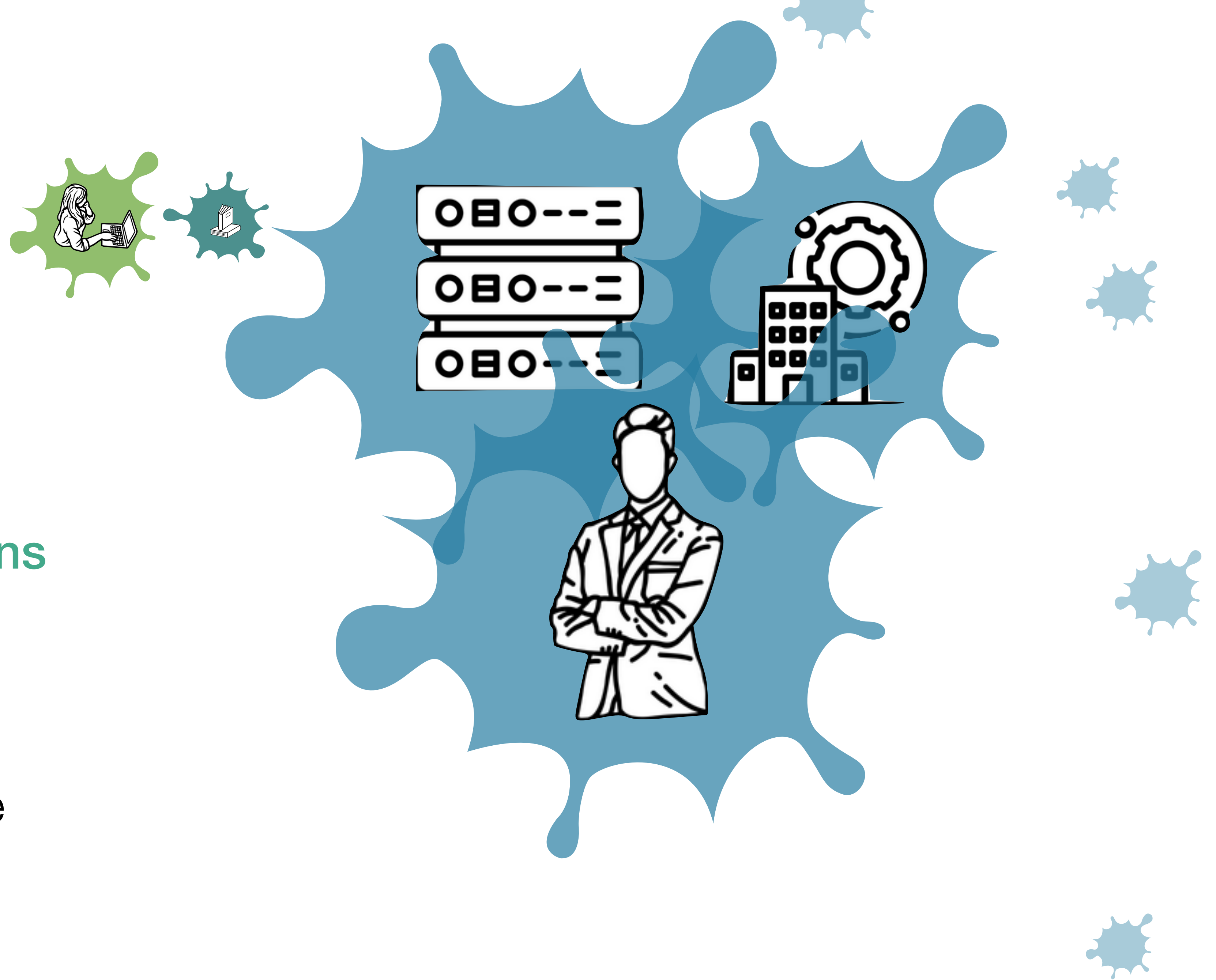
Human Element

Realistic Expectations

Lack of Control

Industry Impacts

Interpretation of Scale



enter into AI

[illegible]

