Minimalist approaches to enforce privacy by design in surveys

Enka Blanchard
Université Polytechnique Hauts-de-France,
CNRS, UMR 8201 - LAMIH, F-59313
Valenciennes, France
Centre Internet et Societé, UPR CNRS 2000
enka.blanchard@cnrs.fr

L. Gabasova
Independent scholar

ABSTRACT

Public institutions and private companies both frequently rely on user surveys for a variety of assessments (e.g. equality issues or quality of work environment). However, many such surveys struggle to garner sufficient responses, especially when they ask about sensitive subjects (such as work harassment), which also makes them exist in a legal grey area when it comes to data protection laws. One important factor in this issue is the perceived threat of deanonymisation, compounded by the frequent lack of transparency on how the data is used. The proposals seeking to address this issue often focus on complex cryptography (e.g. homomorphic encryption), without addressing the fears of non-technical users.

This paper explores a radically different approach which minimises data collection on multiple fronts, partially by limiting the power of survey organisers. By design, it prevents generic attempts to deanonymise participants, as the server never stores even pseudonymised information. We also try to address questions of inclusivity, once again through a minimalist approach. Finally, we report on the first live test of a prototype developed following this approach.

Author Keywords

Privacy by design, Survey, Methodology, Anonymisation, User experience

CCS Concepts

•**Security and privacy** → **Social aspects of security and privacy; Usability in security and privacy; •Theory of computation** → *Theory of database privacy and security;*

•**Social and professional topics** → Privacy policies;

DOI: 10.37789/rochi.2023.1.1.15

INTRODUCTION

Workplace surveys are seeing increased use in both public institutions and private companies, to get employee feedback on indicators ranging from the quality of the work environment to diversity or harassment issues [22]. Those surveys suffer from multiple issues, especially low and further declining response rates (unless the survey is mandatory) and the risk of self-censorship [7]. This is particularly true when there is a risk that participants could be deanonymised, which has led to a lot of work in the field of differential privacy [8].

Due to the specificities of workplace surveys and the varying and sometimes conflicting regulations, some of the questions asked can also be in a legal grey area (despite the rise of legal frameworks such as GDPR). For example, although medical data is subject to strict confidentiality rules, some institutions approve the use of questions about discrimination which ask the reason for such discrimination[1] (with the option to choose medical reasons or disability status). Certain questions can also create some friction depending on how they're phrased. For example, asking the participant's gender is generally seen as mandatory, but runs into multiple issues on questions regarding inclusivity and politicisation [24].

An additional important reason to be mindful with the phrasing and inclusion of specific questions is that, even if the immediate impact of a given survey is limited, its framework may be reused in the future, especially in public institutions. This means that any badly phrased elements will be carried forth into an indeterminate number of future surveys, become standardised and have long-lasting consequences [24]. This situation can then be hard to correct, as simply removing a badly phrased question from a survey can lead to knock-on changes due (for instance) to priming. This echoes ethical issues with datasets with personal information used in computer science, which can continue being used despite major flaws being found [16].

This paper stems from an experiment done in 2022. The authors were tasked by a French university with designing and developing a survey system that had stronger privacy guarantees than off-the-shelf systems[2]. This was an opportunity to follow privacy-by-design minimalist approaches and best

[1] Some of the surveys we were given as examples of what was done previously were operated using LimeSurvey through RENATER (the French National Research Network). The RENATER terms of use expressly forbid questions on health and sexuality [21], rendering uncertain the status of questions on harassment for cause of sexual orientation or disability — which feature in some surveys.

[2] The system was developed to be released as free software.

practices from both usable security and social sciences [23, 17]. We propose the following contributions:

- reflections on minimalist approaches in survey system design;
- one proposed design of an anonymous survey system with potential extensions;
- summary feedback from a live test of a prototype of said system.

This article is structured as follows. We start by general questions on survey design and how this interacts with questions of transparency, data handling, and privacy. We then look at how minimalist approaches can help propose an alternative to existing survey systems by focusing on anonymity and the various components and questions that must be addressed when doing so. We then move on to questions beyond pure anonymity such as usability and inclusion, as they can have negative interactions with the former. We conclude with summary feedback from the first live test of the prototype and future work.

SETTING GOALS

The first task of any survey planner is to decide what the exact goals of the survey are, impacting the design constraints and hence the question of which survey systems are adequate. This is not a trivial task, for multiple reasons. First, there can be competing interests in the outcome of the survey. For example, in practice, there can be a difference between the announced goal of surveys on workplace harassment — finding it if it exists — and the real goal: showing that it is rare or non-existent. This means that some potential objectives such as transparency (e.g., promising to make the results entirely public) can be resisted by stakeholders wanting to avoid bad publicity.

