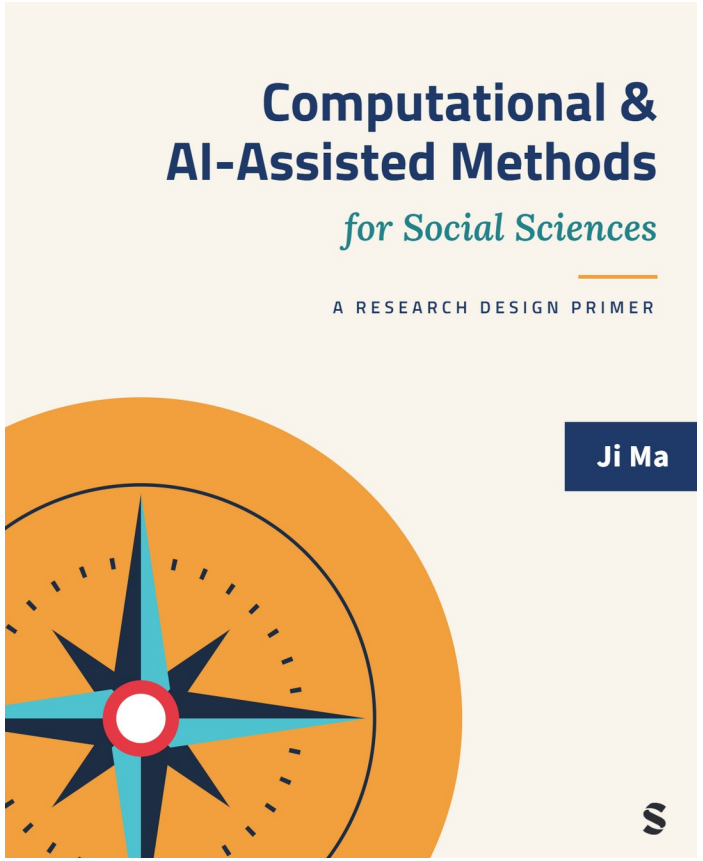
## Computational & AI-Assisted Methods for Social Sciences

*A Research Design Primer*




**Chapter 11**

# Ethical Concerns in Computational and AI-Assisted Research

Ji Ma

cssprimer.org

# Learning objectives

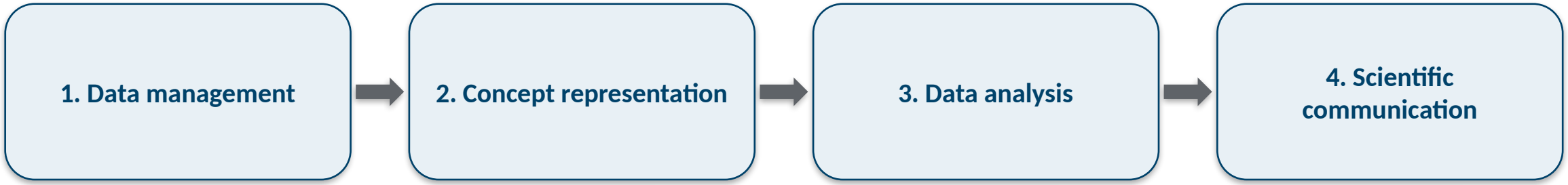
*By the end of this chapter, you can:*

- Identify the ethical concerns at each design stage, and at the AI infrastructure layer cutting across them
- Recognize why the biomedical framework behind IRB review fits CSS poorly, and what an embedded alternative looks like
- Apply safeguards for born-digital data — contextual integrity, consent, tiered access — and for AI workflows: disclosure, traceability, oversight
- Develop a project-level ethical practice integrated with your workflow, audits, and reproducibility commitments

**Ethics in CSS is a property of how a project is designed, not an addendum applied at the end.**

# Every stage raises its own ethical question






Whose context? Whose errors? Who can contest? What is hidden? — and an AI layer cutting across all four.

**Takeaway:** The ethics chapter is organized by the same four stages as the course.

§11.1

# Data management: privacy as contextual integrity

*Stop thinking of privacy as a property of a file; start thinking of it as a property of how data circulate.*

# The technical framing of privacy is part of the problem



- The standard move: anonymize, encrypt, add calibrated noise
- Technically compliant research can still be ethically tone-deaf
- A “de-identified” release of teens' mental-health messages can pass every checkbox

# Contextual integrity: what did the author intend?



- Working rule: data should travel no further than its intended context and audience (Hoser & Nitschke 2010)
- A public tweet is loud talk in a coffee shop, not a press release
- Platform settings mark the upper bound on visibility, not intended audience

**Takeaway:** Users cannot defend a boundary they do not know is being crossed; the judgment falls on you.

# Context collapse and the inferred-attribute problem



- Imagined audience vs. “whole actual audiences” (Williams et al. 2017)
- ~80% of surveyed Twitter users expect to be asked before publication; >90% want anonymity
- Platforms infer ideology, ethnicity, orientation the user never volunteered
- Publishing an inferred attribute discloses what was never shared

