



Lecture 7: Data as Specification

CS120: Introduction to AI Safety - October 13, 2026, Zachary Robertson

- Who makes the dataset? We return to the crowd-sourced toxicity predictor and examine how development decisions about the dataset can produce shortcuts, such as a spurious correlation between dialect and toxicity. The platform intends to remove targeted abuse while allowing other comments.

Who specifies the task?

Generalization Performance on future samples from a specified population.

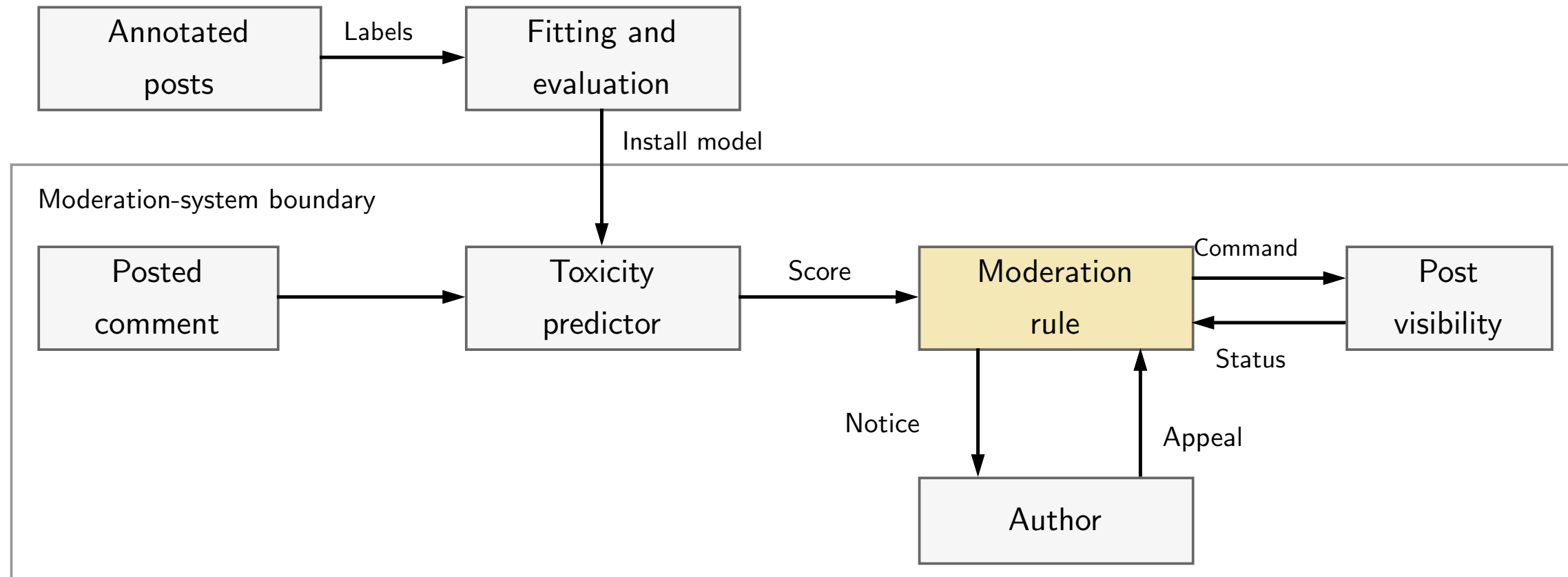
Robustness Behavior under specified changes in inputs or conditions.

Who chose the population, labels, groups, and permitted changes?

Today: trace those choices into data production and model selection.

- In this module we have considered generalization and robustness and why neither alone is sufficient to claim the data specifies the intended task. Now we examine how choices about populations, labels, groups, and permitted changes enter data production and model selection.

A score does not remove a post



Intended policy: remove targeted abuse while allowing other comments.

