Privacy & VSD

Q1 (Privacy): According to the definitions of privacy, which kind is at stake here and why? What solution would you propose?

The kind of privacy at stake in this case is **Privacy II**: **Privacy as the proper flow of personal information**. This involves controlling how personal information is shared, who has access to it, and the conditions under which it is disseminated【96†source】. In Gillian's case, personal information about her pregnancy journey—including the emotional impact of her stillbirth—was used inappropriately by algorithms to target ads, without her consent and under a misguided assumption of a positive outcome (successful pregnancy). The **contextual flow** of this information was violated since it was used in a way that did not align with Gillian's needs or expectations during a highly sensitive time.

Proposed Solution:

- **Contextual Privacy Control**: Implement stronger mechanisms to recognize and adjust the flow of personal information based on **contextual cues**. Algorithms should be able to account for changes in context (such as the emotional impact of a stillbirth) and adjust their behavior accordingly.
- **User Input on Context Changes**: Give users an option to easily indicate significant changes in their life (e.g., stillbirth, death in the family) that would trigger the adjustment or halt of personalized content, thus minimizing harm.
- **Differential Privacy Filters**: Use differential privacy filters to limit how much personal information is shared across different advertising algorithms, ensuring that sensitive information is protected and used appropriately.

Q2 (Privacy): Identify three reasons why privacy should be preserved in this case. Give reasons why.

1. **Avoiding Informational Harm**:

   - Misuse of information about Gillian's stillbirth led to emotional harm by showing her ads related to pregnancy and infant care, which was deeply painful. Preserving privacy can help prevent such **informational harm**, ensuring that personal information is not used in ways that cause distress【96†source】.

2. **Maintaining Moral Autonomy**:

   - Privacy is crucial for allowing individuals to make decisions freely without undue influence or emotional manipulation. The use of Gillian's data to deliver ads took away her **moral autonomy**, forcing her to confront unwanted content that she had no control over. Protecting privacy ensures that personal information is used in ways that respect individual autonomy【96†source】.

3. **Preventing Informational Inequality**:

   - Companies like the one responsible for these ads often hold much more power over personal information than individuals do, leading to **informational inequality**. Preserving privacy helps level the playing field, ensuring that users retain some degree of control over how their personal information is used, thus reducing the power imbalance between corporations and individuals【96†source】.

Q3 (Value-Sensitive Design - VSD): Suppose you are responsible for the Value-Sensitive Design section in a company that deals with algorithms such as the one just discussed. Which empirical, conceptual, and technical investigations would you carry out to avoid cases such as Gillian’s?

1. **Empirical Investigation**:

    - Conduct research to understand the **stakeholders** involved—such as individuals who have experienced pregnancy loss—and gather data on their needs and values. Surveys, interviews, and focus groups would be valuable tools for identifying how these individuals perceive algorithmic recommendations and the privacy of their sensitive information【95†source】.

2. **Conceptual Investigation**:

    - **Conceptualize Relevant Values**: Define and conceptualize key values like **privacy**, **emotional well-being**, and **user autonomy** in the context of personal data use. This would involve understanding how users value the appropriate flow of information, especially during sensitive life events.
    - **Trade-offs**: Investigate potential trade-offs between data utility (for advertising purposes) and privacy, and define acceptable boundaries that prioritize user well-being in cases of highly sensitive events【95†source】.

3. **Technical Investigation**:

    - **Develop Privacy-Respecting Features**: Implement technical features such as **context recognition** that allow the system to understand life changes (e.g., from pregnancy to bereavement) and adapt its behavior accordingly.
    - **Data Minimization and Consent**: Explore technical solutions that allow for **data minimization**, ensuring that only the necessary data is collected, and that explicit consent is obtained when data is used for advertising purposes. Develop tools that allow users to control the flow of their information more effectively【95†source】.

Q4 (Value-Sensitive Design - VSD): Identify the three values to be designed in an algorithm such as the one discussed. Why do you think these are the most important? Mention at least two potential conflicts among them. Give a brief explanation why they are in conflict.

Three Values to Design In:

1. **Privacy**:

    - Users should be able to control how their personal information is used, especially during emotionally vulnerable times. Privacy is fundamental to protecting individuals from informational harm and maintaining control over their personal experiences.

2. **Emotional Well-being**:

    - The algorithm should be designed to consider the **emotional impact** of its content on users. For example, in cases of stillbirth, the algorithm should be aware of the negative emotional impact of displaying ads related to pregnancy or infant products.

3. **Personalization Utility**:

    - Maintaining the **utility of personalization** is also important for providing users with relevant content, which can enhance user experience when properly aligned with their needs and current context.

Conflicts Among Values:

1. **Privacy vs. Personalization Utility**:

    - Maintaining **privacy** may limit the amount of data available for personalization, which in turn could reduce the **accuracy and utility** of the content being presented. Balancing privacy and the ability to offer useful personalized content is a constant challenge.

2. **Emotional Well-being vs. Personalization Utility**:

    - Designing for **emotional well-being** may require the algorithm to avoid certain types of ads (e.g., pregnancy-related ads after a stillbirth), which may conflict with the goal of providing **highly personalized** and potentially profitable content. There is a trade-off between maximizing ad relevance and ensuring the user's emotional state is respected.

Balancing these values requires careful **trade-off decisions** to ensure that personalization is effective while also respecting user privacy and emotional well-being, especially in sensitive situations.