**Takeaway:** Reproducing a post in a paper collapses audiences the speaker never anticipated.

# Consent for subjects who never agreed to be subjects

§11.1

- Most CSS data was generated before the researcher arrived
- The modal move — “it was public” — asserts that scrapeable means usable
- Compensations: minimize collection, aggregate reporting, consult the affected community (Salganik 2017)

**Takeaway:** Human subjects are now data subjects; the protections must be reworked, not waived.

# Re-identification is a risk profile, not a status



- Zip code, birth date, and gender can single a person out
- CSS data are large, sparse, longitudinal: behavior over months is a fingerprint
- Model plausible attacks; hold sensitive joins in controlled environments; release aggregates or synthetic products

**Takeaway:** Anonymization is a probability of disclosure under attack, not a property of the dataset.

# Workflow is the enforcement mechanism



- Weak: the PI tells the team to drop identifiers
- Stronger: a documented workflow step names the columns to remove
- Strongest: an automated step selects only approved columns; deviations show up in version control

**Takeaway:** A privacy commitment that lives only in the IRB protocol is a commitment in name only.

# Three layered controls, each catching what the next cannot



- Minimum-necessary collection: the unshared attribute is the uncollected one
- Tiered access: raw (controlled) → analytic (fewer identifiers) → public (aggregated or synthetic)
- Output-side review: could anything published re-identify a hidden record?

§11.2

# Concept representation: the geography of measurement error

*Where errors concentrate, who they harm, and how the categories were constructed.*

# Data is never innocent



- Numbers always advance someone's way of seeing the world (Shugars & Meyer 2024)
- U.S. Census race categories changed with political commitments, not populations
- Tokenization, stopword lists, and corpus composition are value-laden choices

# Word embeddings transmit stereotype with high fidelity

§11.2

- GloVe reproduces Implicit Association Test results: career–male, family–female (Caliskan et al. 2017)
- Occupations' gender association tracks actual workforce composition at r = 0.90
- “Man is to programmer as woman is to homemaker” (Bolukbasi et al. 2016)

**Takeaway:** The embedding does not introduce the stereotype; it faithfully transmits the corpus, which transmits the world.

# Two implications no technical fix dissolves



- Downstream measures inherit the structure: “professional language” built on embeddings prefers male-coded language, by construction
- Deployed systems amplify: resume search ranks the historical pattern higher
- Using such systems is a choice about participating in that amplification

# Errors have a geography; audit it by group



- High overall accuracy often concentrates errors at category boundaries and rare cases
- Boundary cases are not random: minorities, underrepresented discourse styles, lives off the template
- Disaggregated error audit: ask “for whom is this measure worst?” as routine

↗ CSS Empirical Studies Database · cssprimer.org/studies/

**Takeaway:** The technical problem and the equity problem are the same problem.

§11.3

# Data analysis: opacity and algorithmic accountability

*In consequential settings, opacity is the breakdown of the accountability chain itself.*

# Opacity becomes an accountability problem



- Child welfare, recidivism, credit, hiring: high stakes, routine deployment
- Held-out accuracy justifies the model to developers, not to the people it scores
- Thick trust needs explanation and accountability; opaque tools permit only reliance (Koskinen 2024)

**Takeaway:** The trust must come from the human structures around the model.

# Saliency tells you where the network looked, not whether it had a reason

§11.3


Near-identical saliency maps “explain” the same image as a Siberian husky and as a transverse flute.

Source: Rudin (2019), Fig. 2, Nature Machine Intelligence. © Springer Nature. Included for instructional use.

**Takeaway:** An explanation that fits both labels explains neither.

# COMPAS and the false hope of explainability



- COMPAS: proprietary recidivism score used in U.S. bail and parole decisions
- ProPublica's audit fitted its own linear mimic — an explanation in name only (Rudin 2019)
- A mimic agreeing 90% of the time misdescribes one decision in ten
- 130+ intake items: typos alone have flipped bail outcomes

**Takeaway:** Affected people need a justification they can contest, not a description of model behavior.

# Semantic gaps and domain drift



- A label's social meaning changes across populations even when the statistics hold
- Sentiment trained on product reviews means something else on political speech
- Drift shifts the burden of false positives onto people never validated for

# Technologies are moral mediators

§11.3

- Research technologies reinforce some values at the expense of others (Carusi 2012)
- Caseworkers come to ask first what the model says; departures need justifying
- Risk scores redefine the normal case and the flag

**Takeaway:** The researcher who builds the system is part of the chain that produces its effects.

§11.4

# Scientific communication: disclosure, uncertainty, inspectability

*Communication is the practice that makes scrutiny possible at every step.*

# Disclosure as ethical commitment



- Four layouts of the same network tell four visual stories (Ch. 9's karate club)
- Document what the visual does not show: missing data, exclusions, seeds, parameters
- For AI-assisted steps: prompts and outputs, with traceability and human oversight (Ivanov et al. 2025)

**Takeaway:** Treating pipeline choices as private craft hides them from the people who would catch the mistakes.

# Honest communication of uncertainty



- Four kinds: statistical, model, measurement, algorithmic
- The worst failure is tight confidence intervals over unmodeled systematic error
- Headline accuracy with a 95% CI erases the group the model fails