- Development supplies a fitted predictor; the score does not itself remove a post. The moderation rule controls visibility and receives status feedback. The author receives a decision notice and can appeal, but restoration may come too late to preserve participation. This simplified workflow leaves the data-production decisions to be examined below.



Roadmap

Whose judgments become labels?

Document consequential choices

Change data and trace consequences

- Sap et al. study annotator attitudes and their associations with annotations. We also ask about researchers' own choices.

A recorded rating is not the intended policy

Recorded judgment	How offensive or toxic does an annotator perceive this comment to be?
Intended decision	Does the comment violate the platform's policy against targeted abuse?

Accurate prediction of ratings does not establish accurate application of the policy.

- Sap et al. measure perceived offensiveness of online comments. The platform instead intends to remove targeted abuse while allowing other comments. Reusing the ratings as policy-violation labels assumes these constructs align. Accurate prediction of recorded ratings does not resolve that assumption.

Annotators with Attitudes

Sap et al. study perceived offensiveness and its associations with dialect, prejudice, and vulgarity.

Studied comments include African American English (AAE), racism, and vulgarity.

Higher racist-belief scores	Associated with lower toxicity ratings for anti-Black posts.
The same attitude measure	Associated with higher toxicity ratings for AAE posts.

Sap et al., NAACL 2022

- In *Annotators with Attitudes*, Sap et al. measure perceived offensiveness and its associations with dialect, prejudice, and vulgarity. They focus on comments exhibiting AAE, racism, or vulgarity. Higher scores on their racist-belief scale were associated with lower toxicity ratings for anti-Black posts and higher toxicity ratings for AAE posts.

Where can a dialect shortcut originate?

A possible mechanism upstream of model fitting:



A dataset can reward the association before a model learns it.

- The paper suggests a mechanism upstream of model fitting: the composition of the annotator pool can embed spurious associations directly into the training labels. In our original crowd-sourced predictor example, we focused on a model exploiting a spurious correlation between dialect and toxicity. Here we examine how annotation decisions can create or reinforce that correlation.

Researchers also make specification choices

Breadth-of-posts study: 173 annotators label 571 tweets.

Measurement	One questionnaire item per attitude dimension, selected using the earlier study.
Assignment	Two White conservatives, two White liberals, and two Black annotators per post.

Which comparisons and measurement claims does this design support?

Sap et al., section 3.2 and Appendix D

- In the breadth-of-posts study, 173 annotators label 571 tweets. Each attitude dimension uses a single questionnaire item selected from the earlier study. The assignment scheme politically differentiates White annotators but not Black annotators. These decisions influence which comparisons the design supports and raise questions about the reliability and scope of the attitude measurements. Pointing out study decisions is not itself a critique.

What did the study establish?

Evidence reported	Associations among measured attitudes, identities, and annotations in the study.
Further model claim	Training on those labels produces a predictor that relies on dialect.
Further system claim	That predictor causes permitted posts to be removed with consequential delays.

- The paper studies annotator attitudes and their associations with annotations. Tracing those associations into harm requires evidence about model behavior and the moderation rule. A predictor using dialect as a proxy for toxicity could contribute to speakers losing timely participation because their allowed posts are removed.



Roadmap

Whose judgments become labels?

Document consequential choices

Change data and trace consequences

- Datasheets document researcher decisions. The goal is to identify evidence that would change a decision about using the data.

Datasheets make choices traceable

Gebru et al. propose documenting how a dataset was made and how it should be used.

Purpose and composition

Why this dataset? What does it contain?

Collection process

How were instances and judgments obtained?

Processing and labeling

Which transformations produced the usable data?

Uses

Which uses are recommended or discouraged?

Gebru et al., Datasheets for Datasets

- Gebru et al. propose documentation as a way to give researcher decisions traceable provenance. Recommended components include motivation, composition, collection, preprocessing, cleaning, labeling, and uses to recommend or avoid. A datasheet is more than checking boxes: its goal is to identify information or evidence that would change a decision about using the data.

