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

Tanusree Sharma[1], Lin Kyi[2], Yang Wang [1], Asia J. Biega [2]
[1]University of Illinois at Urbana-Champaign
[2] Max Planck Institute for Security and Privacy
{tsharma6@illinois.edu, lin.kyi@mpi-sp.org, yvw@illinois.edu, asia.biega@mpi-sp.org

Abstract

Data minimization is a legal and privacy-by-design principle mandating that online services collect only data that is necessary for pre-specified purposes. While the principle has thus far mostly been interpreted from a system-centered perspective, there is a lack of understanding about how data minimization could be designed from a user-centered perspective, and in particular, what factors might influence user decision-making with regard to the necessity of data for different processing purposes. To address this gap, in this paper, we gain a deeper understanding of users' design expectations and decision-making processes related to data minimization, focusing on a case study of search engines. We also elicit expert evaluations of the feasibility of user-generated design ideas. We conducted interviews with 25 end users and 10 experts from the EU and UK to provide concrete design recommendations for data minimization that incorporate user needs, concerns, and preferences.

Our study (i) surfaces how users reason about the necessity of data in the context of search result quality, and (ii) examines the impact of several factors on user decision-making about data processing, including specific types of search data, or the volume and recency of data. Most participants emphasized the particular importance of data minimization in the context of sensitive searches, such as political, financial or health-related search queries. In a think-aloud conceptual design session, participants recommended search profile customization as a solution for retaining data they considered necessary, as well as alert systems that would inform users to minimize data in instances of excessive collection. We propose actionable design features that could provide users with greater agency over their data through *user-controlled data minimization*, combined with relevant implementation insights from experts.

1 Introduction

Data minimization is a principle specified in Article 5(1)(c) of the European Union's General Data Protection Regulation (GDPR) requiring that "personal data shall be [...] adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed" [3]. It is a legal and privacy-by-design principle mandating that online services collect only data that is necessary for pre-specified purposes.

Today's consumer-facing web-based systems, such as recommender systems, search engines, or Internet of Things (IoT) systems, often provide personalized services through the collection of vast amounts of user data [13, 62]. Data minimization can be seen as a risk management framework, ameliorating concerns about the misuse of data and societal impacts on data-driven economies [74], while balancing the value of data-driven services for individuals and enterprises. A recent techno-legal analysis of data minimization revealed, however, that one of the main obstacles to compliance with this principle in the context of such data-driven systems is the scarcity of appropriate computational operationalizations [24].

An additional challenge lies in the fact that the definition of data minimization ties the necessity of data to the purposes of data collection – and how to formalize this relationship is still largely an open question in data-driven systems. Several recent computational approaches to data minimization in personalized systems propose, for instance, to tie purpose collection to improvements in the service [14,57]. However, such an automated approach may not always capture the context-dependent perception of the necessity of data by users [24]. Hence, *contrary to traditional system-based approaches that attribute the sole responsibility of data minimization to service providers, it might be vital to involve users in shaping the appropriate data minimization practices*.

While previous literature has highlighted the importance of data minimization in protecting user privacy [14, 24, 57] and complying with regulations [31, 66], there is a lack of understanding about how data minimization could be designed from a user-centered perspective, and in particular what factors might influence user decision-making with regard to the necessity of data. Consequently, to address this gap and to complement existing system-centered approaches, our paper explores a user-centered approach to compliance with data minimization in data-driven systems.

Towards this goal, in this paper, we conducted semi-structured interviews with 25 EU and UK-based end users to examine users' understandings and impressions, decision-making factors, and conceptual design recommendations for implementing data minimization measures. We then interviewed 10 experts, also located within the EU and UK, specializing in technical, legal, and research aspects of search engines, data minimization, and privacy, to assess the practicality of the design concepts proposed by users. Acknowledging the context-dependent nature of data minimization implementations, we grounded our interview in the domain of web search engines – a prototypical example of a data-driven system utilizing, beyond individual user attributes, behavioral and observational data, such as clicks and queries. We investigate how users expect data minimization to be designed in search engines. More specifically, we address the following questions:

- **RQ1**: How do users think data minimization *currently* works in search engines?
- **RQ2**: What factors influence user decisions on what data is necessary for search engines to collect in the context of data minimization?
- **RQ3**: How do users think data minimization ought to work in search engines?
- **RQ4**: How do experts evaluate the feasibility of user-generated designs of data minimization in search engines?

Findings. Our study indicated an evolving perception of data minimization as some users shift from traditional search engines (such as Google) to the next-generation search engines (such as ChatGPT and Gemini). Participants generally held the belief that these newer engines inherently implement more effective data minimization strategies. When it comes to data minimization decision-making, participants considered several factors, including the type of information being searched, their expectations of good service, and the amount of data assumed to be necessary for a given search data type. Our findings indicate that users imagine trade-offs between the amount of information collected by the service and the resulting improvements they can expect in the search results. In particular, for political, and medical search data, participants favored minimal personalization and maximal data minimization due to the sensitive nature of the information. However, our results also indicated that some participants believed data minimization might not be feasible and that, practically, they are paying for services with their data. Additionally, participants expressed the need for minimization features either in the general system settings or just-in-time on the search webpage, to customize their strategies for minimization for different types of search data. The quantity of data and the temporal proximity of search data also influenced the participants' perceptions of necessity with respect to data processing purposes.

In our user interviews, we identified several practical design features for data minimization in search engines. For instance, sensitivity-adjusted keyword search, detailed customization options for data selection based on type and amount, and more straightforward opt-out methods for specific data collection scenarios. However, experts also raised concerns that some user-suggested data minimization designs could inadvertently increase privacy risks instead of enhancing privacy. For example, features that separate search data based on whether the search is for oneself or others, could potentially lead to more detailed data (e.g. binary annotation of oneself vs other) being collected, thereby increasing the likelihood of re-identifying uninvolved parties (such as children or parents).

Contributions. Overall, the contributions of this paper are:

(1) Our user interviews ($n = 25$) identified how participants perceived (*i.e. their prior understanding of*) data minimization, which is under-studied in the current literature.

(2) We identified participants' decision-making factors when it comes selecting data minimization strategies under different scenarios.

(3) Based on user interviews and expert interviews ($n = 10$), we suggest actionable and practical data minimization measures to address users' needs, concerns, and preferences regarding the necessity-quality trade-offs in data collection and processing.

To our knowledge, this is the first study that seeks to develop data minimization interpretations based on end-users expectations of data necessity, complementing the traditional system-centered approaches. Our findings emphasize the need to consider user preferences when regulating and operationalizing data protection principles.

2 Related Work

2.1 Legal Background of Data Minimization

In a data-driven economy, user privacy is often the cost for using data-based systems [57, 74]. Therefore, to protect users, privacy regulations often require that data controllers, those deciding how user data is to be collected and processed [2], exercise *data minimization* measures which involve collecting only the minimum amount of data necessary to fulfill a particular task [55, 57].

The EU General Data Protection Regulation (GDPR) outlines the principle of data minimization in Article 5(1)(c) which requires that data should be *"adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed."* This means that firstly, user data may only be processed when it is *adequate*, meaning only necessary data required to fulfill a task may be collected. More data may be processed, for instance, if the existing data is not adequate to represent groups that are underrepresented [57].

Secondly, the data collected from users must be *relevant* to the purposes of the task. Thirdly, data collection must be *limited*, where only data necessary for the task may be processed. In addition, the data minimization principle also involves i) limiting the categories of data that are collected, ii) limiting the duration for which personal data is retained, and iii) deleting data after it has been used for its intended purpose [22, 71].

Consequently, Article 4 of GDPR provides a broader definition of personal data, encompassing *"any information relating to an identified or identifiable natural person."* As demonstrated in prior research, movie ratings, such as those found in MovieLens 20M dataset [33] facilitate identification by linking private and public datasets [47]. Thus, it is reasonable to infer that a significant portion of data employed in data-driven systems qualifies as personal data and thus falls under the purview of the GDPR dependent on its geographic scope. To rectify privacy concerns, data minimization is useful as data controllers can still collect and retain data that are needed for a certain period of time and provide expected services.

2.2 Operationalizing Data Minimization

There has been speculation that data minimization is incompatible with the current data economy as it may inhibit technical advances in machine learning which are reliant on repurposing user data [36,44,57,71]. As a result, data controllers may take advantage of some loopholes to evade the data minimization principle, such as *pseudonymizing* data, which means preventing re-identification of data subjects using technical means [71]. Unfortunately, pseudonymizing personal data may reduce the quality, since the data would be altered and it would be difficult to aggregate different datasets [49].

