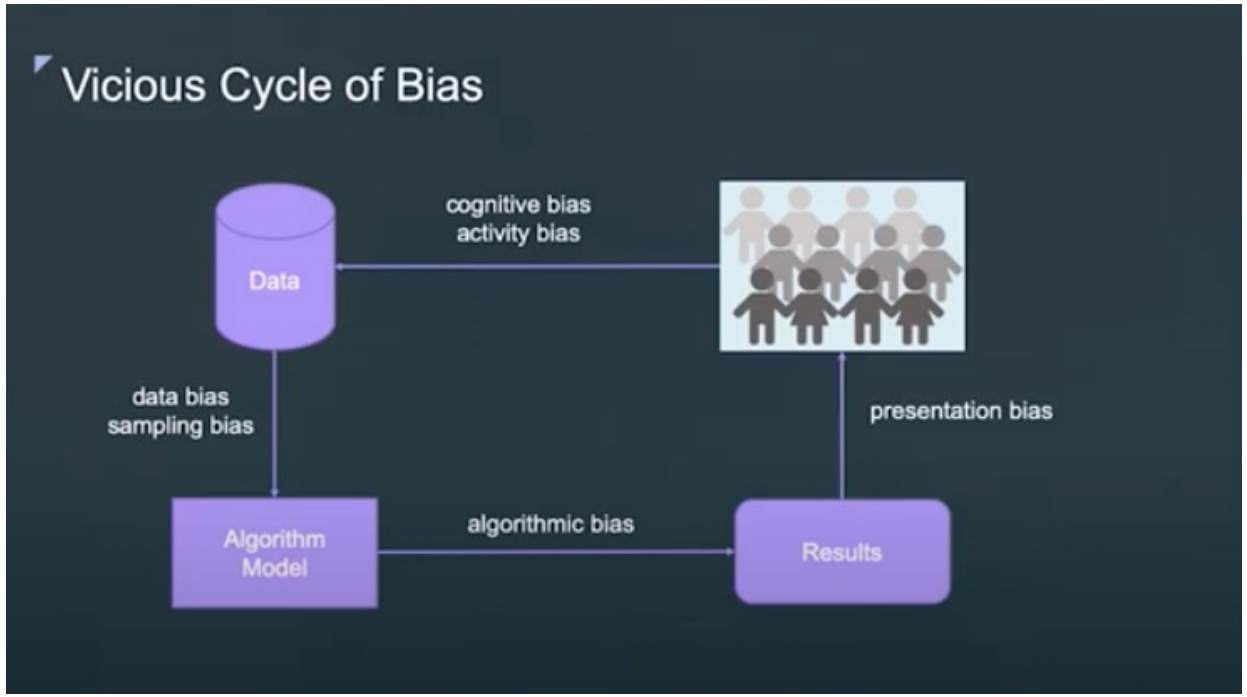
レコメンドシステムによる
バイアスと公平性に関する問題

- 参考1) https://www.youtube.com/watch?v=TtF6exuBbSU, RecSys 2020 Tutorial: Counteracting Bias and Increasing Fairness in Search and Recommender Systems）
- 参考2) https://arxiv.org/pdf/2010.03240.pdf、 Bias and Debias in Recommender System: A Survey and Future Directions
- 参考3) https://arxiv.org/pdf/1908.09635.pdf、 A Survey on Bias and Fairness in Machine Learning
- 参考4) https://link.springer.com/content/pdf/10.1007/s11257-019-09256-1.pdf、 Multistakeholder recommendation: Survey and research directions
- 参考5) https://eng.uber.com/uber-eats-recommending-marketplace/、 Food Discovery with Uber Eats: Recommending for the Marketplace
- 参考6) https://www.deeplearning.ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI.pdf、 MLOps: From Model-centric to Data-centric AI

レコメンドシステムによるフィードバックループとバイアスの拡大

- バイアスを含んだレコメンドがユーザーの行動を決定し、さらにそれがレコメンドの学習データとして戻ってくる「正のフィードバックループ」が存在（正のフィードバックであることが問題）
- 偏りを生むだけでなく、時間の経過とともに偏りを強める