Bias and Fairness

Here are the answers to the questions based on the uploaded PDFs and the given case study:

Q1 (Bias): What form(s) of bias can you identify in this case? Give reasons as to why these are the most likely form(s) of bias.

The forms of bias present in this case include:

1. **Pre-existing Bias**: This bias stems from social inequalities and discrimination that already exist in society, such as systemic **racial disparities in healthcare access and quality**. The algorithm is affected by the fact that black patients tend to have lower healthcare spending, which is a reflection of their access to care being influenced by historical and societal biases【99†source】.
2. **Technical Bias**: The choice of using **healthcare spending as a proxy** for medical need introduces technical bias. This decision was based on data that was inherently biased due to pre-existing social inequities. The metric used—healthcare spending—does not equally reflect medical needs for different racial groups due to underlying systemic inequalities【99†source】.

These biases are present because the algorithm relies on healthcare spending data, which itself is influenced by pre-existing social biases, and thus leads to biased outcomes. The decision to use spending as a metric represents technical bias, while the root cause—unequal access to healthcare—is a pre-existing societal bias.

Q2 (Bias): The article stipulates some reasons as to why this algorithm discriminates against black patients. Could you identify which source of bias this is? Of the remaining two sources,

could you indicate how such an algorithm could discriminate?

The primary **source of bias** discussed in the article is **pre-existing bias**. This arises from societal and structural inequalities that affect black patients' access to healthcare. Pre-existing biases are embedded in the data used to train the model, leading to disparities in risk assessment between black and white patients.

The other two types of bias that could also play a role are:

1. **Technical Bias**: The algorithm's **use of healthcare spending as a proxy** for medical need is a source of technical bias. The data collected and the choice of proxy do not accurately represent the medical needs of different racial groups, leading to discriminatory outcomes. This is an example of bias arising from poor design choices or using incorrect metrics.
2. **Emergent Bias**: **Emergent bias** can develop when the system interacts with new contexts of use. In this case, the algorithm may have worked differently when applied to a broader or more diverse population compared to the training data, leading to biases not initially foreseen during the design phase. It can result in discrimination as the system fails to adapt to differences in how healthcare is accessed by different demographic groups【99†source】.

Q3 (Fairness): Evidently, a biased algorithm leads to forms of unfairness. Could you mention at least two forms of (un)fairness perpetrated by the algorithm? How could these forms of (un)fairness be addressed by the algorithm in order to be reduced (or utterly eliminated)?

Two forms of **unfairness** in this case are:

1. **Outcome Unfairness**: Black patients were **less likely** to be identified as needing high-risk care compared to white patients, even when they had the same level of health need. This results in unequal access to healthcare benefits and programs designed to manage high-risk conditions. This outcome is unfair as it directly impacts patient care.

   Solution: Replace the spending-based metric with a more appropriate measure of **medical need** that is not influenced by race or socio-economic status. This could include clinical markers of disease severity, which are more directly tied to a patient’s health.

2. **Representational Unfairness**: The algorithm failed to adequately **represent the needs of black patients**, which means it underestimates the actual healthcare requirements of a significant portion of the population. The use of healthcare spending as a proxy overlooks how systemic inequalities lead to different levels of spending, thereby perpetuating disparities.

   Solution: Use **demographic parity** techniques to ensure that all racial groups receive an equal chance of being identified as needing high-risk care, regardless of their healthcare spending. This can help ensure that the model's decisions are more equitable across groups.

Q4 (Fairness): Suppose you have been hired by the National Health Agency (NHA) of some country to implement a “risk-care management” algorithm nationwide. Based on the previous experience gained from the algorithm just discussed, which considerations would you recommend NHA to take into account in order to avoid current and potential sources of unfairness? Which provisions would you take up for increasing fairness in risk-care management algorithms?

To avoid sources of unfairness in implementing a “risk-care management” algorithm nationwide, I would recommend the following considerations:

1. **Data Quality and Representation**:
    - Ensure that the training data is **representative of all demographics**. Specifically, it must include diverse patient records to avoid biases linked to underrepresentation. The data should cover different racial, socio-economic, and geographic groups to prevent systemic disparities from becoming embedded in the algorithm.
2. **Appropriate Proxy Metrics**:
    - Choose **appropriate metrics** that accurately reflect medical needs. In the previous case, healthcare spending was used, which reflected economic rather than medical necessity. Instead, I would recommend metrics such as disease severity scores or physician assessments that directly measure patient health.
3. **Bias Auditing and Testing**:
    - Conduct **bias audits** to identify and correct any disparities that may arise in the model’s predictions. Regular testing should be done to ensure that the algorithm treats similar cases similarly, regardless of racial or socio-economic differences.
4. **User Feedback Mechanism**:
    - Incorporate a mechanism for **user feedback** and **contestability**. This allows healthcare providers and patients to flag potentially incorrect predictions, leading to continuous improvement in the model’s performance and fairness.

Provisions to Increase Fairness:

1. **Fairness Constraints**:
    - Apply **fairness constraints** such as **equal opportunity** or **demographic parity** to ensure that patients from different demographic groups have the same probability of being flagged for high-risk care, regardless of their race or socio-economic status【97†source】.
2. **Human Oversight**:
    - Implement **human oversight** mechanisms that allow clinicians to override algorithmic decisions. Physicians should have the final say in care management decisions to account for factors that algorithms may not fully understand.

By adopting these measures, NHA can ensure that the risk-care management algorithm is designed in a way that addresses the fairness issues encountered in the previous case, leading to more equitable and effective healthcare for all patients.