Despite these speculations, previous work has shown that is it possible to strike a balance between data minimization and high-quality machine learning models [1, 14, 57]. For example, Biega et.al. [14] proposed a performance-based approach to data minimization that reconciles the principle of data minimization with performance metrics, addressing the lack of a consistent interpretation of data minimization and the complexities associated with personalization. Similarly, the FIDO Framework utilizes machine learning and law to limit data collection by iteratively updating an estimate of the performance curve, providing accurate stopping criteria, and offering practical recommendations to implement data minimization [57]. Another approach to reduce personal data for machine learning predictions is via feature removal or generalization using knowledge distillation [28]. This approach has the potential to preserve accuracy while meeting data minimization regulations. Additionally, Senarath et al. propose a way to minimize user data in software from a design science perspective to encourage developers to consider privacy engineering for data storage and sharing and evaluate the compatibility of this approach with existing practices [55].

2.3 Data Handling in Search Engines

Most information retrieval studies explored users' search behavior and evaluated systems [68] along a number of dimensions including search complexity or time-sensitivity of search [43] to demonstrate how variations in search tasks properties can impact users' behavior [16, 63, 65]. Personalized search or recommender systems often use browsing sessions, SERP clicks, and user information to build profiles of user interests [26]. This data combined with user queries, has potentially high value with respect to personalization [61] and advertising [64] and can be traced to an individual via linkage of *pseudo-anonymous* data, such as the IP address of the home router being used. Several search engines (e.g. Duck Duck Go) now promise not to collect personal data and web browsing tools (e.g. Tor) are also available to prevent identification, however, evidence suggests only a minority of people make use of 3rd-party blocking software to handle their data in search engines and, as a result, data processing is often not clear [20, 58, 70]. To address this, studies explored client-side personalization [60], privacy warning labels in search systems inspired by decision-making research on food nutrition labels [73], intent-aware query obfuscation [8], or creating distorted user profiles with noise [17].

When it comes to user preferences regarding data handling by search engines, one study highlighted that users in India exhibit a slight preference for personalization in their search results but are usually willing to give up personalization when searching for topics they deem sensitive [50]. While examining data protection practices in search engines are limited, some closely aligned research in online advertisement and web-tracking indicated the complexity of configuring for opt-in/out and privacy settings, confusing interfaces, and tools' default settings which are often minimally protective [32, 41]. Studies on consent have found that users do not often agree with how their data is being used by data controllers [37, 39], and would like less data to be collected for purposes they do not deem to be necessary, such as personalized advertising purposes [39]. Unfortunately, users who deny consent for data collection are penalized for this and have been shown to receive unjustified, lower prediction scores, raising concerns over how users who wish to protect their privacy may be negatively impacted by not consenting [35].

A notable example of a data minimization strategy was introduced by Google in 2020, with web activity of new users being deleted by default after 3 or 18 months [29]. Our results in this paper, however, surface the need for users to have more fine-grained control over data collection and retention practices, depending on data type, individual personalization preferences, and other factors.

2.4 Decision-Making in Privacy Contexts

Users' privacy behaviors are not consistent, nor are they rational [6]. An example of this is the *privacy paradox*, where users tend to find tracking and other technologies creepy and invasive, yet their behaviors are not in line with their beliefs, and

sometimes may behave against their privacy beliefs [6, 12, 18]. Often, users are subjected to their individual constraints, such as time, effort needed to protect their privacy, knowledge of privacy, among other factors, which results in users having to trade the long-term benefits of protecting their privacy in exchange for short-term benefits, such as convenience [7].

In the context of personalized services, there is often an *acceptability gap*, where users tend to be more accepting of using personalized services, but are less accepting of sharing their data for these services [37]. Users are often not interested in paying money to block tracking or ads, but would like more control over the data they share, and whom this data is shared with [18]. Unfortunately, current methods aimed at giving users more choice over their data, most notably through cookie banners, are not effective [38, 45, 48]. Instead, users rely on the site's reputation and the services it provides compared to the text of a cookie banner [38]. This overload of information available to users often forces users to engage in *privacy calculus*, which is where users consider the costs and benefits of any information disclosure before deciding to provide their personal information. A user's behavioral intentions are based on the expected benefits and costs of that privacy violation [19, 42, 69]. In this paper, we explore the decision-making strategies users adopt for search engines in the context of data minimization.

3 Method

In this section, we outline the method to study and analyze search engine users as well as experts in fields related to search engines, data minimization, and privacy.

3.1 Recruitment

User interviews. We recruited participants through Prolific [4]. Participants had to be: a) 18 years or older; b) living in the EU or UK; and c) have experience using a search engine. The principle of data minimization under the GDPR is considered an individual's right over their data for the European Union (EU) and UK, thus we recruited participants exclusively from these regions. Participation was voluntary, and participants were allowed to quit anytime. Each participant received £20 upon completion of an hour-long interview.

Expert interview. We interviewed experts for their input about how to implement data minimization in practice, and to provide feedback on the feasibility of our participants' data minimization suggestions. Our inclusion criteria for experts was they had to have expertise in either technical, legal, policy, or research in the domain of search engines, data minimization, or privacy, either in academia or industry, and be living in the EU or UK. We leveraged our professional connections, research communities (e.g., SIGIR), and Prolific. We received 12 responses in total: four were from Prolific, three were from the SIGIR community with a research background (e.g., differential privacy, conversational search system, law/policy for data minimization), and five were from industry backgrounds including, internet technology, banking, health insurance, and donation distributor. All participants were from the EU or UK. Each participant received £20 upon completion of an hour-long interview.

3.2 Participants

Users. Out of our 25 participants (Table 1), 48% self-identified as women, and 52% as men. The majority of the participants (44%) were in the age range of 30-40, followed by 20% were 20-30 years old, 20% were over 50 years old, and 16% in the age range of 41-50 years old. The participants were from various European countries, namely Poland, Portugal, Spain, Germany, the Netherlands, Italy, Estonia, Ireland, and the United Kingdom. All participants had attained at least a high school diploma, with 37% participants holding a Bachelor's degree and 27% holding a graduate degree. This aligns with prior research showing most users on crowdsourcing sites are typically between 20 and 30 years old and highly educated [5, 34, 56] A majority of the participants (60%) reported having a full-time job, while 26.67% of them had a part-time job, and the remainder were either unemployed or not in a paid job. All participants reported using mobile or desktop devices on a weekly basis, with the majority (63%) indicating multiple daily usages. All participants reported using Google as their primary search engine, with some mentioning the occasional use of Bing, Mozilla, Microsoft Edge, and local search engines (e.g., Sapo) for specific countries. Table 1 presents the demographics of our participants. We refer to end users as P1,. . . ,P25.

Experts. Participants were from Germany, the United Kingdom, Portugal, France, Netherlands. Seven of them were men and three were women. The majority (5) of them were 30-39 years old, followed by 40-55 years old (3) and 26-29 years old (2). Table 2 presents the demographics of our participants. We refer to experts as E1,. . . ,E10. Most participants had two or more years of experience, with three of them having 10 or more years of experience. Their current role involves software product managers, database security, IT security in banking, health, software engineering, and policy research (Table 2).

3.3 Pilots

We conducted four pilot interviews with general participants, and two with expert participants to test our study design. Three of the general participants were graduate/undergraduate students and one worked in banking. Of the two experts, one was PhD researcher and one was a privacy engineer in a technology company. We adjusted our user and expert interview questions based on pilot feedback and our observations.

3.4 Semi-Structured Interview Procedure

General user interview. We designed the general user interview protocol based on the research questions outlined in

Section 1. The full interview script is available on GitHub.[1] The protocol was structured according to the following topical sections:

1. Prior understanding of data minimization and search engines;
2. Watching an educational video on data minimization followed by a knowledge check to assess participant understanding of data minimization;
3. User expectations regarding data minimization in search engines; and
4. User design suggestions for data minimization in search engines.

In the first section, we asked about participants' understanding of data minimization and search engines. When participants indicated certain usage of search engines, we asked participants-*"How do you think [earlier mentioned search engine] works?"* We further asked about their thoughts about information collected by search engines after providing context, such as-*"To improve search results, search engines may collect your GPS location, and personal information, such as name, email address, gender, and birth date."*. We then asked them how this data collection of search engines impacts their thoughts on data minimization. The majority of the participants could not provide substantial responses in the context of data minimization, therefore, before starting the second part of the interview, we showed them an informational video on data minimization. This video encompasses an overview of the concept of data minimization, describing the principles of relevance, limitation, and adequacy and processing of personal data within the framework of the General Data Protection Regulation (GDPR). We created the video with accessible visuals and audio to make the concepts easily understandable by users. Given that Section 2 of the interview was designed to answer RQ3 on investigating users' preferences regarding the implementation of data minimization in search engines, it is crucial for participants to possess a basic understanding of data minimization, despite the fact that end-users often lack familiarity with the associated regulations.

