"I'm not convinced that they don't collect more than is necessary": User-Controlled Data Minimization Design in Search Engines

A Interview Protocol

A.1 General User Interviews

Section 1: General Questions

1. Have you heard of the term data minimization? What do you think it means?
2. Do you use any search engines?
3. How do you think a search engine works?
4. How do you assess whether search results are good?
5. What information do you think the search engine collects about you to show you the results?
    - *Follow-up:* How do you feel about search engines collecting your data? Why?
6. Do you think the search engine needs the data that it collects from you?
7. What factors specifically do you consider when making this judgment?

   To improve search results, search engines may collect your GPS location, and personal information, such as name, email address, gender, and birth date.
8. How do you think search engines get a hold of this data?
9. How do you feel about the search engines collecting this additional information?
10. How does this additional data collection impact your thoughts on data minimization?
    - *Follow up:* What are the pros and cons of search engines collecting and processing this additional information?
    - *Follow up:* when was the last time you had a positive/ negative experience with the search engine?
    - What can potentially go wrong with search engine companies' managing or storing users' data?
    - Do you think there's an opportunity to apply data minimization in this scenario?

Section 2: Watch the Video Data Minimization Tutorial, available here: https://drive.google.com/file/d/16fbTaik_uhDiyMT2CWtnQAIbliI-Chn4/view?usp=sharing

Knowledge questions to assess Data Minimization comprehension [participant selects one answer for the following questions]:

1. What is the primary goal of data minimization under GDPR?
    - To collect as much data as possible for future use.
    - To ensure data storage systems are filled to capacity.
    - To limit personal data collection to what is necessary for its intended purpose
    - To minimize the cost of data storage
2. Which of the following practices is NOT in line with the principle of data minimization?
    - Collecting only the data that is essential for a service.
    - Storing data indefinitely just in case it might be useful later
    - Regularly reviewing and deleting unnecessary data.
    - Asking users only for the information needed to complete a transaction.
3. Why is data minimization important?
    - It ensures faster data processing speeds.
    - It reduces storage costs, and security risk and builds trust with users
    - It allows for more data to be stored in databases.

- It ensures that companies have a competitive advantage.

4. Which of the following best describes "adequate" data in the context of data minimization?
    - Data that is excessive and beyond what is needed.
    - Data that is sufficient for the intended purpose.
    - Data that is outdated.
    - Data that is irrelevant to the task at hand.

Correct Answers:

1. C
2. B
3. B
4. B

Section 3: Data Minimization Decision-Making Factors

The search engine creates a personalized profile based on your browsing history, search history, and SERP (search engine results page) clicks. Subsequently, it personalizes your search results based on your interests.

1. How do you feel about your data being collected/used to provide good search results to you? Why?
2. What types of your search query data would you allow the search engine to retain to obtain better search results? We also ask participants to rank their willingness for each type of data.
    - Medical (i.e., disease symptom search queries, medication for diseases)
    - Political (i.e., international politics, specific updates on the political party),
    - Financial (i.e., contact info of your bank),
    - Entertainment (i.e., sport, movie),
    - Personal identification (i.e., photo id, digital id, name),
    - Location revealing search queries (geolocation)
    - Behavioral (i.e., online activities, website visits),
    - Communication (messaging, email),
    - Observational (photographs, recordings of voice)
3. When the participant mentioned certain types of search query data from 2(a), for example, [entertainment, location, etc.], we asked the set of following questions:

    [If [Location] is mentioned in 2(a), then]

    - How would you expect a search engine to use [Location] search query data after you provide it? Why?
    - To what extent do you think the search engine needs [Location] search queries to provide good service? Why?
    - What amount of your [Location] search query data are you willing to share to get good service? Why?
    - How recent data of [Location] search query data are you willing to share? why/why not?
    - Are you willing to have the search engine retain your [Location] search query data, so that it provides improved results for other users?
    - Suppose that the search engine doesn't need past search history data to improve your search results.
        - Would you be comfortable if the search engine retains your [Location] search history data? why?
    - Suppose that we are not certain that past search history data is needed to provide you with good search results.
        - Would you be comfortable if the search engine retains your [Location] search history data? why?
    - If participants don't mention data minimization-follow-up: What thoughts do you have regarding data minimization in the context of search engines retaining your past [Location] search queries when it is unclear if they need it to provide good results?
    - Often asked the set of design suggestions questions from Section 4 of the user interview protocol here to maintain a natural flow of conversation.

    [If [Entertainment] is mentioned in 2(a), then]

    - How would you expect a search engine to use [entertainment] search query data after you provide it? Why?
    - To what extent do you think the search engine needs [entertainment] search queries to provide good service? Why?
    - What amount of your [entertainment] search query data are you willing to share to get good service? Why?
    - How recent data of [entertainment] search query data are you willing to share? why/why not?
    - Are you willing to have the search engine retain your [entertainment] search query data so that it provides improved results for other users?
    - Suppose that the search engine doesn't need past search history data to improve your search results.
        - Would you be comfortable if the search engine retains your [entertainment] search history data? Why?

- Suppose that we are not certain that past search history data is needed to provide you with good search results.
  - Would you be comfortable if the search engine retains your [entertainment] search history data? Why?
- If participants don't mention data minimization-follow-up: What thoughts do you have regarding data minimization in the context of search engines retaining your past [entertainment] search queries when it is unclear if they need it to provide good results?
- Often asked the set of design suggestions questions from Section 4 here to maintain a natural flow of conversation.

