

# COSI-230B: Natural Language Annotation for Machine Learning

## Lecture 22: Ethics, Labor, and Power

Jin Zhao

Brandeis University

Spring 2026

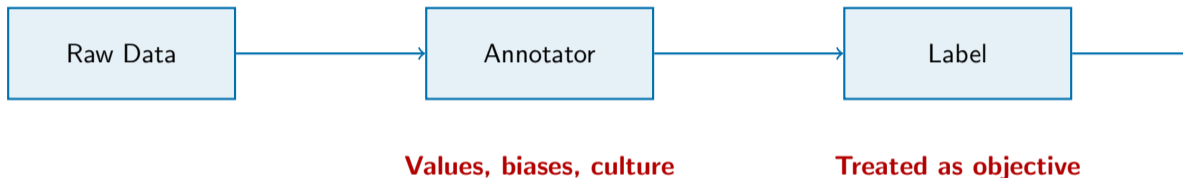
# Today's Agenda

- 1 Annotator positionality — who are the annotators?
- 2 Cultural assumptions embedded in annotation
- 3 Harm and consent in annotation work
- 4 **Discussion:** Should this task exist?
- 5 LLM angle: Hidden labor, alignment datasets, privilege and silence
- 6 Ethical frameworks and principles for annotation

## Core Question

When we ask someone to label data, whose worldview becomes ground truth?

# The Central Tension



- Annotation is not neutral — it is an **act of interpretation**
- Every label carries the annotator's perspective
- Models inherit these perspectives as “ground truth”

# What is Annotator Positionality?

**Positionality:** The social, cultural, and personal position from which a person interprets the world.

## Dimensions that shape annotation:

- Race, ethnicity, nationality
- Gender and sexuality
- Socioeconomic class
- Language and dialect
- Education and disciplinary training
- Religious and political beliefs
- Disability and neurodivergence

## Key Insight

Annotators do not “discover” labels. They **construct** them through interpretation shaped by who they are.

## Question

When we aggregate annotator judgments, whose voice gets amplified? Whose gets erased?

# Case Study: Hate Speech Annotation

**Example:** “That’s so ghetto.”

Annotator Background	Label	Confidence
White, suburban, monolingual English	Not offensive	High
Black, urban, AAVE speaker	Offensive/racist	High
Hispanic, bilingual, first-gen college	Offensive	Medium
International student, L2 English	Unclear	Low

- Same text, four different readings — all defensible
- Majority vote would likely label this “not offensive”
- The people *most affected* by the language are in the minority

# Positionality in Practice: Who Gets Hired?

## Typical crowdworker demographics (US platforms):

- Disproportionately young, white, college-educated
- English-dominant, Western cultural norms
- Often from higher socioeconomic brackets
- Low representation of marginalized communities

**Consequence:** The “gold standard” reflects a narrow slice of human experience.

## Demographic Gaps

- Global South annotators paid less, underrepresented in design
- Disability perspectives rarely sought
- Indigenous language speakers almost absent
- LGBTQ+ perspectives systematically missing from many datasets

# Cultural Assumptions in Annotation

**Every annotation schema encodes cultural assumptions.**

Task	Hidden Assumption	Who is excluded?
Sentiment analysis (pos/neg/neutral)	Sentiment is universal	Cultures with different affect norms
Toxicity detection	“Toxicity” is objective	In-group reclamation of slurs
Named entity recognition	Entity boundaries are clear	Agglutinative languages, oral cultures
Emotion classification	Ekman’s 6 basic emotions	Non-Western emotion taxonomies

# Emotion Across Cultures and the Universality Trap

## Concepts that defy Western categories:

- **Schadenfreude** (German) — pleasure from others' misfortune
- **Amae** (Japanese) — sweet dependence on another
- **Saudade** (Portuguese) — melancholic longing
- **Litost** (Czech) — tormented self-pity

## Research

Wierzbicka (1999): Emotions are not universal categories but culturally constructed concepts.

## The Universality Trap

When we force annotators to use an English-centric schema:

- Researchers assume categories are universal
- Schema fails on non-dominant cultures
- Annotators force-fit or skip items
- Models systematically misrepresent non-dominant perspectives

**Result:** Culturally specific experiences are flattened and Western frameworks are privileged.

# Harm in Annotation Work

## Annotation can cause real harm to annotators.

### Psychological harm:

- Content moderation: graphic violence, CSAM, hate speech
- Trauma exposure without adequate support
- PTSD-like symptoms documented among content moderators (Roberts, 2019)

### Economic harm:

- Below minimum wage compensation
- No benefits, no job security
- Piece-rate pay incentivizes speed over care

### Social harm:

- Asking marginalized annotators to label their own oppression
- Emotional labor unrecognized and uncompensated
- Power asymmetry: annotators cannot challenge the schema

### Key Stat

Kenyan content moderators for ChatGPT were paid \$1.32–\$2/hour to label graphic content (TIME, 2023).