Before proceeding with the second part of the interview, we assessed participants' understanding of data minimization with knowledge questions in a multiple-choice question format as well as qualitatively. Most participants correctly responded to the knowledge questions. Every participant accurately identified the main objective of data minimization under GDPR as limiting personal data collection to what is necessary for the intended purpose. Regarding the question *Why is data minimization important?*, 95% correctly stated that it reduce the risk of data breaches and safeguards individual privacy. For the question "Which practice does not align with data minimization principles?", 90% correctly chose the option *"Keeping data indefinitely on the off chance it might be useful in the future."*

In the second part, we asked a series of questions to understand participants' strategies for how they would share different types of search information (i.e., medical, financial, political, entertainment, etc) with search engines to obtain good search results. This was to identify contextual differences in their expectation and rationale for decision-making. We asked how much data they would be **willing** to share with search to get good service and what amount of data they think search engine **needs** for a specific purpose. Another functionality of search engines is that they improve results for other users with similar search behaviors or interests. Thus, we asked if they are willing to share their search data to improve results for other users.

To examine shifts in decision-making tendencies, we incorporated a scenario wherein data minimization was framed as the *necessity*, for instance, *"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 retained your [location, medical, political] search history data? why?"* In the third section, we encouraged participants to consider possible contextual privacy implications of search engines based on the amount and types of search data used or processed. Then, we asked them to consider potential solutions and present their ideas. We instructed participants to generate solutions by sketching their ideas on pen and paper, using a think-aloud protocol [21]. The purpose of this exercise was to understand their thought processes and visualize the measures they would implement to minimize their data.

Expert interview. To evaluate the practicality of data minimization designs suggested by users, we interviewed experts in technical, legal/policy, and research areas related to search engines, data minimization, and privacy. We started by asking about the expert participant's current role and the recent projects they have worked on. Then, we posed similar general questions we asked users on search engines and data minimization. We then asked about their organizations' practices for complying with data minimization requirements and the technical measures they implement.

Finally, we showed experts three user-generated design sketches and the description of the sketch from the first round of general interviews. We asked them to assess the sketch's feasibility in the context of search engines, considering current technical, legal, and infrastructural available to implement data minimization. To scope the interview in one hour and ensure a balanced evaluation, each expert reviewed three randomly selected sketches from a total of five and all the five sketches were evenly distributed among experts. This allowed all designs to receive a comparable level of expert analysis.

[1] https://github.com/Sree0270/usenix2024-supplimentary

3.5 Data Analysis

Both user and expert interview data were obtained through the audio of the interviews recorded on Zoom upon participants' permission and transcribed. We collected each participant's interview audio, think-aloud responses of possible solutions, and image files of design sketches they created during the session. We performed a thematic analysis of our transcriptions [15, 23]. Two researchers independently read through the transcripts of 20% of the interviews, developed codes, and compared them until we developed a consistent codebook. The inter-coder reliability of the two researchers' codes was calculated (Cohen's Kappa = 0.91), which is considered good [25]. They met regularly to discuss the coding and agreed on a shared codebook before coding the remaining data. After completing the coding for all interviews, both researchers spot-checked the other's coded transcripts and did not find any inconsistencies. They grouped lower-level codes into sub-themes and further extracted main themes. Finally, they organized codes into higher-level categories.

4 RQ1: Participants' Impressions and Experiences with Data Minimization

Current search engine landscape & data minimization. When discussing search engines, many participants described them as having *"algorithms," "existing databases and indexes," "internal APIs," and "search keyword to web mapping."* They often equated search engines with recent advancements in large language models, such as ChatGPT, Gemini, and Claude, viewing them as next-generation search engines (while Google and Firefox are seen as traditional search engines). P1 said, *"Bard (now Gemini), ChatGPT automatically minimize data because they don't need my information to provide results as the others do."* Despite this, they mentioned why they still rely on traditional search engines for tasks for reliability, as P10 explained- *"If I know what I am searching then nowadays I go for ChatGPT search. I'm a DevOps and I use AWS services a lot. To configure a load balancer on Amazon, I used Google to find articles and many links to find the relevant ones. Now I can find that on ChatGPT search just configuring an elastic load balance. But for specific code snippets for development, I don't yet rely on GPT. I still look for search engines, StackOverflow sites because the ChatGPT search system relies on datasets that are a few years old, and give outdated, incorrect syntax or deprecated code. So I go to Google or Edge."* Participants' perception of LLMs as similar to search engines or conversational search engines [52] are arguably accurate with notable differences [10] such as that LLMs do not index and retrieve real-time from the web directly [53], are context-aware, and provide human-like responses [54]. Some studies suggest that LLMs augment search experience for higher user satisfaction when querying [46], or in domain-specific search [27].

We found that participants tended to see tradeoffs between performance and privacy when choosing search engines among various options, including, Google, Bing, Firefox, DuckDuckGo, Brave, and Microsoft Edge. P13 said *"I think Google is getting a bit nosy, all those ads. But you know you make the tradeoffs with performance. They don't even think about data minimization."* In contrast, P20 finds Bing more appropriate to her search as she said *"I can pinpoint it (search results) more with Bing. Google would try and show me the top 5 (results) based on what they think is right or how those sites pay them and collect my device ID, name, gender, everything. Bing is only matching with my keywords. So it might have some (data) minimization thingy."* On the same note, P14 appreciates DuckDuckGo and Brave for the agency and control *"Brave provides more privacy and even provides an incentive if I choose to provide my data so I have a choice there. I would imagine this as (data) minimization because you can say 'no' or say 'yes' and get an incentive."* P24 said that Edge is providing incentives for an Xbox game by asking for user data, *"I use Edge to buy games. Edge doesn't work well like Google, and you need to give very specific keywords. But that's because they don't collect data, and they are working on improving with this incentive program to ask you to spend more time on their engines so they can train the model."*

What are the metrics for good search results? Data minimization is defined with respect to data processing purposes, and prior work has suggested that the purpose of data processing in personalized data-driven systems is service quality improvement [14, 57]. Thus, an important consideration in user-driven data minimization is how users might judge the quality of results. In our study, participants identified several criteria for what constitutes a "good search result." These include an exact match of keywords in top-ranking results, the credibility of the website URL, presence of location-specific results, and a diversity of results.

P7 discussed the significance of topic-specific searches, particularly when researching scientific topics. She expects to see links from scientific journals rather than blogs or untrusted sites, noting, *"If I'm researching a scientific topic, I tend to trust results from scientific journals. Currently, seeing results to unfamiliar sites as a response to those searches which I don't trust and consider as good results."* Similarly, P11 highlighted the value of diverse search results, especially when shopping for products. *"product from various suppliers, showing different prices, different links, providing more options."* A key issue emerged from this discussion where many participants, including P13 mentioned the challenge of balancing specificity in search queries. This leads to the question of how much information is needed for reasonable results which is often unknown to users. As P13 said, "Being too specific might exclude useful information as Bing does while being too vague results in an overload of irrelevant data as Google does. Perhaps, need a way to determine how much user data is used in the backend."

5 RQ2: Factors Influencing Data Minimization Decisions

We present how participants expect data minimization to function in search engines. We asked participants three key questions in the light of preference, willingness, and perceived need for search engines to provide good results. The analysis showed that their decision-making process involves trade-offs of different factors related to data minimization, including different types and volume of search data.

5.1 Types of Search Data

Most participants had reservations about medical, political, financial, behavioral, and personal identification-related search queries, preferring to limit data collection for these sensitive topics. P2 discussed current volatile world politics to illustrate this: *"So many crises, like COVID-19, now the Palestine-Israel conflict, those are tricky. If my search says I am supporting Israel or I'm supporting Russia, later this data could be used against me."* P6 similarly was expecting medical searches not to be collected, mentioning the potential negative impacts on insurance benefits due to chronic or psychological conditions. We found participants often relied on trusted sources rather than search engines for financial information, expressing concern for potential risks of fraud. P25 illustrated this by sharing an experience where, after searching for cryptocurrency investments, they received suspicious emails and messages, including a dubious link to redeem an Airdrop. Despite recognizing the risk, P25 clicked on one of these links. This experience led P25 to question the necessity of search engines collecting financial query data, as they preferred to make investment decisions independently without guidance from search engines.

Regarding location and behavioral searches, participants had specific preferences. While participants were open to search engines retaining general location data (like city or country), they avoided sharing exact locations. P13, for example, used tactics like setting her home address to her neighbor's to mislead search engines: *"When I use Google Maps to search for nearby stores, I set my home address to that of my next-door neighbor so the search doesn't have my actual location."* Opinions on behavioral search data were mixed. Some saw the benefits of product purchasing and receiving relevant ads based on past searches, while others found it invasive and irrelevant. This highlights users' expectations regarding the minimization of data by search engines are primarily influenced by the type of search and further tailored to the individual's specific circumstances.