引用元：参考1

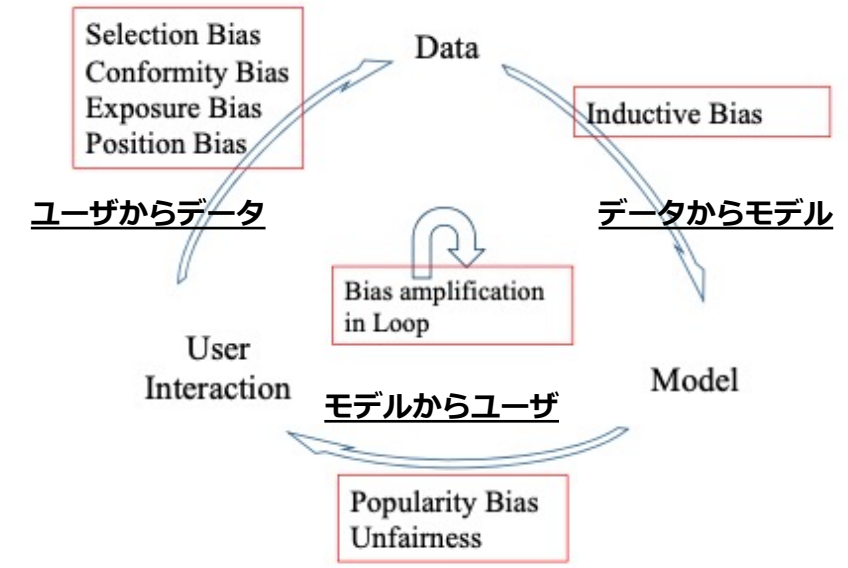


Fig. 2. Feedback loop in recommendation, where biases occur in different stages.

引用元：参考2

フィードバックループによるバイアス拡大の例 ～エコーチェンバーとフィルターバブルの発生

- 例）Popularity Bias
  - フィードバックループは人気バイアスを増幅させ、人気のあるアイテムはさらに人気が増し、人気のないアイテムはさらに人気がなくなることがシミュレーションで検証されている
  - これらの増幅されたバイアスは、ユーザーの多様性を減少させて同質化を強め、いわゆる「エコーチェンバー」や「フィルターバブル」を生じさせる（参考2から）

- エコーチェンバー効果
  - “「エコーチェンバー効果」とは、エコーチェンバーのような閉じたコミュニティの内部で、誰と話しても自分と同じ意見しか返って来ないような人々の間でコミュニケーションが行われ、同じ意見がどこまでも反復されることで、特定の情報・アイデア・信念などが増幅・強化される状況のメタファー（隠喩）となっている[4]。” (Wikipediaから引用)
- フィルターバブル
  - “フィルターバブル (filter bubble) とは、「インターネットの検索サイトが提供するアルゴリズムが、各ユーザーが見たくないような情報を遮断する機能」（フィルター）のせいで、まるで「泡」（バブル）の中に包まれたように、自分が見たい情報しか見えなくなること。”（ Wikipediaから引用）

レコメンドシステムにおけるバイアスの種類

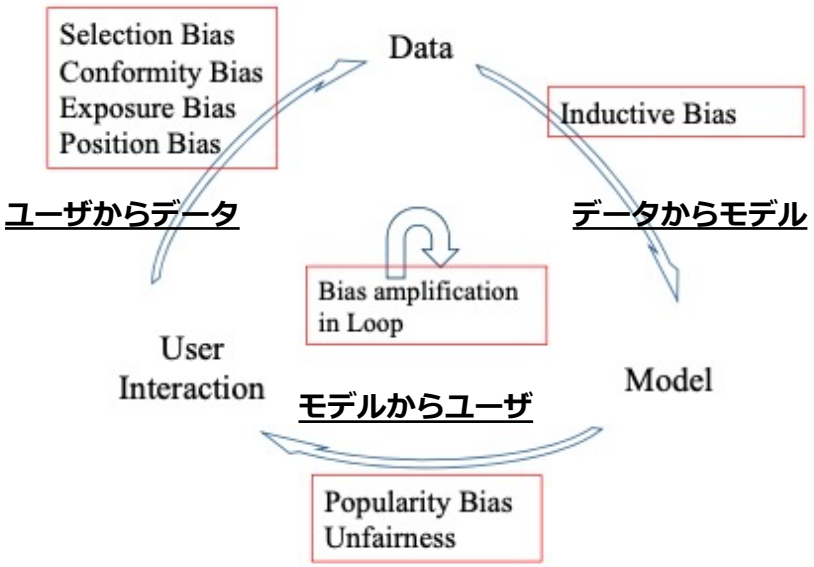


Fig. 2. Feedback loop in recommendation, where biases occur in different stages.