4. Are there any circumstances under which you would NOT like the search engine system to collect/process information about you? Why?

Section 4: Design Suggestions Thinking about the search engine,

1. Can you explain potential privacy implications in the aforementioned scenarios of search engines in the context of additional data use/processing?
   - Can you think of privacy implications regarding the **volume of search data** retained by the system?
   - Can you think of the privacy implication of **types of search data** retained by the system?
   - Can you consider the privacy implication of **the recency of search data** retained by the system?
2. Please consider different kinds of potential solutions in the context of data minimization. Can you present your ideas for the solution?
   - What features/functionalities would you like to have in the search engine that will allow you to minimize your data? Or to set your preferences regarding data minimization?
   - Now, considering the concept of data minimization, can you sketch a solution that you would apply to reduce the potential privacy implication you mentioned earlier?

Ranking Questions

Please rank your willingness and preferences towards some types of search queries:

1. How do you rank this information based on your sharing preference (1: less likely to share, 5: more likely to share)?
   - Medical (i.e., disease symptom search queries, medication for diseases)
   - Political (i.e., international politics, specific updates on the political party),
   - Financial (i.e., contact info of your bank),
   - Entertainment (i.e., sport, movie),
   - Personal identification (i.e., photo id, digital id, name),
   - Location revealing search queries (geolocation)
   - Behavioral (i.e., online activities, website visits),
   - Communication (messaging, email),
   - Observational (photographs, recordings of voice)
2. How recent/past search data are you willing to share with search engines for good service/search result for the following types? (None, few day, few weeks, few months, few years, indefinitely)
   - Medical (i.e., disease symptom search queries, medication for diseases)
   - Political (i.e., international politics, specific updates on the political party),
   - Financial (i.e., contact info of your bank),
   - Entertainment (i.e., sport, movie),
   - Personal identification (i.e., photo id, digital id, name),
   - Location revealing search queries (geolocation)
   - Behavioral (i.e., online activities, website visits),
   - Communication (messaging, email),
   - Observational (photographs, recordings of voice)
3. How recent/past search data do you think the Search Engine system needs to provide good service/good search result?(None, few day, few weeks, few months, few years, indefinitely)
   - Medical (i.e., disease symptom search queries, medication for diseases)
   - Political (i.e., international politics, specific updates on the political party),
   - Financial (i.e., contact info of your bank),
   - Entertainment (i.e., sport, movie),
   - Personal identification (i.e., photo id, digital id, name),
   - Location revealing search queries (geolocation)
   - Behavioral (i.e., online activities, website visits),
   - Communication (messaging, email),
   - Observational (photographs, recordings of voice)

4. Which search data are you willing to share, so the Google search engine can provide improved results for other users? (Select all that apply)
    - Medical (i.e., disease symptom search queries, medication for diseases)
    - Political (i.e., international politics, specific updates on the political party),
    - Financial (i.e., contact info of your bank),
    - Entertainment (i.e., sport, movie),
    - Personal identification (i.e., photo id, digital id, name),
    - Location revealing search queries (geolocation)
    - Behavioral (i.e., online activities, website visits),
    - Communication (messaging, email),
    - Observational (photographs, recordings of voice)

A.2 Expert Interviews

Section 1: General Questions

1. Could you briefly describe your current role? and What are two recent projects you have worked on recently?
2. Have you heard of the term data minimization? What do you think it means?
3. Do you use/build/work with any search engines?
4. How do you think a search engine works?
5. How do you assess whether search results are good?
6. What information do you think the search engine collects about its users to show them the results?
    - *Follow-up:* How do you feel about search engines collecting their data? Why?
7. Do you think the search engine needs the data that it collects from its users?
8. What factors specifically do you consider when making this judgment?

    To improve search results, search engines may collect users' GPS location, and personal information, such as name, email address, gender, and birth date.
9. How do you think search engines get a hold of this data?
10. How do you feel about the search engine collecting this additional information?
11. How does this additional data collection impact your thoughts on *data minimization*?
    - *Follow up:* What are the pros and cons of search engines collecting and processing this additional information?
    - What can potentially go wrong with search engine companies' managing or storing users' data?
    - Do you think there's an opportunity to apply data minimization in this scenario?

Section 2: Current Practices and Possible Solutions

1. Are there any specific guidelines or best practices you follow in your organization to remain compliant with the data minimization requirement?
2. Are there any specific methodologies you know of that operationalize data minimization?
    - Are there any specific encryption or anonymization techniques you employ as part of data minimization?
    - Are there any Machine learning techniques you employ as a part of data minimization?
3. What tools or technologies do you currently use to implement data minimization?
4. Can you walk me through the process of how you decide which data to retain and which to minimize?
5. Can you share an example where you had to refactor or redesign a system to better align with the data minimization principle?
6. What are the most common challenges organizations face when implementing data minimization? How did you overcome them?
7. For organizations just starting their data minimization journey, what steps would you recommend?
    - Are there any resources, courses, or communities you'd recommend for developers looking to deepen their understanding of data minimization?
8. How should organizations balance the need for data for business insights and the principle of data minimization?

Section 3: Showing General Participants' Data Minimization Suggestions and Asking About Feasibility

I'm going to present three design concepts, illustrated through screenshots, that depict potential user perspectives on data minimization in search engines. Please take a moment to review each screenshot and its accompanying description. Afterward, we'd appreciate your insights on the feasibility of these designs.

[Each expert went through three screenshots of conceptual designs by end users, randomly assigned].