Some participants saw value in sharing certain types of medical, behavioral, and personal data. For instance, P2 mentioned how his search on certain medical symptoms could improve the search for others or help different medical institutes match website content so those are easily found by users' search keywords: *"Search queries in a medical context is okay because this can help some medical institutions to have a good database. I understand from my experience how difficult sometimes to find relevant information on a legitimate medical topic."* Similarly, P7 was open to sharing personal details like professional experience and gender, seeing it as necessary for obtaining more accurate search results: *"LinkedIn and search engines need my professional details if I am searching for a job in event management, banking, or in a tech company, etc. Some personal data will help avoid getting unrelated search results."*

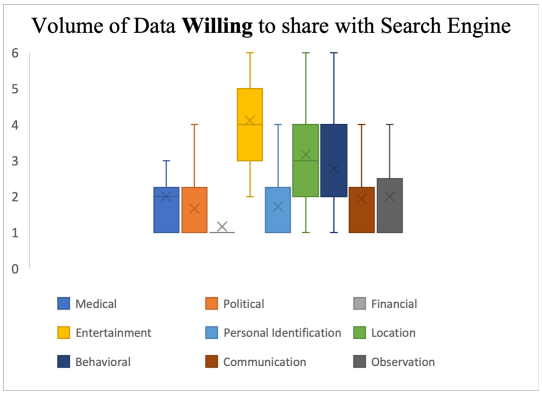

Figure 1: Willingness to share data with search engines: How long are you willing to let a search engine store the following types of data to get a good service/search results? (1: Not at all, 2: A few days, 3: A few weeks, 4: A few months, 5: A few years, 6: Indefinitely).

Participants also ranked their willingness to share various types of data to improve search results for themselves vs others (Figure 7 and 8 in Appendix). For instance, participants were more willing to have their medical data retained when it improved the results for other users (for example, when affected users search for symptoms of rare diseases) compared to just improviding their own results.

5.2 Volume of Search Data

Participants conveyed their comprehension regarding the amount of data required for good results from search engines. They discussed: (1) what they consider as reasonably good search results; and (2) how they perceive the amount of data needed to generate such results. P21 expressed his perspective on the relevance of past political and location data searches and the amount of data required for providing satisfactory service by stating, *"I understand why search engine needs my regular or daily location data, for example, if I were looking for restaurants nearby my place or new locations, But for politics, I don't think my preference change ever and I don't need my results needs any improvement. So I don't see why they need search queries on whatever I search about relating to global politics of wars, crises, and local politics."*

To further elucidate the participants' viewpoints, Figure 2

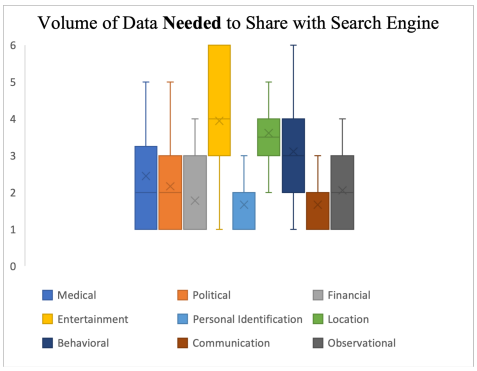

Figure 2: Necessity to share data with search engines: How long do you think a search engine needs to store the following types of data to provide you with a good service/search results? (1: Not at all, 2: A few days, 3: A few weeks, 4: A few months, 5: A few years, 6: Indefinitely).

presents self-reported quantitative value for each type of search data and their understanding of the duration for which data needs to be retained. For instance, financial information was rated with a mean score of 1.77 for how long the data needed to be retained for good search results, followed by political (2.16), and medical (2.16) information. In contrast, some participants thought search engines needed past search data related to user location, behavior, and entertainment to provide good results.

5.3 Tradeoffs Between Performance and Data Minimization

Many participants discussed their decision-making regarding the minimization of data collection and usage for their intended service. They considered the trade-offs between the level of information and the extent of improvement in the search result that could be achieved by sharing their search data. P21 reported being satisfied with the current search results when searching for medical symptoms while he preferred to rely on medical professionals for detailed consultations rather than search engines. As P23 stated, *"I am satisfied with the current search results. I do not believe that the search engine needs any further information to store and improve results. I already know how to get to a trustworthy website and have visited diabetes.co.uk several times before. If I need more medical advice, I will prefer to consult a doctor rather than depending on the search."*

Some participants recognized a trade-off between the necessity, usage of service, and trust when considering data minimization. Despite understanding the risks associated with storing and using search data, they accepted it as part of using a free service, viewing data collection as a form of **payment for the service**. P3 expressed this by noting the importance of trusting the service providers to some extent when using their service *"I think most of it is necessary. The risk concerns the use of such information, but if we want to use these services, we must also give some trust."* Meanwhile, P16 highlighted the dilemma between data minimization and service usage, pointing out that avoiding sharing personal data could mean refraining from using online services altogether *"Data minimization or service. Even though I don't want to share and I think its not needed, but only way to avoid using online."*

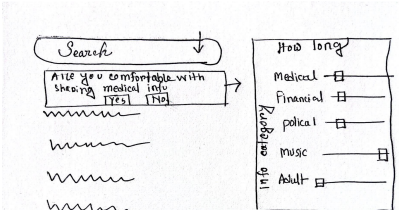

Figure 3: Profile customization or just-in-time pop-up to configure data minimization preferences for different data types. This design consists of a form with different categories of search queries with (i) a checkbox for enabling/disabling the collection of data of a given type and (i) a slider for configuring the volume and retention duration of data of a given type.

6 RQ3: Conceptual Designs

In this section, we introduce participants' conceptual designs (CDs) for data minimization in search engines. Five such designs were identified through user-provided think-aloud protocols and design sketches. Following this, we present the practicality and feasibility of these designs for data minimization in search engines from experts assessment.

6.1 User-Driven Data Minimization Design

6.1.1 CD1: Profile Customization for Data Types and Data Volume

Most participants suggested a user-end profile customization option for setting the amount and type of search data, which they think would be needed for a reasonable search result. They proposed two customization methods: (1) a profile status customization in the settings page, allowing users to select different data (e.g., medical, politics, music, etc) and volumes (e.g., current, past few days, months) for various data types using checkboxes; (2) a just-in-time pop-up near the search bar for users to specify their data sharing preferences based on type and volume. To illustrate this model, P3 presented a sketch that depicted a just-in-time pop-up design for users to input their preferences (Figure 3). As P3 noted, *"Some kind of checkbox or anything similar where I can click or slide on for information category and storing preferences. So if I'm searching for a movie, for example, which is in category of entertainment, there will be pop up right side of search bar where I check how much data to collect, none to all."*

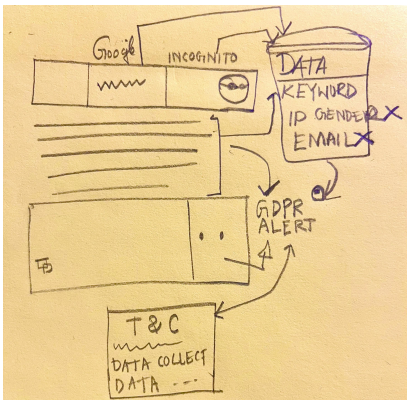

Figure 4: Designing a GDPR alert to provide users an accessible way to see data collection and retention. This would consist of an automatic detection of irrelevant data collection by search engines and cross-checking with their privacy policy page regarding data collection and retention.

Figure 5: A design that separates searches that the device owner conducts for themselves and others. The idea is to minimize data collection (preserving the privacy of other users) while maintaining personalization and accuracy of search results for the device owner.

6.1.2 CD2: GDPR Alerts for Data Minimization

Many participants suggested implementing an extension-like feature in search engines to provide clear explanations when data collection or retention exceeds what's necessary and relevant. Some expressed uncertainty about what data is considered essential for collection and the duration of its retention. As P10 remarked, *"Terms and policy (privacy policies) might have a retention and collection policy, but who reads those? Those pages seem like they are pushing you during service usage to click on 'I agree'."*

To this end, P6 proposed a solution (Figure 4) for data minimization in terms of collection and retention: *"A search engine page or a stand-alone extension could probably use some sort of firewall or program that would alert you to what kind of personal engine is collecting and keeping. And somehow the GDPR principle could be coded there so the rules cross-check the list of collections and give alerts to users if unnecessary data is being collected. This is certainly not completely new, but can minimize data."* P19 expanded on this idea by proposing the addition of a chatbot in place of GDPR alerts which would communicate information to users and help them manage their data-sharing preferences through a conversational system.

6.1.3 CD3: Separating Searching Sessions for Oneself and Others

Data minimization (anonymity) for others. Several participants expressed concerns about receiving ads in the search results that were not relevant to their own search history. P21 highlighted that after searching for gifts for friends, he received constant ads related to their search result on the top three results. To address these issues, P12 suggested the implementation of two tabs within the search engine (as shown in Figure 5), enabling users to define whether a search was intended for themselves or others. As P12 stated, *"In search engines, you type into the search box, what's the best medicine for eczema? And then it could be something within you press, a little drop-down box– is this information pertinent to you or for someone else, you click a relevant box, could be a quite effective way."* Similarly, P22 suggested separating users' search history and making search-related information entirely anonymous, ensuring that search channels remained focused on the users' needs.