Documented, inferred, or unknown?

Documented	An attributed claim about what was done or observed.
Inferred	A conclusion requiring reasoning and explicit assumptions.
Unknown	A question the available sources do not answer.

An unanswered question does not establish that a step was omitted.

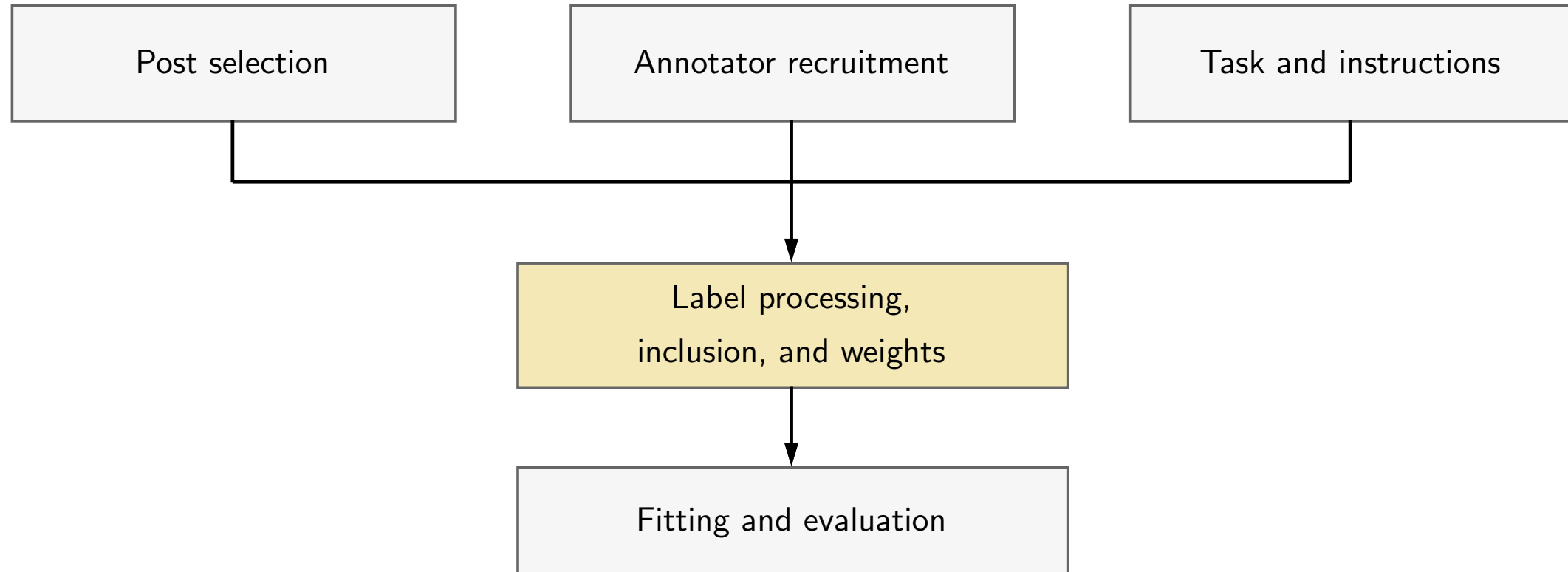
- It is important to document attributed claims, inferences, and what remains uncertain. Just because a question is unanswered does not establish that a step was omitted. A documented fact may support a proposed use or limit its suitability; an unknown identifies where further inquiry is needed.

Follow provenance into a reuse decision

Source	Ratings elicited to measure perceived offensiveness.
Assumption	Those ratings can stand for violations of the intended policy.
Check	Compare ratings with independently justified policy judgments.
Decision	Substantial mismatch may require new labels or a narrower use.