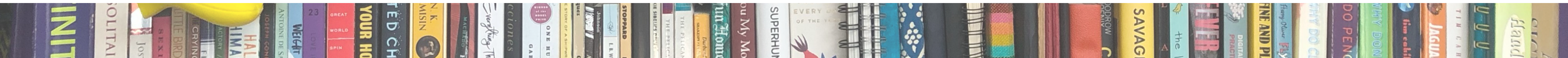
Scholarly Values

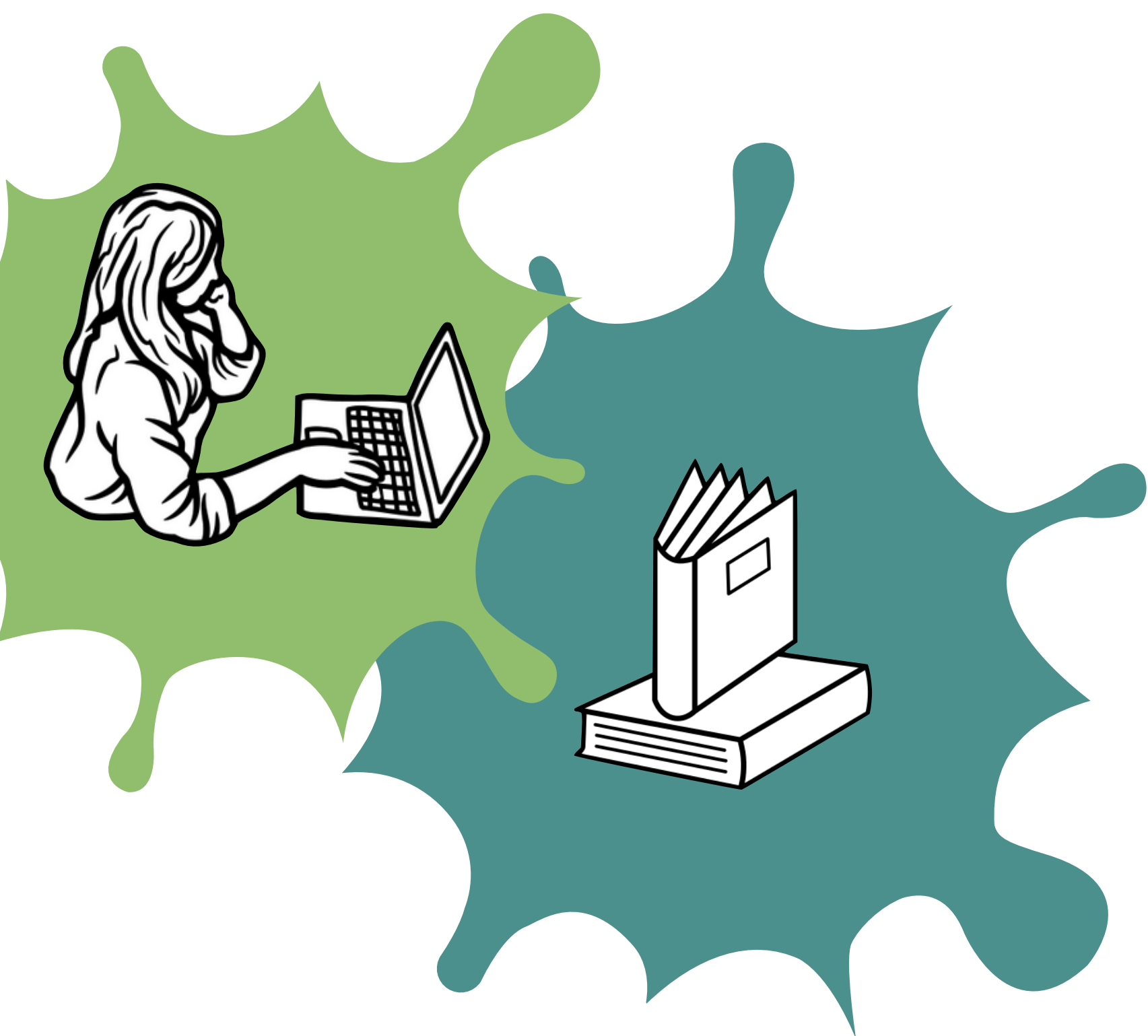
“Yes, I view it as an opportunity for the work to be integrated into a **universal knowledge bank** that can be synthesized with other works for positive benefit.” 378A

“**Regulations** that can protect the livelihoods of authors and publishers are useful even if difficult to devise and difficult to monitor.” 484B