Additionally, several participants expressed concerns regarding children's privacy when using search engines and recommended that separate search options be made available when children are using the same device. P1 suggested *"Google search could do the same thing as YouTube kids, and then they can't collect data from children's searches because it's illegal for search engines to collect data from kids."*

Data minimization (dimension reduction, PCA-like feature) for performance. Some participants mentioned data minimization in the context of better performance and non-biased search results. They noted that their use of search engines extends beyond personal needs, such as searching for gifts for friends or information for siblings. As P24 said, this could create an outlier in the search data and potentially influence the overall data profile.

One potential improvement for search engines could be adding an option under the search bar to specify whether the search is for personal use or for others. This distinction could help in managing the relevance of collected data and ads shown. To illustrate the design, P24 described the concept of principal component analysis (PCA), a mathematical technique: *"I hope search engines could apply PCA to simplify data in multidimensional (many users) by reducing the complexity by retaining only the necessary ones. With mathematical expressions under the hood, there could be a UX design for users only a click way to reduce the data retention*

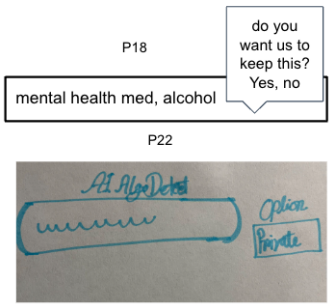

Figure 6: Minimized data collection based on keyword sensitivity in a search engine: redirecting users to a data-minimized search mode or showing a just-in-time pop-up when a sensitive search query is detected.

from the user end."

6.1.4 CD4: Sensitivity-Based Search Customization

Many participants expressed the need for a search option based on the level of sensitive words. For example, they expected a search option that would prevent search engines from collecting and storing information related to personal topics, such as disease information, and children care, violence (e.g. guns, alcohol) related topics out of curiosity. P7, P18, and P22 suggested that the search engine's underlying functionality could be designed to identify search sensitivity based on the keywords entered into the search bar and prompt or remind users to switch to private mode or simply ask users with a pop if those searches should be collected or not. This would allow for a more dynamic and user-friendly private search experience. As P18 states, *"They could just have a pop up triggered by the sensitive word on the search bar and ask me if those to collect or not"*. P18 added that this option would help easily allow or disallow sensitive data collection during a search. Similarly, P22 said *" they could just remind me to use private mode based on sensitive keyword identified."* To illustrate, P22 presented a sketch design for *"**sensitive search**,"* (Figure 6) in which there would be buttons for normal search and search without collecting data (sensitive search).

6.1.5 CD5: Data Donations & Incentives

Some participants, particularly those using Bing, Microsoft Edge, and Brave, discussed an incentive model where users are asked for permission to collect certain data during searches. For instance, Brave rewards users with cryptocurrency tokens, while Microsoft Edge offers points based on the time spent on the search engine. These participants suggested implementing a similar structure in mainstream search engines, allowing users to see the data being collected and choose what to share, potentially receiving credits for their data. This highlights the ownership and control design expectations of users. To expand on this, P16 said- *"Currently Firefox, Google Search doesn't have any ownership structure for the user and no accountability to the user. I don't even know if my medical queries are collected. I generally don't want them to collect these queries and links I visit, but, probably there is a chance if they give me some choice of share or no share. even say this info will help build some medical databases for researchers, and doctors. I perhaps would give a thought on providing."*

6.2 Expert Evaluation of the Practicality of User Design Suggestions

6.2.1 Experts' Current Practices for Data Minimization

Based on the expert participants' area of expertise, we found several practices related to Data minimization and privacy.

Data encryption & anonymization. All our expert participants frequently mentioned data encryption as a part of data minimization practices. As E1 noted as their company practice,*"I work on data encryption to secure information during both transmission and storage when collecting data and inverse query processing to restore information and ensure the storage of personal and necessary information only exclude irrelevant data that doesn't contribute to the system's functionality. Another aspect of our approach is the use of anonymization techniques and tracking user sessions to enhance privacy. I employ the PVC method to make it more challenging to associate data with specific users. This is to communicate privacy policies and terms of service to users.*

Consent before collecting sensitive data. Interestingly, all expert participants from the industry emphasized the need for sensitive or personal data for the performance of the service. However, for the purpose of data minimization, they mentioned consent while asking for any personal data. As E9 noted *"We get consent for collecting any sensitive personal data. Our policy is - clear on data retention and making sure my team does the same. We have protocols to **automatically delete** or anonymize user data after a certain period to minimize risks associated with long-term storage."*

Avoiding centralized execution of data. As a part of data minimization practice, a few experts mentioned decentralized execution on users' devices as a best practice in their product. As E7 said *"Our service considers processing data on the user's device whenever possible because it's health-related data, we reduce the reliance on centralized data storage. We have a client side to boost privacy to limit the exposure of sensitive information. This is what we do for data minimization."*

GDPR compliance for organizations. Three experts who had 10-15 years of experience in the industry as damage security, information security, and financial security experts in a European bank mentioned that they followed the principle of data minimization before GDPR existed. They mentioned the economical aspect of keeping data long period of time for cloud, computation, and maintenance costs. As E2 mentioned, *"We ensure data accuracy and collect only what is necessary for*

our specific purposes, like fundraising. For instance, knowing someone's pet's name isn't relevant to fundraising, so we avoid collecting those. isn't just about respecting privacy; it's also cost-effective for maintenance and quality control. we are conscious of how long we retain data. For example, after a fundraising event, I assess whether the collected data is still needed or if it can be deleted or archived. It's also an economic decision for us. We try to balance privacy, legal compliance, and economic considerations. But we also know there are companies for them this data is a gold mine."

Access control. As a practice of data minimization, some experts mentioned access control which ensures training and using data by certain entities for certain applications strictly based on the requirement. E3 and E9 who have a role in banking and finance, particularly emphasized on authorization and access of data use as a part of their data minimization practice while also reporting that they don't have official direction based on GDPR for data minimization.

As E3 said, *"For data minimization: sensitive data obscure from external applications. We use a layered database approach to control access to sensitive information, so we have specific permissions to query the database, and even then, certain data like fiscal codes or IBANs (since I work in banking) are masked and inaccessible even when you search in a 'query-like feature". There are protocols for accessing older data, which require another level of authorization. Essentially, each employee's access to the database is tailored to their role, ensuring both data security and privacy compliance, not necessarily GDPR."*

6.2.2 Practicality & Feasibility Assessment

Experts agreed on the viability of the data minimization design concepts CD1, CD2, and CD4 provided by users, with ideas for their practical applicability and available resources for implementation. In contrast, they found CD3 and CD5 having numerous drawbacks, including both immediate and long-term security risks associated with their practical implementation. In this section we present the experts' assessment of the users' designs ideas.

Assessment of CD1: profile customization for data types and data volume. Most experts find the profile customization feature useful and practical to implement. E1, a software product manager, shared insights from a cloud product he worked on. He explained *"we ensure secure and efficient data storage. This involves clients' and their users' consent, especially in cloud computing. We deal with a variety of rental agreements. At times, we encounter users who prefer to keep their information private, due to the nature of our services spanning multiple countries and continents, and require a certain level of confidentiality for certain types of data. We have a consent-like form design for the product pipelined with db instances to cater to these various requirements efficiently, adding certain noise, adding gibberish for unique identifiers, and setting limits to the data life."* E1 also noted that such a feature is practical for search engines, allowing users to customize data types and volumes. E3, a specialist in banking IT security, stated, *"Search engines should be capable of this level of discretion. Theoretically, they could use algorithms to filter and avoid storing sensitive data, setting timelines for collecting specific data types. In banking, we strictly adhere to data collection timelines. We don't just follow GDPR for data minimization; it's our default system. Implementing this in the user interface shouldn't be a major challenge. The real question is their willingness to implement such measures. There's no technical excuse."* E4, a conversational search engine researcher, evaluated the design concept considering storage efficiency and performance. She commented, *"using data for longer durations is not the best way to present complex search results, scattered and skewed user queries in politics of many years is not reliable."*

Assessment of CD2: GDPR alerts for data minimization. For CD2, experts acknowledged its potential but expressed reservations about the feasibility of its implementation, especially given the challenges in quantifying the GDPR principle. E9 commented, *"This idea is great and would be something groundbreaking, but I don't see how as of now it is possible. To make Data minimization GDPR quantifiable against companies' Data policy, we need benchmarking which is not available to my knowledge. With ML, it will be possible to summarize the text content of these two sources, but I don't see actual data collection and retention being detected to provide GDPR alerts to users. will the company be willing to opt in for such data flow analysis? that's a question mark?."* Similarly, E3 and E7 reported similar feedback on the practicality of expecting companies to develop such UX in monitoring data, analyzing policies, and identifying key indicators for compliance.