- Using perceived-offensiveness ratings as policy-violation labels introduces an assumption about their suitability. Compare the ratings with independently justified policy judgments, documenting how those judgments were made. A substantial mismatch may call for new labels, narrower use, or declining reuse. This proposed check connects provenance to a decision about using the data.

Expand the development boundary



- Instead of conceptualizing data production as a black box, we ask which decisions and assumptions flowed into its construction. Expanding the development boundary brings post selection, annotator recruitment, task instructions, label processing, inclusion, and weighting into view. Documentation helps identify where to investigate the effects of these choices on model behavior.

Documentation guides inquiry; it does not certify safety

Which fact or empirical result would change the decision to use this data?

Can support

Tracing choices, exposing limitations, and targeting further investigation.

Does not establish alone

Construct validity, acceptable predictor behavior, or timely correction of removals.

- The end-goal of using a datasheet is to identify what information or evidence would change a decision about using the data. Such documentation can help identify where to investigate; it does not by itself establish that a dataset or its downstream use is safe.



Roadmap

Whose judgments become labels?

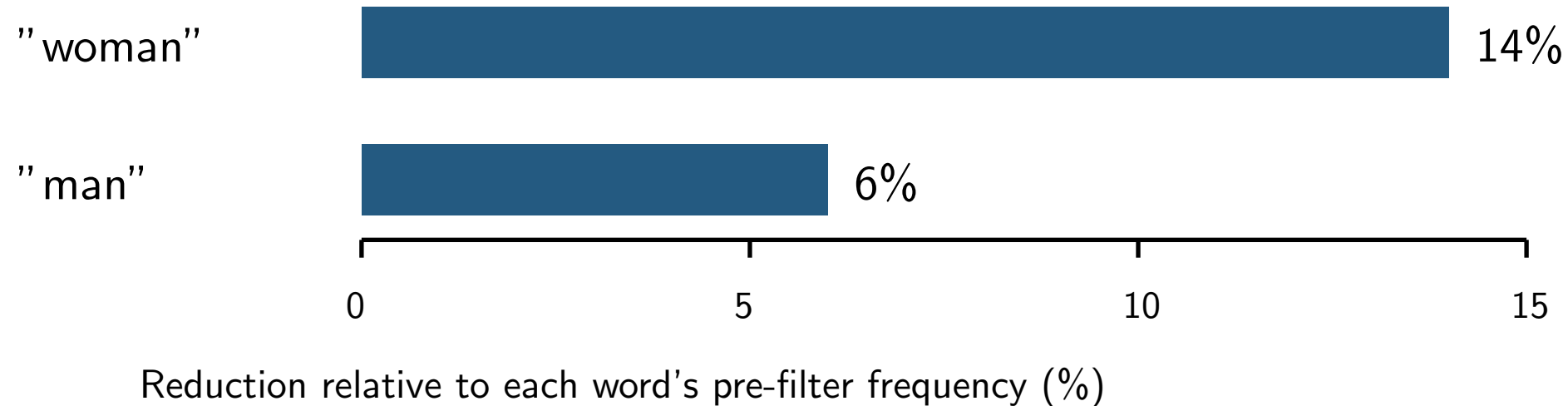
Document consequential choices

Change data and trace consequences

- Filtering and reweighting help specify the training objective. We extend the moderation example into a hypothetical harm scenario.

Filtering can change more than its target

DALL-E 2 pre-training: filter sexual and violent imagery to reduce problematic outputs.



OpenAI, DALL-E 2 Pre-Training Mitigations, 2022

- OpenAI's DALL-E 2 report describes filtering sexual and violent imagery from training data to reduce problematic outputs. Filtering reduced caption-word frequencies for "woman" by 14 percent, versus 6 percent for "man," suggesting amplified gender skew in dataset composition. Each reduction is relative to that word's pre-filter frequency. The chart shows these reported values.

Reweightings is a second intervention

Filtering

Changes which examples are retained.

Reweightings

Changes the influence of retained examples during fitting.