# Consent in Annotation

## Meaningful consent requires:

- ① **Informed:** Annotators know what they will encounter
- ② **Voluntary:** Real ability to refuse without consequence
- ③ **Ongoing:** Can withdraw at any time
- ④ **Specific:** Consent to this task, not blanket permission

## Common Failures

- Vague task descriptions hiding sensitive content
- Economic coercion: “consent” when you need the money
- No opt-out for individual items
- No debriefing or mental health support

## Better Practices

- Content warnings before sensitive tasks
- Fair wages that allow genuine refusal
- Item-level skip options
- Access to counseling services
- Regular check-ins with annotators

# Data Subjects: The Other Consent Problem

## Who else is affected?

The people *whose data is being annotated* also face ethical risks.

Data Source	Consent Issue	Potential Harm
Social media posts	Users did not consent to being labeled	Stigmatization, surveillance
Medical records	Patients may not know data is used for NLP	Privacy violations
Court transcripts	Defendants have no say	Reinforcing carceral bias
Indigenous texts	Community did not authorize use	Cultural appropriation

**Question:** Should annotation projects require consent from data subjects as well as data annotators?

# Discussion: Should This Task Exist?

**Scenario:** A company wants to build a classifier to detect “mental instability” from social media posts, hiring crowdworkers to annotate 50,000 posts for signs of psychological distress.

## Not every annotation task should exist

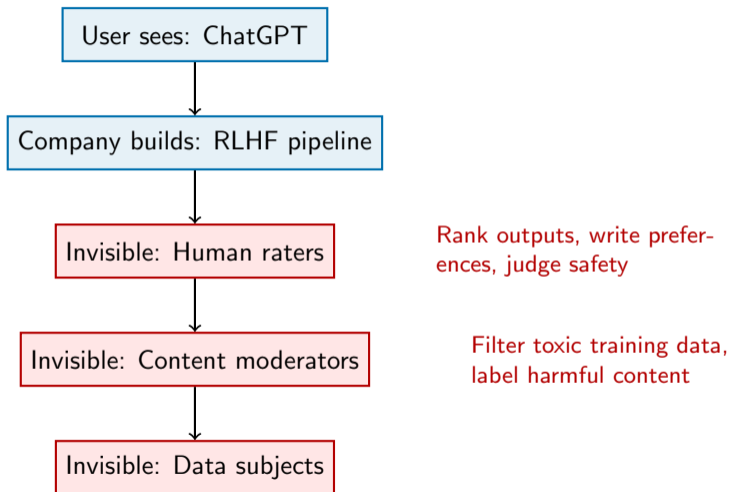
- The *ability* to annotate something does not mean we *should*
- Ethical review should happen **before** annotation begins
- Some constructs resist valid operationalization

## Questions to ask before any project

- 1 Who is affected by this annotation?
- 2 Do annotators and data subjects meaningfully consent?
- 3 Does the schema impose culturally specific values?
- 4 What happens when the model is wrong?
- 5 Who has the power to stop this project?

**Discussion:** Who benefits from this task? Who bears the risk? Could the same goal be achieved without this annotation?

# LLM Angle: The Hidden Labor Behind LLMs



# Alignment Datasets as Moral Infrastructure

**RLHF and its descendants encode human values into models.**

## How alignment works:

- ① Human raters compare model outputs
- ② They judge which response is “better”
- ③ A reward model learns their preferences
- ④ The LLM is optimized against this reward model

**But “better” according to whom?**

## Critical Questions

- Who are the raters? (Mostly US-based, English-speaking)
- What instructions do they follow? (Company values, not universal ethics)
- What counts as “helpful” vs “harmful”? (Culturally contingent)
- Can raters push back on the rubric? (Rarely)

## Insight

Alignment datasets are not neutral technical artifacts — they are **moral documents** that encode specific cultural values as universal norms

# Who LLM Annotations Privilege and Silence

## Privileged perspectives:

- English-language norms
- Western liberal values
- Middle-class sensibilities
- Corporate-friendly viewpoints