**Takeaway:** This is “confidently wrong” translated into a published claim.

# Open science within real constraints



- Tiered access and platform terms pull against openness; law constrains further
- Practical rule: share what you can, restrict what you must, explain both
- Full pipeline code, schema, summary statistics, and synthetic samples meet most replication needs

**Takeaway:** The boundary between shared and withheld should itself be part of the inspectable record.

§11.5

# The AI layer: a transformed research environment

*Not a new method but a changed environment — including who produces the data.*

# Three roles of generative AI; the third is genuinely new



- Measurement tool and simulation tool: validity questions from earlier chapters
- Research-environment transformer: participants, coders, and institutions use AI too
- Human findings on unverifiable human provenance are claims you cannot defend

# Synthetic respondents are not human respondents



- Synthetic ANES responses match averages but invert higher-order relationships (Bisbee et al. 2024)
- Machine bias: strong but socially random skews; near-identical answers across personas (Boelaert et al. 2025)
- Dictator Game: a human persona does not produce human behavior (Ma 2026)

**Takeaway:** They do not stand in the same epistemic relation to social phenomena as human respondents.

# Mixed-origin data and model collapse



- LLM respondents pass standard attention checks at near-perfect rates
- Nontrivial fractions of crowdworkers admit to being LLMs
- Model collapse: synthetic content contaminates future training data, compounding bias

# Disclosure is not accountability

§11.5

- “We used an LLM; prompts in the appendix” is visibility, not answerability
- A human RA who miscoded could explain and be replaced; the model offers no handle
- What survives: procedural accountability — bounded use, recorded use, exposed use

**Takeaway:** Disclosure on top of Rules–Memory–Knowledge is verifiable; without them it is an assertion of good faith.

# Ethics is institutionally grounded practice



- Technical functions are only necessary conditions of unethical practice (Jeon et al. 2025)
- The same AI use: exploitative in one institution, empowering in another
- Deliberation at field, university, and department scales — checklists complement, never substitute

**Takeaway:** A field that skips the deliberation adopts its model providers' values by default.

§11.6–11.7

# Why CSS demands its own ethics

*The inherited framework misidentifies, at every stage, what CSS has at stake.*

# Ethics creep, and boards ruling outside their depth



- Regulation spreads outward and tightens simultaneously (Haggerty 2004)
- A journalist scrapes freely; a university researcher waits months for approval
- IRB administrators: 93.1% see unique online-data issues; 55.2% call their board well versed; 5.3% think scraping needs consent (Vitak et al. 2017)

**Takeaway:** “The Common Rule itself does not provide appropriate guidance for research with online data.”

# Ethics as embedded, covenantal practice

§11.7

- Principles are reminders of considerations, not premises for deriving judgments (Hammersley 2015)
- The covenantal view: obligations created by the relationship, not calculations of harm (May 1980)
- Learned like other judgment: emulation, deliberation with peers, accumulated cases

**Takeaway:** Ask not what the regulation requires, but what you owe the people on the other side of this data.

# Practices you can adopt now



- Treat data source as ethical context — document origin, intent, audience
- Build privacy into the pipeline: minimum collection, tiers, output review
- Audit measurement error by group, in the headline result
- Prefer interpretable models in high-stakes settings
- Record AI-assisted steps in promptbooks and promptlogs
- Share what you can; restrict what you must; explain both

**Takeaway:** Reminders that make the right considerations harder to forget — not a substitute for judgment.

# Discussion



- Your OpenAlex corpus holds records about identifiable researchers who never agreed to be subjects. Which uses preserve contextual integrity, and which breach it?
- Whose cases did your pipeline's errors fall on? Is the geography of your measurement error ethically neutral for your question?
- Draft the AI-disclosure paragraph for your pipeline. Does it make a verifiable record visible, or assert good faith?
- When an AI-coded variable turns out wrong, where does accountability sit: the prompt, the model, the audit sample, or you?

# Key concepts from Chapter 11

**Contextual integrity**: Data should travel no further than the context and audience its author intended

**Context collapse**: Content made for an imagined audience reaches whole actual audiences

**Re-identification**: Recovering identities from "anonymized" data; a risk profile, not a status

**Tiered access**: Raw data controlled; analytic data restricted; public release aggregated or synthetic

**Disaggregated error audit**: Reporting where errors concentrate, by group, alongside aggregate accuracy

**Post-hoc explanation**: LIME, SHAP, saliency: shows where a model attended, not whether it had a reason

**Mixed-origin data**: Responses that may come from humans, LLMs, or both; validate accordingly

**Ethics creep**: Review apparatus spreading and tightening without producing better protection

# Materials on the companion site



**CSS Empirical Studies Database** cssprimer.org/studies/

Audit published studies with this chapter's stage-by-stage ethical questions

**Materials by chapter** cssprimer.org/chapters/

Return to any stage's materials when an ethics question points back to practice

# Where to go next

- Read: Chapter 12, the conclusion — a century of methods controversies and what they teach
- Revisit: your exercise pipeline with the practices list; formalize one layer of privacy enforcement
- Draft: the AI-disclosure paragraph for your project, and test it against today's standard