引用元：参考2

Selection Bias （選択バイアス）	ユーザは興味を持ったものしか評価/選択しないため、観測された評価データが真のユーザの代表的なサンプルではない。ユーザは自分が気に入っているアイテムを評価したり、特に良いものや悪いものを評価しやすい
Conformity Bias （同調バイアス）	自分の判断に反して、周囲の判断に迎合する。みんなの評価が高いと厳し過ぎる評価を避けるなど。評価値がユーザの真の評価とは限らない
Exposure Bias （露出/曝露バイアス）	ユーザが一部のアイテムにしか接触しないために生じるバイアス。評価のないアイテムが必ずしもネガティブというわけではない。ユーザの興味に合わない、そもそも知らない、の２パターンある
Position Bias （ポジションバイアス）	レコメンドの上位にあるアイテムについて本来は興味がなくてもクリックしたりする。特定の位置にあるものに気付きやすかったりする
Inductive Bias （帰納バイアス）	機械学習手法が汎化のために使用するデータ以外の仮定や制約。望ましい特性を得るためにモデル設計に意図的に加えたりする
Popularity Bias （人気バイアス）	ごく一部の人気アイテムがインタラクションの大半を占める（ロングテール）。ロングテールのデータで学習すると、人気のアイテムには過剰に高く、逆に人気のないアイテムは低くなる。
Unfairness （不公平/不当性）	不均衡なデータで学習を行うと、モデルはこれらの過大に表現されたグループを学習し、ランク付けされた結果でそれらを強化する可能性が高く、不利なグループに対する組織的な差別や可視性の低下を引き起こす可能性があります（例：マイノリティの過小表現、人種や性別によるステレオタイプ）

バイアスを取り除く方法～fairnessとloop effectの一例

- Fairness（公平性）に関するバイアスへの対応
  - 対応例）リバランシング
    - データまたは推薦結果を、人口統計学的平準化のような特定の公平性目標に対してバランスさせること
    - 機械学習の研究で集中的に採用されている戦略には，保護されたグループと保護されていないグループで正のラベルの割合が等しくなるように訓練データを再ラベリングすることや，統計的に同等になるように訓練データを再サンプリングすることがある

- Loop effectへの対応
  - 対応例）uniform dataの導入
    - 各ユーザに対して，アイテム配信のための推薦モデルを使用せず，代わりにいくつかのアイテムをランダムに選択し一様にランク付けしたものを含める
    - 当然、ランダムなものが含まれればクリック率の低下が起こるため、トレードオフを両立させるための研究が続けられている
  - 対応例）強化学習（バンディットアルゴリズム等）の活用
    - 探索と活用を同時に行う

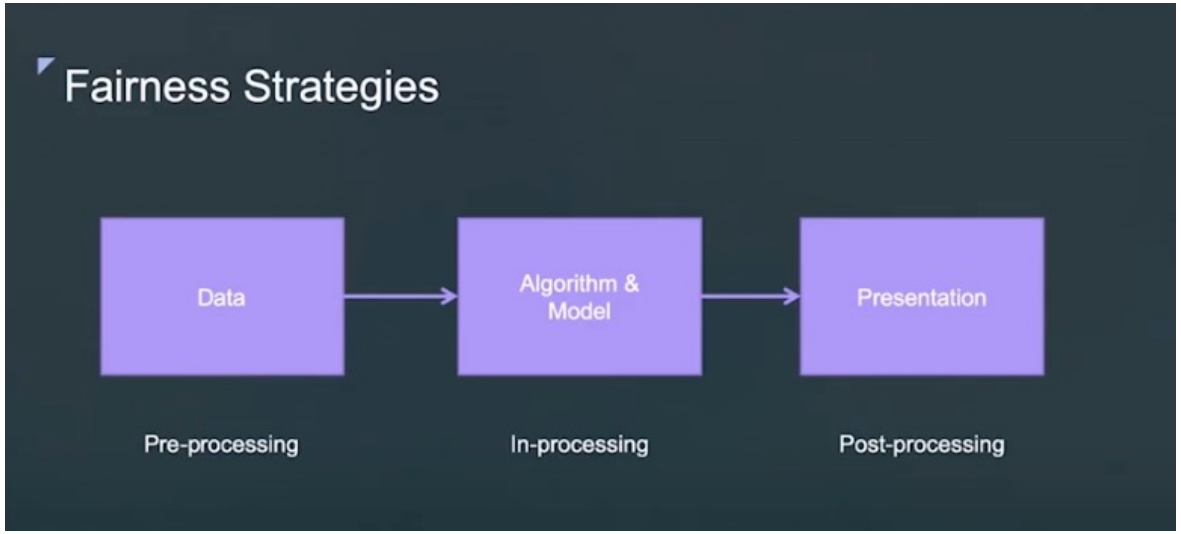
公平性に関して考慮すべき項目の例

- 考慮すべき項目の例：人種、肌の色、国籍、宗教、性別など
- Pre-/In-/Post-processingの中で処理する

Attribute	FHA	ECOA
Race	✓	✓
Color	✓	✓
National origin	✓	✓
Religion	✓	✓
Sex	✓	✓
Familial status	✓	
Disability	✓	
Exercised rights under CCPA		✓
Marital status		✓
Recipient of public assistance		✓
Age		✓

Table 4. A list of the protected attributes as specified in the Fair Housing and Equal Credit Opportunity Acts (FHA and ECOA), from [29].

引用元：参考3



引用元：参考1