Assessment of CD3: separating searching sessions for oneself and others. While users showed a strong preference for the design concept of having the option to conduct searches differently for themselves and others to minimize data, most experts considered this feature potentially more invasive. E7, a privacy and security expert, explained the risks *"Using two different search options for yourself and another person, like your child, essentially provides the search engine with annotated data to create a shadow profile. You're giving them a detailed footprint of behavior and possibly medical issues of your child under the 'others' search option. This creates a binary flag from your IP. This 'shadow' profile, built over lets say 20 years, could be misused. Unless it's guaranteed that the search engine doesn't collect any information from the 'others' search option, this design won't serve its intended purpose."* Similarly, experts E2, E5, and E10 acknowledged the technical feasibility of implementing this option but raised concerns about its effectiveness and privacy implications. E5 stated, *"Technically, it's possible to implement such a feature. However, I doubt the search results for the two options would*

be differentiated in the system for access permission and model training in the long run." E4 also commented, *"This might help unnecessary data collection. Specifying search queries about another person could allow re-identification by combining with other data sources, violating privacy of uninvolved parties. Keeping all searches together maintains privacy due to the noise. I wouldn't recommend this approach."*

Assessment of CD4: sensitivity-based search customization. Most experts supported the design concept of sensitivity-based search and confirmed that it's easy to implement with current technology and has the potential for incremental improvements over time. E8 explained, *"programming queries to identify keywords, analyze search frequencies and adapt to user behavior. While it may not yield perfect results immediately, it progressively improves through 'dynamic update mechanisms', 'intent recognition', and 'query parsing."'* Similarly, E2 emphasized the importance of distinguishing between personal and non-sensitive information. E2 outlined how this could be integrated into existing search engine frameworks: *"Current engineering or algorithm teams could anonymize or apply access control, adding noise after data collection at the query stage, triggering sensitive search alerts or options. This prevents data from reaching the server, with execution done on the client's search engine. It requires UX design and existing tech stacks in the backend to instantly discard sensitive searches made locally. It's akin to using Trusted Execution Environments (TEEs) for search queries. Cryptography could even be employed for private local execution. From a technical standpoint, this is very feasible."*

Assessment of CD5: data donations & incentives. Experts find the concepts of data donation and incentives intriguing, primarily in terms of enhancing data ownership and control, rather than data minimization. E9 noted, *"Platforms like Brave and Microsoft Edge offer users some agency and ownership of their data, providing incentives to create a sense of choice. However, this approach doesn't address the reduction of unnecessary metadata collection. It doesn't seem to alleviate that concern."* Similarly, E7, E3, and E10 echoed the sentiment that while the concept fosters a sense of control and ownership over data, it does not directly contribute to the goal of data minimization.

7 Discussion

Our study elucidated participants' perceptions of data minimization, the factors that shaped their decision-making processes, and end users' proposed conceptual designs assessed by experts for the feasibility for data minimization in search engines. Additionally, we discuss how the findings can facilitate system administrators in operationalizing data minimization capabilities.

7.1 Main Findings

Performance (service quality)-based interpretation to evaluate search results. Our study indicated a variety of metrics to evaluate the quality of search results, including exact keyword matches, no sponsored ads in search results, and results appearing on the first page, etc. Prior literature by Biega et al. has proposed performance-based interpretations of the data minimization principle, linking restrictions on data collection to quality metrics [14] where it is crucial to consider both global (average) and local (per user) minimum performance [14] while minimizing data collection that may not align with users' needs but rather those of companies. In the line with data minimization principle of GDPR which recommends retaining only relevant and necessary data [22, 71], our study contributes unique granular metrics for data minimization specific to users in the context of search engines which are pivotal for implementing user-driven data minimization.

Varying expectations, perceptions of necessity, and trade-offs for data minimization decisions. The principle of data minimization recommends that organizations should only collect and retain the minimum amount of personal data necessary for their stated purposes [30], but our findings indicate that users have more fine-grained expectations of how data minimization can be implemented in practice. Our research, when juxtaposed with earlier studies on privacy and the willingness to share data [59, 64, 67], reveals a nuanced, context-sensitive understanding among participants regarding the necessity of data minimization for different types of search queries. For example, participants recognized the value of location-based queries for good service, yet they were not comfortable with exact location coordinates. While users see value in medical search queries and consider sharing for better results in rare medical symptoms for other users if it is collected, they want more care put into ensuring this data is private. Furthermore, our study highlights a perceived trade-off by users between data minimization and free access to search services, viewing their personal data as a form of currency for the services received, echoing sentiments found in prior research [40, 51, 72]. Despite understanding that sharing data is a trade-off, users believe the granular quantity of data being collected, in addition to the time it is retained, can be reduced to better operationalize data minimization which is unique to previous privacy literature.

Incentives and data minimization. Our research highlighted a trend among users in response to advancing technologies to exchange personal data for incentives. This trend is exemplified by new data models in systems like Brave or Edge, which offer rewards for usage duration and opting in for advertisements. Although this resembles the "privacy paradox" [6], our study suggests that users view this trend as a means to gain ownership of their data, which they can then trade for improved services or rewards. Users attribute this inclination to the pervasive interconnectedness of modern technologies, especially AI. They perceive this choice as a way to have a degree of control over data by accepting these

incentives, as opposed to completely abstaining from online services for privacy which they deemed as impossible. On the other hand, providing incentives can act as an influence that makes users act in the company's best interests. Future research can further investigate user mental models of such concepts of data ownership and control.

The role of generative AI as a search engine. We found that through the proliferation of generative AI services such as ChatGPT and Gemini, some users are increasingly using these tools in lieu of traditional search engines. According to these participants, these generative AI services may become the next generation of search engines, but participants still held reservations about them, such as how reliable they truly were, and often using traditional search engines alongside AI for a more reliable search. Given the advancements in AI, data minimization strategies in a search engine context should be applied to these services to better protect user privacy across different services.

7.2 Design Implications

Drawing upon users' data minimization needs needs and experts' input regarding technical feasibility, we suggest actionable design implications for data minimization.

Data-minimized search mode with keyword sensitivity. Our findings show that users are skeptical of sharing data for sensitive queries, such as data relating to health, finances, politics, behavior, or data that is personally identifiable. Therefore, minimizing data collection for *sensitive queries* is an important design implication. One possible approach is to employ techniques based on keyword detection and natural language processing to identify sensitive keywords entered by users in the search bar. Second, the search engine can prompt the user to switch to a Sensitive Search Mode when a user inputs sensitive information. This would ensure no data is collected during the search to provide users with a customizable sensitive search experience. Another interactive design could involve triggering a just-in-time pop-up when typing sensitive queries in the search bar and providing an option to allow or disallow during a search.

Profiles customizable to minimize data based on type, quantity, and length of storage. Our results with users reveal that users have considerations about not only the sensitivity of data being collected, but also how much data, and how long data is retained for. To enable developers to implement a design that minimizes data collection and still offers personalized search results, we suggest implementing a profile configuration on the setting page or in the search page with a list of search data types and how long users are willing to share the search data. The scale could range from 0 (no data), 1 (only current) to data from last week month, or year. As user's preferences towards data minimization could change over time, this design can ask users to customize their preferences within certain time intervals. By providing users control over their data, search engines can address concerns about inadvertent data collection, and minimize data with users' intent and expectations.

Data donation and incentive structure. Users in our study mentioned that certain web browsers incentivized them to share data in exchange for a small reward. While experts view this design as somewhat unrelated to the core goal of data minimization, there are similar infrastructures within decentralized application ecosystems. In those settings, models from game theory and value exchange have been effectively utilized to design incentives for companies and users, facilitating a fair trade of data sharing in promoting user autonomy and self-sovereignty [9, 11]. The underlying concept is to allow a collaborative space where users and companies can negotiate, building trust and establishing a sense of ownership. This framework, though not directly focused on data minimization, presents an opportunity for search engines to offer users a choice in the degree of data minimization they prefer. Concurrently, it allows users to negotiate the value derived from sharing necessary and adequate data.

7.3 Limitations

Our interview study has several limitations. Firstly, we used Prolific as our recruitment platform, which resulted in over 50% of participants holding bachelor's or graduate degrees [34, 56]. This may restrict the generalizability of our findings, as individuals with higher education levels might exhibit stricter preferences and expectations towards data protection implementations compared to participants with other educational backgrounds. Secondly, the majority of participants primarily relied on Google for their daily search engine use. Consequently, their perceptions and expectations regarding data minimization are largely influenced by their experiences with the Google search engine, and we cannot claim that our findings generalize to other settings.

7.4 Future Work

Our results suggest that some users begin to replace traditional search engines with AI tools, such as Gemini and ChatGPT. Future work could explore users' privacy concerns with generative AI tools, how data minimization can be implemented in those contexts, or whether data minimization considerations for traditional search engines directly translate to generative AI usage.