## Silenced perspectives:

- Non-English speakers and cultures
- Political dissent and activism
- Marginalized sexual and gender identities
- Working-class and poverty experiences

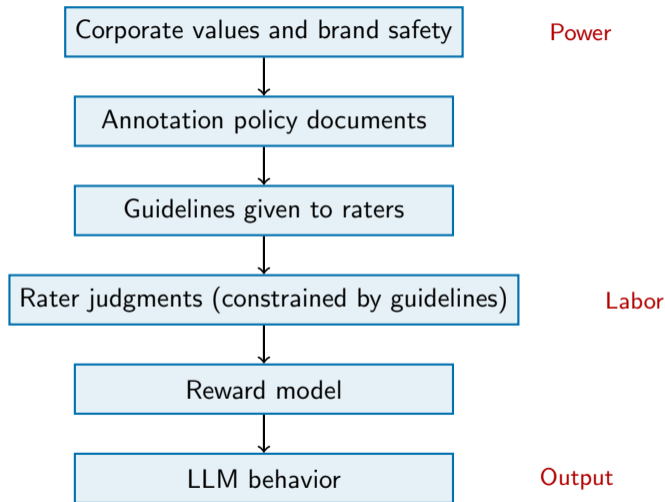
## Refusals as editorial choices:

Prompt	Pattern
Palestinian resistance	Often hedged
Drag culture for children	Sometimes refused
Safe drug use	Typically refused
Critique capitalism	Usually compliant

## The Paradox

Models trained to be “safe” can systematically erase the perspectives of those who most need to be heard.

# The Supply Chain of Values



**Those with power** set the values. **Those who labor** implement them. **Users** experience the

# Ethical Frameworks for Annotation

## Consequentialist

- Focus on outcomes
- Does this annotation reduce harm overall?
- Cost-benefit across stakeholders
- Risk: justifies exploitation for “greater good”

## Deontological

- Focus on duties
- Are annotator rights respected?
- Is consent genuine?
- Risk: rigid rules may miss context

## Virtue Ethics

- Focus on character
- Does this project reflect integrity?
- Would we be proud of this process?
- Risk: subjective, hard to operationalize

## Recommendation

Use all three lenses. No single framework is sufficient for the complexity of annotation ethics.

# Toward Ethical Annotation: Principles and Practice

## A proposed ethics checklist for annotation projects:

- 1 **Positionality audit:** Document who your annotators are and who they are not
- 2 **Schema review:** Identify cultural assumptions in your categories
- 3 **Consent protocol:** Ensure informed, voluntary, ongoing consent
- 4 **Fair compensation:** Pay at least living wage, not piece-rate
- 5 **Harm mitigation:** Provide mental health support for sensitive tasks
- 6 **Data subject rights:** Consider consent of people in the data
- 7 **Power analysis:** Who decides the schema? Who can change it?
- 8 **Disagreement as data:** Preserve annotator disagreement rather than forcing consensus

## Practical Reality

Perfect ethics is impossible. The goal is **reflective practice** — knowing the trade-offs and documenting them honestly.

# Disagreement as Signal, Not Noise

## Traditional view:

- Disagreement = annotator error
- Solution: more training, majority vote
- Goal: single “correct” label

## Critical view:

- Disagreement = legitimate difference in perspective
- Solution: preserve and model disagreement
- Goal: understand the *landscape* of interpretations

## Approaches

- **Jury Learning** (Gordon et al., 2022):  
Model individual annotator perspectives
- **Perspectivism** (Basile et al., 2021):  
Reject single ground truth
- **Distributional labels**: Report label distributions, not majority

## Connection

Respecting disagreement is an ethical stance: it refuses to silence minority perspectives.

# Key Takeaways

- ➊ **Annotation is not neutral.** Every label carries the annotator's positionality, and schemas encode cultural assumptions.
- ➋ **Harm is real.** Annotators and data subjects face psychological, economic, and social harm that is often invisible.
- ➌ **Consent must be meaningful.** Informed, voluntary, ongoing, and specific — not a checkbox.
- ➍ **LLMs inherit annotation politics.** RLHF alignment datasets are moral infrastructure that privileges some worldviews and silences others.
- ➎ **Disagreement is valuable.** Preserving multiple perspectives is both better science and better ethics.
- ➏ **Ask: should this task exist?** The ability to annotate does not imply that we should.

## Questions?

Office Hours: Wednesdays 1–3pm, Volen 109

✉ [jinzhao@brandeis.edu](mailto:jinzhao@brandeis.edu)