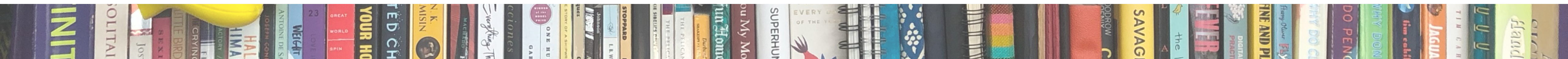
Exploitation & Harm

“They are short-sighted deals based on near-term gains for publishers in exchange for greatly weakening publishers ... in the mid-term, while also potentially **threatening its very foundations** and existence ... in the long-term.” 609B



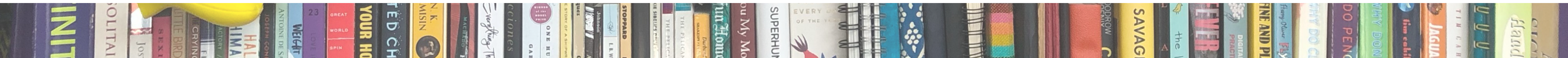


What is the
role of LLMs?
**And who gets
to decide?**

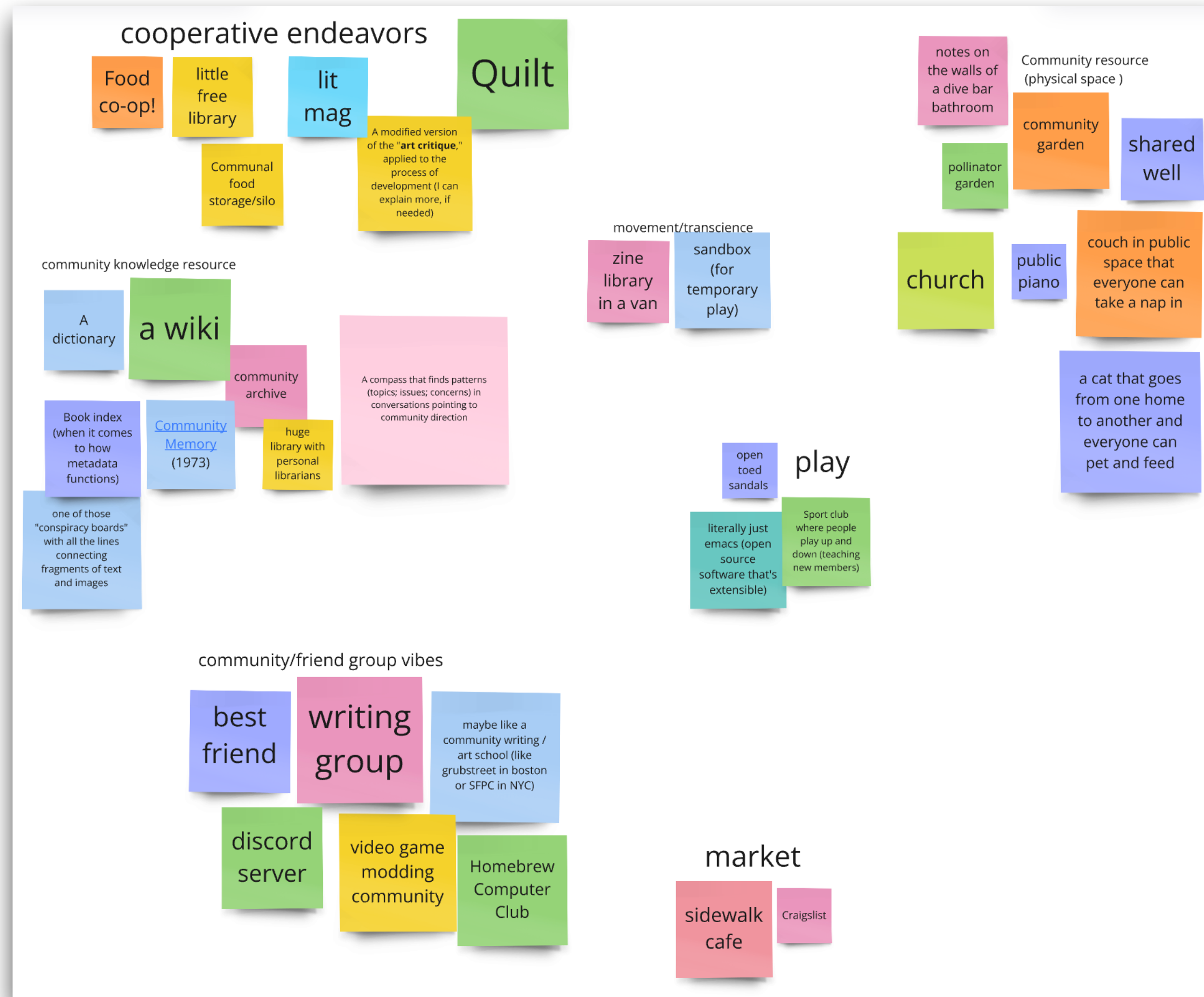


“You should **work with other presses** to develop and implement a unified, expert-informed, non-exploitative strategy such that tech companies can't play presses off against each other.” 670B