8 Conclusion

Our study sheds light on users' perceptions, experiences, and needs regarding data minimization in search engines. Our findings surface different user needs for data minimization depending on the type of search engine, data type, and various contextual factors. Based on our interviews, we propose several expert-assessed conceptual designs for user-centered data minimization in search engines. Users prioritize transparency and interactive alerts to guide them in data minimization and

expect fine-grained data minimization controls in practice. Overall, this paper is the first to propose a user-driven interpretation of data minimization, highlighting the opportunity and need to involve end users as stakeholders in data protection implementations.

References

[1] Artificial intelligence and privacy: Report, 2018.

[2] What is a data controller or a data processor?, 2023.

[3] Accessed on 2023. https://gdpr-info.eu/art-5-gdpr/.

[4] Accessed on 2023. https://www.prolific.co/.

[5] Turkers in this canvassing: young, well-educated and frequent users, Accessed on 2023. https://www.pewresearch.org/internet/2016/07/11/turkers-in-this-canvassing-young-well-educated-and-frequent-users/

[6] Alessandro Acquisti. Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM conference on Electronic commerce*, pages 21–29, 2004.

[7] Alessandro Acquisti and Jens Grossklags. Privacy and rationality in individual decision making. *IEEE security & privacy*, 3(1):26–33, 2005.

[8] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. Intent-aware query obfuscation for privacy protection in personalized web search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 285–294, 2018.

[9] Darcy WE Allen, Chris Berg, Aaron M Lane, Trent MacDonald, and Jason Potts. The exchange theory of web3 governance. *Kyklos*, 2023.

[10] Beatriz Botero Arcila. Is it a platform? is it a search engine? it's chatgpt! the european liability regime for large language models. *J. Free Speech L.*, 3:455, 2023.

[11] Joost Bambacht and Johan Pouwelse. Web3: A decentralized societal infrastructure for identity, trust, money, and data. *arXiv preprint arXiv:2203.00398*, 2022.

[12] Susan B Barnes. A privacy paradox: Social networking in the united states. *First Monday*, 2006.

[13] Nik Bessis and Ciprian Dobre. *Big data and internet of things: a roadmap for smart environments*, volume 546. Springer, 2014.

[14] Asia J Biega, Peter Potash, Hal Daumé, Fernando Diaz, and Michèle Finck. Operationalizing the legal principle of data minimization for personalization. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 399–408, 2020.

[15] Virginia Braun and Victoria Clarke. *Thematic analysis.* American Psychological Association, 2012.

[16] Katriina Byström and Kalervo Järvelin. Task complexity affects information seeking and use. *Information processing & management*, 31(2):191–213, 1995.

[17] Jordi Castellà-Roca, Alexandre Viejo, and Jordi Herrera-Joancomartí. Preserving user's privacy in web search engines. *Computer Communications*, 32(13-14):1541–1551, 2009.

[18] Farah Chanchary and Sonia Chiasson. User perceptions of sharing, advertising, and tracking. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 53–67, 2015.

[19] Tamara Dinev and Paul Hart. An extended privacy calculus model for e-commerce transactions. *Information systems research*, 17(1):61–80, 2006.

[20] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 1388–1401, 2016.

[21] K Anders Ericsson. Protocol analysis. *A companion to cognitive science*, pages 425–432, 2017.

[22] European Commission. 2018 Reform of EU data protection rules, 2018.

[23] Jennifer Fereday and Eimear Muir-Cochrane. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods*, 5(1):80–92, 2006.

[24] Michèle Finck and Asia Biega. Reviving purpose limitation and data minimisation in personalisation, profiling and decision-making systems. *Technology and Regulation*, pages 21–04, 2021.

[25] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. john wiley & sons, 2013.

[26] Fabio Gasparetti. Modeling user interests from web browsing activities. *Data mining and knowledge discovery*, 31(2):502–547, 2017.

[27] Stephen Gilbert, Hugh Harvey, Tom Melvin, Erik Vollebregt, and Paul Wicks. Large language model ai chatbots require approval as medical devices. *Nature Medicine*, 29(10):2396–2398, 2023.

[28] Abigail Goldsteen, Gilad Ezov, Ron Shmelkin, Micha Moffie, and Ariel Farkash. Data minimization for gdpr compliance in machine learning models. *AI and Ethics*, pages 1–15, 2021.

[29] Google. Keeping your private information private, 2020.

[30] Seda Gürses, Carmela Troncoso, and Claudia Diaz. Engineering privacy by design. *Computers, Privacy & Data Protection*, 14(3):25, 2011.

[31] Seda Gürses, Carmela Troncoso, and Claudia Diaz. Engineering privacy by design reloaded. In *Amsterdam Privacy Conference*, volume 21, 2015.

[32] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. An empirical analysis of data deletion and {Opt-Out} choices on 150 websites. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 387–406, 2019.

[33] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[34] Panagiotis G Ipeirotis. Demographics of mechanical turk. 2010.

[35] Gjergji Kasneci, Christian Thomas Eberle, Martin Pawelczyk, and Tobias Leemann. I prefer not to say: Protecting user consent in models with optional personal data. 2022.

[36] Bert-Jaap Koops. The trouble with european data protection law. *International data privacy law*, 4(4):250–261, 2014.

[37] Anastasia Kozyreva, Philipp Lorenz-Spreen, Ralph Hertwig, Stephan Lewandowsky, and Stefan M Herzog. Public attitudes towards algorithmic personalization and use of personal data online: Evidence from germany, great britain, and the united states. *Humanities and Social Sciences Communications*, 8(1):1–11, 2021.

[38] Oksana Kulyk, Annika Hilt, Nina Gerber, and Melanie Volkamer. this website uses cookies": Users' perceptions and reactions to the cookie disclaimer. In *European Workshop on Usable Security (EuroUSEC)*, volume 4, 2018.

[39] Lin Kyi, Sushil Ammanaghatta Shivakumar, Franziska Roesner, Cristiana Santos, Frederike Zufall, and Asia Biega. Investigating deceptive design in gdpr's legitimate interest. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.

[40] Lin Kyi, Abraham Mhaidli, Cristiana Santos, Franziska Roesner, and Asia Biega. " it doesn't tell me anything about how my data is used": User perceptions of data collection purposes. *arXiv preprint arXiv:2312.07348*, 2023.

[41] Pedro Leon, Blase Ur, Richard Shay, Yang Wang, Rebecca Balebako, and Lorrie Cranor. Why johnny can't opt out: a usability evaluation of tools to limit online behavioral advertising. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 589–598, 2012.

[42] Han Li, Rathindra Sarathy, and Heng Xu. Understanding situational online information disclosure as a privacy calculus. *Journal of Computer Information Systems*, 51(1):62–71, 2010.

[43] Yuelin Li and Nicholas J Belkin. A faceted approach to conceptualizing tasks in information seeking. *Information processing & management*, 44(6):1822–1837, 2008.

[44] Mark MacCarthy. In defense of big data analytics. *The Cambridge Handbook of Consumer Privacy*, pages 47–78, 2018.

[45] Dominique Machuletz and Rainer Böhme. Multiple purposes, multiple problems: A user study of consent dialogs after gdpr. *Proceedings on Privacy Enhancing Technologies*, 2020:481–498, 04 2020.

[46] Falk Maoro, Benjamin Vehmeyer, and Michaela Geierhos. Leveraging semantic search and llms for domain-adaptive information retrieval. In *International Conference on Information and Software Technologies*, pages 148–159. Springer, 2023.

[47] Arvind Narayanan and Vitaly Shmatikov. Robust deanonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[48] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. Dark patterns after the gdpr: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[49] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev.*, 57:1701, 2009.

[50] Saurabh Panjwani, Nisheeth Shrivastava, Saurabh Shukla, and Sharad Jaiswal. Understanding the privacy-personalization dilemma for web search: A user perspective. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3427–3430, 2013.

[51] Panagiotis Papadopoulos, Nicolas Kourtellis, Pablo Rodriguez Rodriguez, and Nikolaos Laoutaris. If you are not paying for it, you are the product: How much do advertisers pay to reach you? In *Proceedings of the 2017 Internet Measurement Conference*, pages 142–156, 2017.

[52] Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pages 117–126, 2017.

[53] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten de Rijke. Conversations with search engines: Serp-based conversational response generation. *arXiv preprint arXiv:2004.14162*, 2020.

[54] Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. Leading conversational search by suggesting useful questions. In *Proceedings of the web conference 2020*, pages 1160–1170, 2020.

[55] Awanthika Senarath and Nalin Asanka Gamagedara Arachchilage. A data minimization model for embedding privacy into software systems. *Computers & Security*, 87:101605, 2019.

[56] Eunjin Seong and Seungjun Kim. Designing a crowdsourcing system for the elderly: a gamified approach to speech collection. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2020.