A control aimed at one problem can motivate a further mitigation.

OpenAI, DALL-E 2 Pre-Training Mitigations, 2022

- The report then describes reweighting the filtered data to address filtering-induced bias. Filtering and reweighting are distinct interventions: filtering changes which examples are retained, while reweighting changes their influence during fitting. A control aimed at one problem can introduce another, motivating a further mitigation.

Data choices enter the training objective

$$\hat{\theta}_c \in \arg \min_{\theta \in \Theta} \frac{\sum_{i \in I_c} w_i^{(c)} \ell(\theta; x_i, y_i^{(c)})}{\sum_{i \in I_c} w_i^{(c)}}$$

I_c Which instances are retained

$y_i^{(c)}$ Which label is used

$w_i^{(c)}$ How much each instance contributes

Weights are nonnegative and their total is positive. Here, i can index annotations.

- Development controls such as inclusion, labels, and weights are choices c that help specify the training objective. In moderation, i can index annotations, I_c identifies retained annotations, and $y_i^{(c)}$ is the label used under choice c . The weights $w_i^{(c)}$ are nonnegative, with a positive total weight, and help shape which model is selected. The hypothesis class and loss form are held fixed.

Choose the policy, then fit the model



Intended change Which measured problem should the intervention reduce?

Intended invariance Which relevant behavior should remain acceptable?

- Researchers choose c and then fitting chooses parameters for the model. A safety intervention may change the training distribution. We must ask what changes, what should stay invariant, and how to connect the researcher choice to observable evidence. For example, does filtering or reweighting reduce one measured problem while suppressing subgroup representation or preserving a harmful shortcut?

First fix the baseline pipeline

Hypothetical harm scenario

Fixed policy c_0

Repeated training runs return context-sensitive and dialect-shortcut predictors.

Validation equivalence

Their validation performance is statistically indistinguishable.

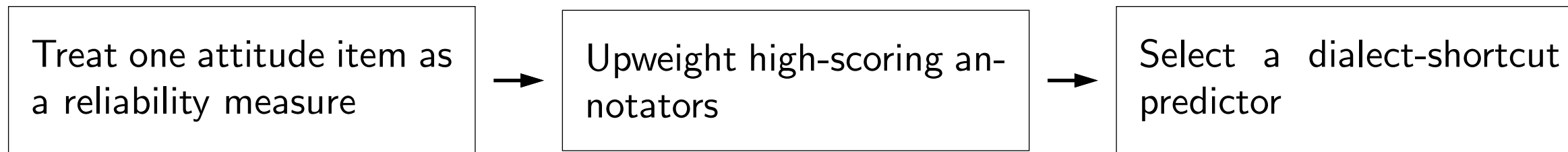
Unresolved behavior

Validation omits non-abusive AAE; the predictors differ on permitted posts.

- Under a fixed baseline development policy c_0 , assume repeated training runs yield context-sensitive and dialect-shortcut predictors with statistically indistinguishable validation performance. Suppose validation omits non-abusive AAE, leaving their different treatments of allowed speech unresolved. This is consequential underspecification under a fixed pipeline. These are assumptions for our hypothetical scenario, not reported findings of Sap et al.

Then change the development policy

Change $c_0 \rightarrow c_1$; keep the same validation check.



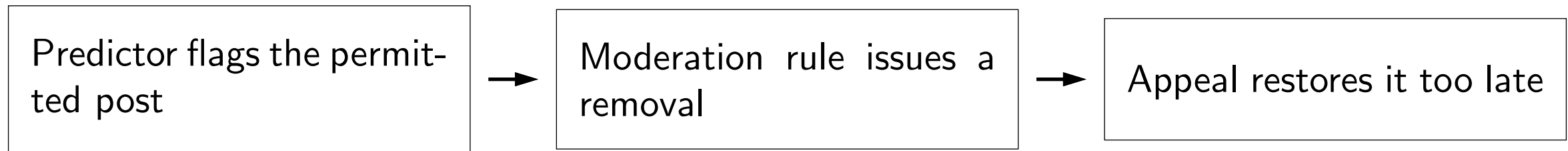
Assume the pilot confounds dialect with toxicity and those annotators rate dialect itself as toxic.

