Detailed Proofs

A.1 Proof of Non-Determinacy of Content-Agnostic Moderation (Intuition from Indistinguishable Inputs / Contradiction)

Claim: A content-agnostic function f cannot guarantee a targeted stance distribution D across all recommendations because it cannot distinguish $\pi_{t,1}$ from $\pi_{t,2}$ based solely on content-agnostic features if they have different underlying stance distributions but identical relational properties.

Proof. Assume for contradiction that there exists a content-agnostic moderation function $f : \Pi \to \Pi$ (where Π is the space of all recommendation configurations π_t) capable of modifying any given π_t to achieve a targeted stance distribution D in the output $\pi'_t = f(\pi_t)$.

Consider two recommendation configurations at a given time t:

1. $\pi_{t,1}$, where the items recommended predominantly belong to a single stance $s_1 \in \mathcal{S}$. The resulting stance distribution of items in $\pi_{t,1}$ is far from D.

2. $\pi_{t,2}$, where the items recommended are already distributed according to the target stance distribution D.

Crucially, let us construct $\pi_{t,1}$ and $\pi_{t,2}$ such that all their relational properties accessible to f are identical. For example, they recommend the same number of items to the same users, and if f uses historical data $\mathbf{C}_t$, assume this history is the same for both scenarios leading up to the generation of $\pi_{t,1}$ and $\pi_{t,2}$. Thus, from f's perspective (which only sees relational data derived from π_t and $\mathbf{C}_t$), $\pi_{t,1}$ and $\pi_{t,2}$ are indistinguishable.

To achieve the target distribution D, $f(\pi_{t,1})$ would require substantial modification of its item composition (e.g., swapping many items to diversify stances). In contrast, $f(\pi_{t,2})$ should ideally result in minimal or no changes, as $\pi_{t,2}$ already meets D.

However, since f is content-agnostic and perceives $\pi_{t,1}$ and $\pi_{t,2}$ as identical inputs due to their matching relational properties, it must apply the same transformation to both. If f modifies $\pi_{t,1}$ to achieve D, it must also modify $\pi_{t,2}$ in the same way, potentially moving it away from D. Conversely, if f leaves $\pi_{t,2}$ largely unchanged (as it should), it must also leave $\pi_{t,1}$ largely unchanged, failing to achieve D for $\pi_{t,1}$.

This leads to a contradiction: f cannot consistently transform all possible inputs to achieve D while treating relationally indistinguishable inputs identically. Therefore, no such content-agnostic function f can guarantee the achievement of D for all $\pi_t \in \Pi$. $\square$

A.2 Proof of Non-Determinacy of Content-Agnostic Moderation (Intuition from Learning Uncertainty)

Claim: If f were capable of producing an output π'_t that conforms to D, consistently training f to realize π'_t using a content-agnostic learning approach is unfeasible, as the learning algorithm cannot differentiate π'_t from other potential outputs π''_t that do not meet D but share the same relational characteristics.

Proof. Assume, for the sake of argument, that a content-agnostic moderation function f could, in principle, take an input recommendation configuration π_t and produce an output π'_t whose items exhibit the target stance distribution D. Now, consider training such an f using a machine learning approach where only relational properties of π_t (and perhaps $\mathbf{C}_t$) are available as input features, and the learning algorithm must learn to produce π'_t.

The core issue arises during the learning process. Suppose the learning algorithm considers π'_t (which achieves D) as a desirable output. However, because the learning process is content-agnostic, it cannot directly assess the stance distribution of π'_t. It can only evaluate π'_t based on its relational properties (e.g., similarity to π_t, diversity of item IDs, etc.).

Now, consider another potential output configuration π''_t that could be generated by f. Let π''_t be relationally indistinguishable from π'_t (i.e., it has the same number of items, perhaps similar diversity scores based on item IDs, etc.), but its underlying (unseen) item stances result in a distribution $D'' \neq D$.

Since the learning algorithm for f only has access to relational features, it cannot differentiate between π'_t and π''_t in terms of their true alignment with the target stance distribution D. Any loss function or reward signal based solely on relational properties would assign similar scores to both π'_t and π''_t.

This means the learning algorithm has no basis to prefer π'_t (which meets D) over π''_t (which does not). The training process would be unable to reliably guide f towards outputs that consistently achieve D, as outputs that fail to meet D but have similar relational characteristics would be equally plausible outcomes. This leads to unpredictable and unreliable moderation with respect to the unobserved stances.

Therefore, it is theoretically unfeasible to train a content-agnostic function f to reliably produce outputs aligning with a specific stance distribution D when the learning signals are restricted to relational properties. □

A.3 Proof that Egalitarian Exposure Leads to Uniform Distribution of Stances

Proposition 1. *If the set of all available items $\mathcal{I}$ is uniformly distributed across stances in $\mathcal{S}$, then achieving uniform exposure for each item in $\mathcal{I}$ results in a uniform distribution of stances among the exposed items.*

Proof. Let $|\mathcal{I}_s|$ represent the number of items in $\mathcal{I}$ associated with stance $s \in \mathcal{S}$.

Since the distribution of stances across items in $\mathcal{I}$ is uniform, we have:

$$|\mathcal{I}_s| = \frac{|\mathcal{I}|}{|\mathcal{S}|} \quad \text{for each } s \in \mathcal{S}.$$

Let "uniform exposure" mean that each item $i \in \mathcal{I}$ receives the same total number of exposures (e.g., appearances in recommendation lists over a period, or total clicks) across all users. Let this uniform exposure count per item be e_{item}.

The total exposure count for all items associated with a particular stance s, denoted E_s, is the sum of exposures for each item in $\mathcal{I}_s$:

$$E_s = \sum_{i \in \mathcal{I}_s} e_{\text{item}}$$

Since e_{item} is constant for all items, this sum becomes:

$$E_s = |\mathcal{I}_s| \times e_{\text{item}}$$

Substituting the expression for $|\mathcal{I}_s|$ from the uniform item distribution assumption:

$$E_s = \left(\frac{|\mathcal{I}|}{|\mathcal{S}|} \right) \times e_{\text{item}}$$

Since $|\mathcal{I}|$, $|\mathcal{S}|$, and e_{item} are constants, E_s is the same value for every stance $s \in \mathcal{S}$. If the total exposure for each stance is identical, then the distribution of stances among the exposed items is uniform. $\square$