Are we just talking about... collective action?



Metaphors for Models



• Think through the 4 C's:

- Consent
- Control
- Credit
- Compensation

Metaphors for Models

Quilt

movement/transcience

zine library in a van

sandbox (for temporary play)

open toed sandals

literally just emacs (open source software that's extensible)

play

Sport club where people play up and down (teaching new members)

notes on the walls of a dive bar bathroom

Community resource (physical space)

community garden

pollinator garden

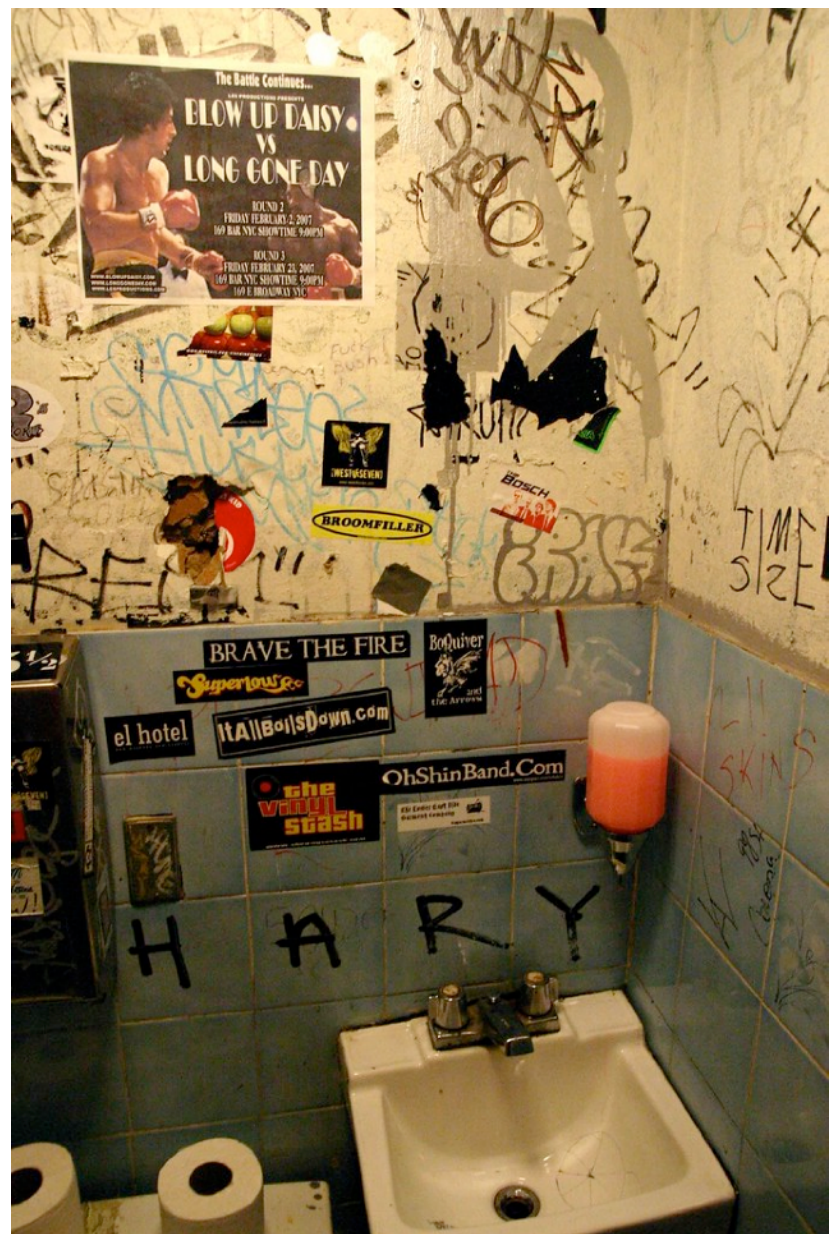
shared well

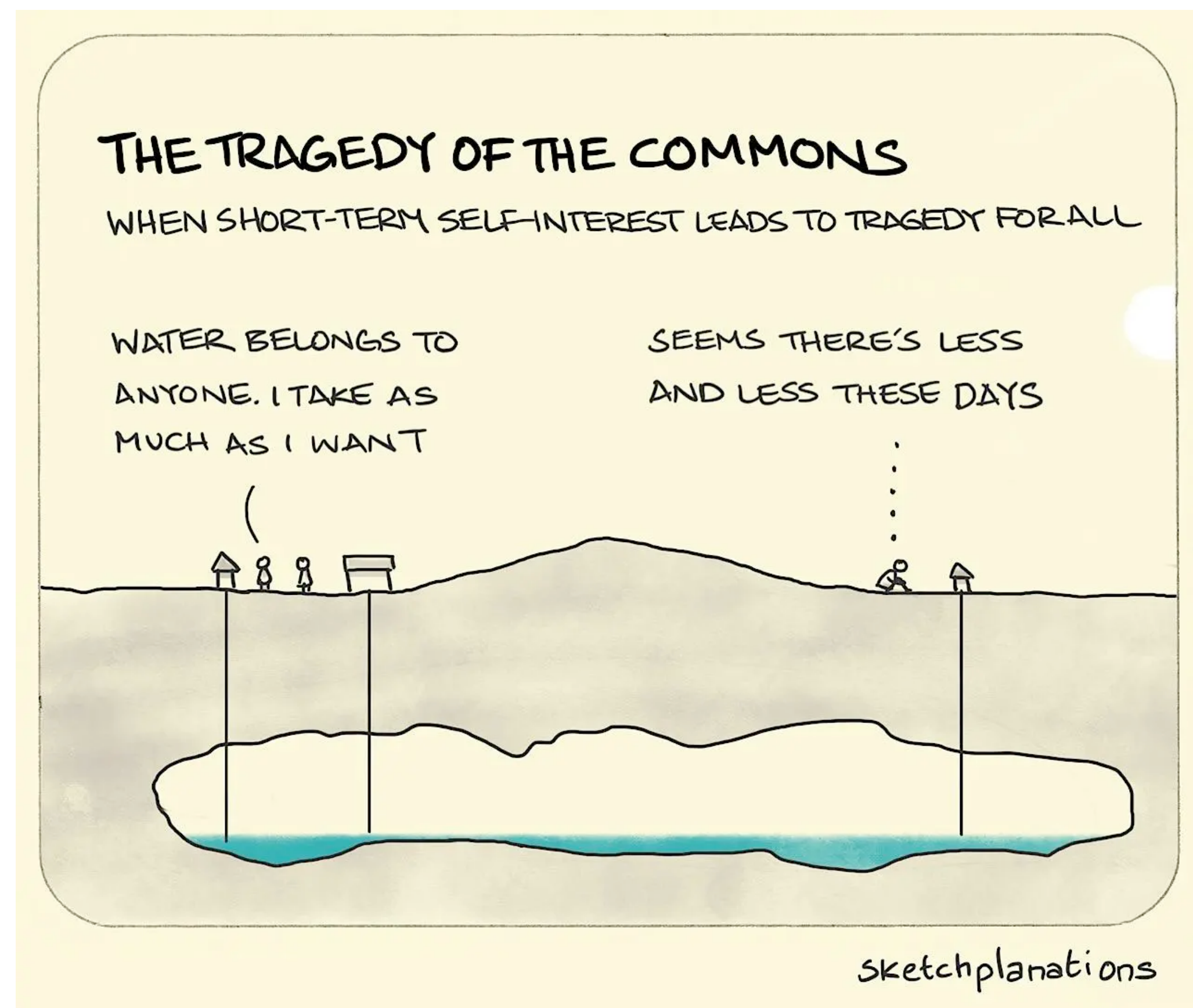
church

public piano

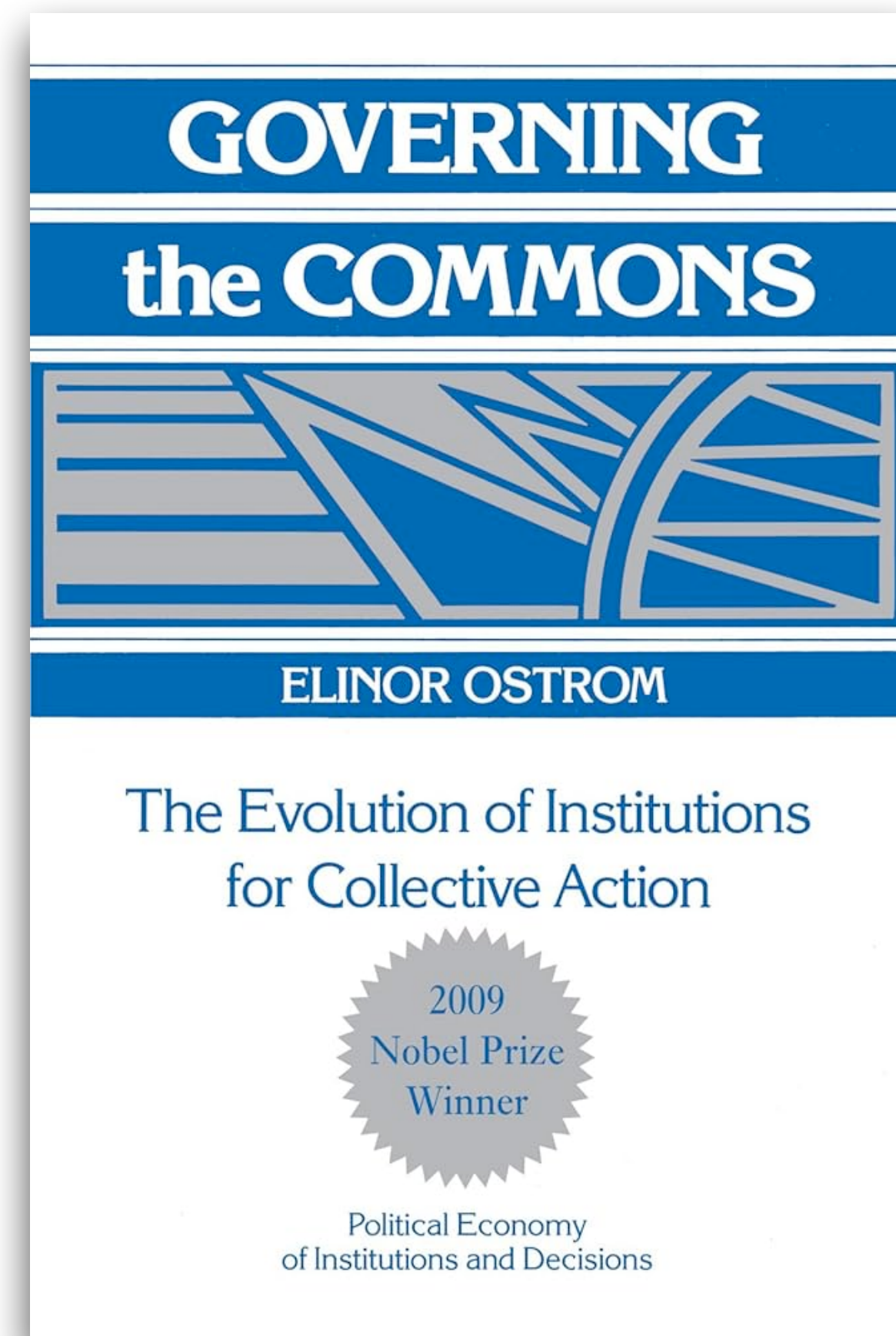
couch in public space that everyone can take a nap in