[57] Divya Shanmugam, Fernando Diaz, Samira Shabanian, Michèle Finck, and Asia Biega. Learning to limit data collection via scaling laws: A computational interpretation for the legal principle of data minimization. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 839–849, 2022.

[58] Tanusree Sharma. Rethinking data minimization from a {User-Centered} approach: A paradigm shift. 2023.

[59] Tanusree Sharma, Smirity Kaushik, Yaman Yu, Syed Ishtiaque Ahmed, and Yang Wang. User perceptions and experiences of targeted ads on social media platforms: Learning from bangladesh and india. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.

[60] Xuehua Shen, Bin Tan, and ChengXiang Zhai. Privacy protection in personalized search. In *ACM SIGIR Forum*, volume 41, pages 4–17. ACM New York, NY, USA, 2007.

[61] Jaime Teevan, Susan T Dumais, and Eric Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 449–456, 2005.

[62] Omer Tene and Jules Polonetsky. Privacy in the age of big data: a time for big decisions. *Stan. L. Rev. Online*, 64:63, 2011.

[63] Elaine G Toms. Task-based information searching and retrieval., 2011.

[64] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *proceedings of the eighth symposium on usable privacy and security*, pages 1–15, 2012.

[65] Pertti Vakkari. Task-based information searching. *Annual Review of Information Science and Technology (ARIST)*, 37:413–64, 2003.

[66] Tijs van den Broek and Anne Fleur van Veenstra. Governance of big data collaborations: How to balance regulatory compliance and disruptive innovation. *Technological Forecasting and Social Change*, 129:330–338, 2018.

[67] Richard Whiddett, Inga Hunter, Judith Engelbrecht, and Jocelyn Handy. Patients' attitudes towards sharing their health information. *International journal of medical informatics*, 75(7):530–541, 2006.

[68] Wan-Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. Grannies, tanning beds, tattoos and nascar: Evaluation of search tasks with varying levels of cognitive complexity. In *Proceedings of the 4th information interaction in context symposium*, pages 254–257, 2012.

[69] Heng Xu, Hock-Hai Teo, Bernard CY Tan, and Ritu Agarwal. The role of push-pull technology in privacy calculus: the case of location-based services. *Journal of management information systems*, 26(3):135–174, 2009.

[70] Zhonghao Yu, Sam Macbeth, Konark Modi, and Josep M Pujol. Tracking the trackers. In *Proceedings of the 25th International Conference on World Wide Web*, pages 121–132, 2016.

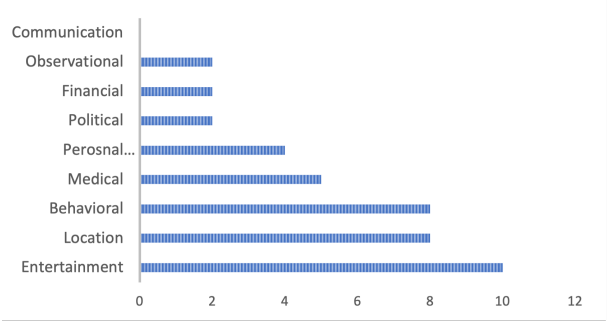

Figure 7: Types of search queries participants would allow the search engine to retain to obtain better search results.

[71] Tal Z Zarsky. Incompatible: The gdpr in the age of big data. *Seton Hall L. Rev.*, 47:995, 2016.

[72] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. User perceptions of smart home iot privacy. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–20, 2018.

[73] Steven Zimmerman, Alistair Thorpe, Chris Fox, and Udo Kruschwitz. Investigating the interplay between searchers' privacy concerns and their search behavior. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 953–956, 2019.

[74] Shoshana Zuboff. *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books, 2019.

A Appendix

A.1 Additional Figures & Tables

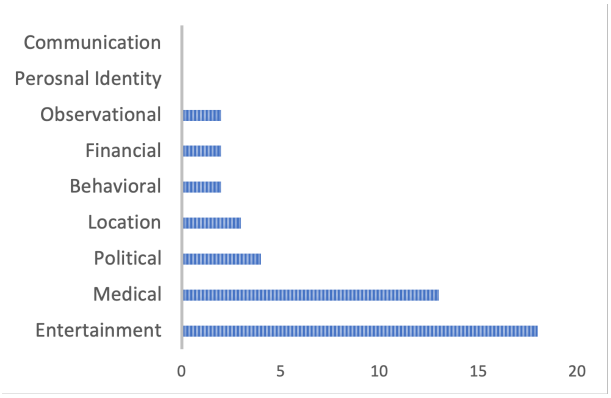

Figure 8: Types of search queries participants would allow the search engine to retain so that other users can obtain better results.

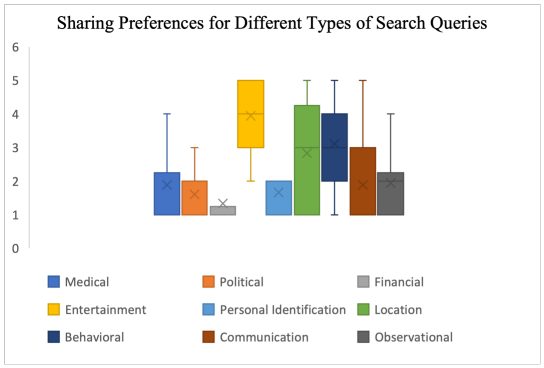

Figure 9: Ranking of participants sharing preferences towards types of search queries where 1: less likely to share, 5: more likely to share.

Participant ID	Sex	Age	Country of residence	Highest education level completed	Employment status	Weekly device usage
P1	Female	32	United Kingdom	Graduate degree (MA/MSc/MPhil/other)	Full-time	Multiple times every day
P2	Male	28	Germany	Undergraduate degree (BA/BSc/other)	Full time	Multiple times every day
P3	Male	40	Portugal	Undergraduate degree (BA/BSc/other)	Full time	Multiple times every day
P4	Male	22	Netherlands	Undergraduate degree (BA/BSc/other)	Part-Time	Multiple times every day
P5	Male	43	Italy	Undergraduate degree (BA/BSc/other)	Full time	Everyday
P6	Female	29	Germany	Undergraduate degree (BA/BSc/other)	Full time	Multiple times every day
P7	Female	38	Netherlands	Undergraduate degree (BA/BSc/other)	Unemployed (and job seeking)	Multiple times every day
P8	Male	51	Estonia	High school diploma/A-levels	Part-Time	Multiple times every day
P9	Male	40	Portugal	Undergraduate degree (BA/BSc/other)	Full time	Multiple times every day
P10	Male	34	Spain	Graduate degree (MA/MSc/MPhil/other)	Part-Time	every day
P11	Female	36	Portugal	Graduate degree (MA/MSc/MPhil/other)	Full time	Multiple times every day
P12	Female	54	United Kingdom	High school diploma/A-levels	Not in paid work (e.g. homemaker')	2-6 times a week
P13	Female	62	United Kingdom	Undergraduate degree (BA/BSc/other)	Full time	Everyday
P14	Male	34	Poland	High school diploma/A-levels	Full time	Multiple times every day
P15	Male	37	Portugal	Undergraduate degree (BA/BSc/other)	Full time Everyday	
P16	Female	24	United Kingdom	Undergraduate degree (BA/BSc/other)	Part-Time	Multiple times every day
P17	Female	47	United Kingdom	Undergraduate degree (BA/BSc/other)	Full time	Multiple times every day
P18	Female	49	Spain	High school diploma/A-levels	Part-Time	Everyday
P19	Male	40	United Kingdom	Undergraduate degree (BA/BSc/other)	Full time	Multiple times every day
P20	Female	40	France	Technical/community college	Full time	Multiple times every day
P21	Male	52	Ireland	High school diploma/A-levels	Full time	Multiple times every day
P22	Male	32	France	Undergraduate degree (BA/BSc/other)	Full time	Multiple times every day
P23	Female	26	United Kingdom	Technical/community college	Full time	Multiple times every day
P24	Female	50	Poland	Undergraduate degree (BA/BSc/other)	Not in paid work (e.g. homemaker')	Everyday
P25	Male	52	Spain	High school diploma/A-levels	Part-Time	Everyday

Table 1: Participant demographics and background.

Expert ID	Sex	Education	Year of Exp	Role
E1	Male	Masters CS	4 year	Software Product Manager (IT)
E2	Female	Masters IT	17 years	Database Security (Fund-Raising)
E3	Male	Masters CS	15years	IT Security (Banking)
E4	Female	PhD Researcher	5 years	Conversational SE (Health)
E5	Male	Bachelors	10 years	Software engineer (Employee Tool)
E6	Male	Bachelors	8 years	Software engineer II
E7	Female	PhD Researcher	6 years	Data Minimization (Law)
E8	Male	Masters	3 years	Software engineer (Search Engine)
E9	Male	Bachelors	2 years	Software engineer
E10	Male	PhD Researcher	4 years	Search Engine Algorithm

Table 2: Experts' demographics and background.