Changing the objective does not repair the validation gap.

- Suppose a moderation pilot confounds dialect with toxicity. The team treats a strongly correlating questionnaire item as an annotator reliability measure and upweights high scorers. Suppose these annotators tend to rate dialect itself as toxic. Changing c_0 to c_1 increases their annotations' weights and can favor a dialect-shortcut predictor; assume it is selected and passes unchanged validation. The original validation equivalence was under c_0 : changing c changes the training objective, not the adequacy of the validation check.

From a label association to lost participation

The deployed predictor encounters an AAE post the policy permits.



Hazard H_T

An allowed post is hidden while timely correction is unavailable.

Harm L_T

The speaker loses timely participation because the post was removed.

- The predictor flags an AAE post that the intended policy permits, and the moderation rule issues a removal instruction. The platform enters hazardous state H_T : an allowed post is hidden while timely correction is unavailable. An appeal restores it after the conversation has moved on. The participant realizes harm L_T : loss of timely participation. Repetition can produce systematic exclusion.

Requirements need evidence at different points

SC_T : **prevent or correct erroneous removals before the participation window closes.**

Development check

Evaluate permitted non-abusive AAE before and after reweighting; compare validation-equivalent predictors.

System check

Compare removal, appeal, and restoration times with a defined participation window.

- A potential constraint is to prevent or correct erroneous removals before the participation window closes. Evaluate non-abusive AAE against the intended policy before and after reweighting, comparing predictors indistinguishable on the original validation check. Testing alone does not prevent a removal; results must inform selection or deployment. Check correction timeliness using removal, appeal, and restoration logs against an explicitly defined participation window, including unresolved cases.

Data specifies learning, not safety by itself

Provenance	Makes data-production choices and uncertainty traceable.
Validation	Tests specified behavior; other differences may remain unresolved.
Selection	Determines which predictor is installed.
System controls	Shape whether prediction errors become participation harm.

Discussion: should the platform reuse the breadth-of-posts data?

- Researcher choices can influence data in ways that are hard to detect within the development pipeline. Available evaluations may leave safety-relevant differences among well-performing models unresolved. Documenting decisions from data collection to intended use makes provenance traceable and helps identify consequential sources of underspecification before model behavior contributes to harm. The discussion applies this approach to a reuse decision.

Discussion: Data as specification

A moderation team is considering reusing the *breadth-of-posts* data from Sap et al., *Annotators with Attitudes*, to train the toxicity predictor from lecture. The platform intends to remove targeted abuse while allowing other comments. Use Sap et al. §3.2 and Appendix D.1–D.2, and Gebru et al., *Datasheets for Datasets*, §3.2 (Composition), §3.3 (Collection Process), and §3.5 (Uses). If the sources do not answer a question, mark it **unknown** rather than inferring an answer. Do not use the lecture's one-item attitude/reweighting example.

1. Proposed reuse

What did Sap et al. collect the ratings to study, and what would the platform ask those ratings to stand for? State one assumption needed for this reuse.

2. Composition and sampling

For the posts, annotators, and ratings, what are the dataset's instance types and how were they sampled? What larger population, if any, can each reasonably represent?

3. Consequential provenance

Choose one documented collection choice or one **unknown** from Question 2. Give a plausible route by which it could contribute to a permitted post being removed. What additional assumption does your route require?

4. Reuse decision

Using *Datasheets* §3.5, identify one additional fact or empirical check that would most affect reuse. What result would make you more or less willing to reuse the data, and what safety claim would still remain unsupported? Conclude: **reuse / reuse with conditions / do not reuse yet.**