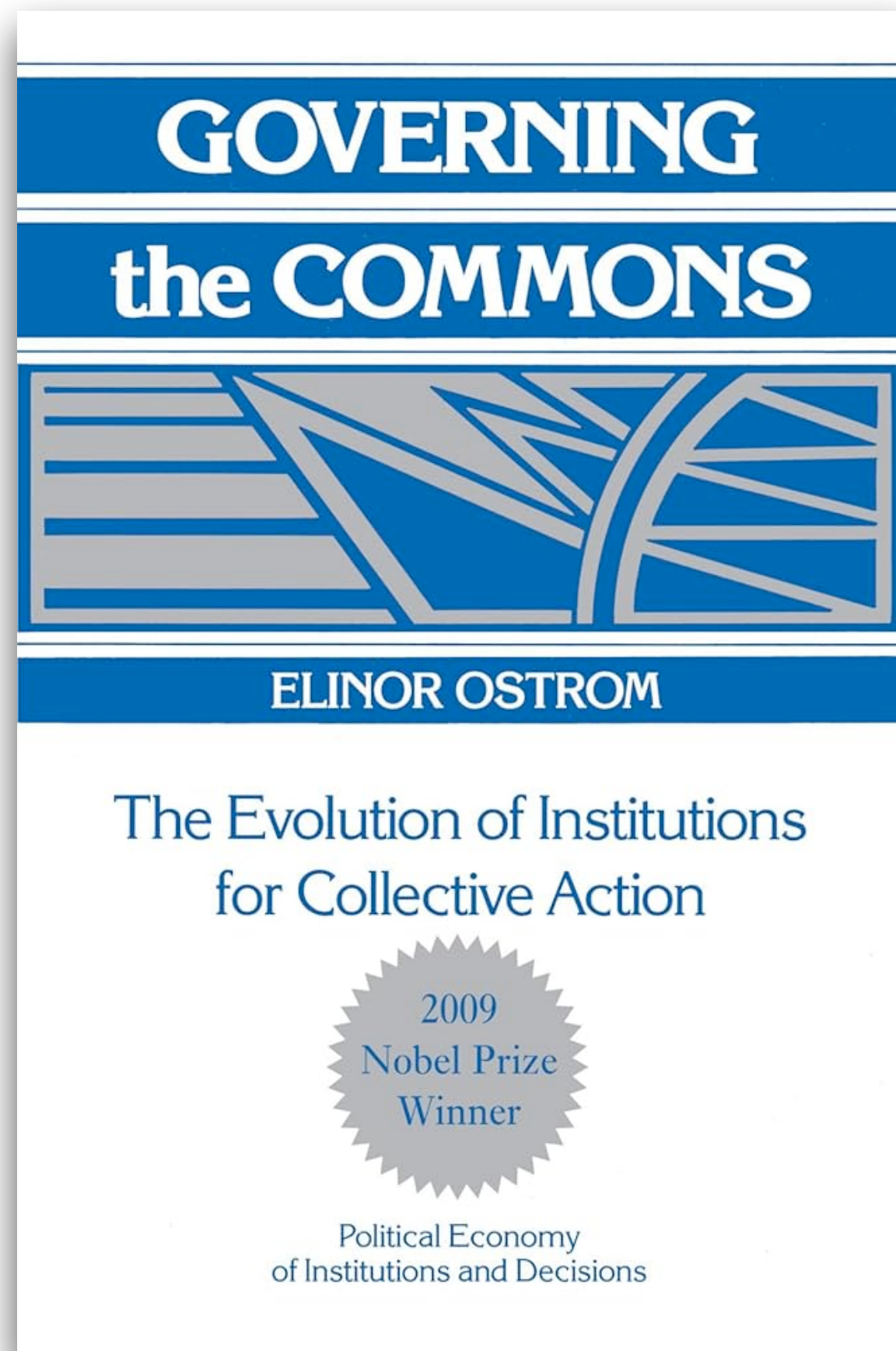
a cat that goes from one home to another and everyone can pet and feed



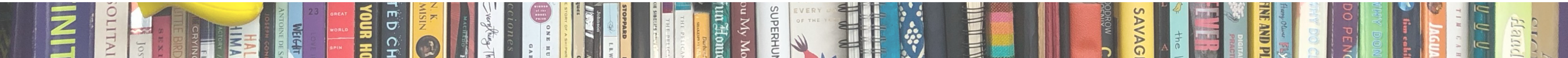


<https://sketchplanations.com/the-tragedy-of-the-commons>





- digital commons have poorly defined boundaries
- unclear what local knowledge means for digital commons
- digital commons lack agreed-upon monitoring protocols
- many problems come from externalities



Who Owns the Future* of AI?

How and why ownership matters for human-AI collaboration.

- ▶ When should AI make writing easier? Can we create “productive struggle”?
- ▶ How should we balance openness while protecting writing-as-data?
- ▶ Can we make AI models by and for creative communities?

Dr. Katy Ilonka Gero
Lecturer at the University of Sydney
katy.gero@sydney.edu.au



THE UNIVERSITY OF
SYDNEY