Second, many workplace surveys are not designed in a fashion similar to surveys intended for purely research purposes, which start from a research question and seek to answer it. Instead, they can come from many different considerations: legal requirements or corporate incentives to enquire about mental health, desire for data/feedback to improve the company's practices to help employee retention, or even the public relations bonus coming from acting on a perceived issue... The people designing the survey can also be entirely removed from the initial decision to have a survey, all of which can increase the confusion as to the objectives.

We observed one one notable consequence of this in past surveys: many of them were designed in an *ad hoc* fashion. This involved adding questions that seemed interesting without necessarily looking at interactions between them, priming effects — or even whether any given question contributes information that cannot be inferred from other questions. This is crucial when considering the main two objectives of such surveys: either getting a description of how things are (with both quantitative aspects and open-ended qualitative approaches), or finding correlations and anomalies between groups of participants. As statistical analysis may not always be possible due to small sample size or low response rate, questions that rely on returning a quantitative measurement can be useless to the survey's goals. The presence of such questions goes against good usability practices, as more questions translate directly to higher user cost and therefore a higher dropout rate.

This *ad hoc* approach to survey creation often falls under the paradigm and characteristic patterns of big data and data mining. It silently assumes that data speaks for itself and that, by collecting as much data as can possibly be collected, one can sift through it for meaning and correlations [12, 18]. This already becomes problematic at the data collection stage due to the above-mentioned user cost. Then, with the wealth of data, the researcher — or survey organiser — is free to test all the possible correlations until they find interesting ones, without necessarily following good statistical practices (such as Holm-Bonferroni methods [1]). This encourages bad practices such as p-hacking, which may not even need to be carried out consciously [11], and can be compounded by organisers whose expertise lies in psychology and not statistics, as some bad statistical practices have been considered standard by professional organisations [5]. This is not just bad scientific practice, but can also have an impact on the results. Any survey with sufficiently many questions (10 is enough in practice) and respondents will statistically find differences between groups of people (e.g., following gender or age), encouraging the organisers to start policies to address effects that can be pure statistical noise.

To address this, one option is to adopt a fully transparent approach and, following social science methods, pre-register the analyses [19]. This means that the methods, stated goals and planned observations and correlations should be recorded before it is conducted, and ideally made available to participants (e.g., on the survey's welcome page).

The following sections will develop a design methodology and a proposed system that seeks to address these issues by enforcing such transparency and pre-registration following a privacy-by-design approach.

MINIMALIST APPROACHES AND ANONYMITY

In almost all cases, the survey's organisers' goals include having both a high response rate and good-quality answers with minimal self-censorship. This depends on many aspects, including trust — both in the organisers and the survey system. Although surveys should always leave open the possibility of declining to answer, and some users can choose to answer inaccurately, both options reduce the information that can be gathered and do not give strong privacy guarantees. Indeed, one of the main risks when handling workplace surveys concerns some of the sensitive data being attributable to specific individuals, especially if the data gets leaked. A survey organiser with access to full answer sheets for each participant can find it trivial to identify the single person who answered a certain way, which increases the barrier to reporting sensitive information. Even when people do not have access to full answer sheets, it can still be possible to deanonymise participants by looking at correlation chains. For example, let's suppose we have a single participant reporting harassment[3]. If one has access to the average age and gender of people

[3] We have observed surveys which allow one to report harassment while also giving demographics information. Although there should

reporting harassment, one has initial elements which, if that person is the only one in their age-gender category, allow one to get increasingly more information and to eventually build a profile.

Before detailing our proposed system, we should give a quick warning about attack frameworks. A first constraint is that, if the full set of survey organisers have bad intentions, no design system can address the issue. Indeed, the organisers could simply lie as to which system is used — especially if they have full control of the servers. Similarly, any person with physical access to the server has a high chance of being able to obtain the data — unless everything is secured through trusted platform modules with correct cryptography, and even this supposes resistance to side-channel attacks which is not guaranteed [20, 13]. Thus, a reasonable attacker's profile in such a context is someone — potentially a manager or an employee from human resources — trying to access data about their colleagues, hence with reasonably limited technical ability.

The central idea behind the system proposed is to fully **get rid of answer sheets**. If correctly implemented, this boosts anonymity and has multiple positive consequences.

First, we can observe that any system that does not rely on answer sheets needs to store data in separate non-linked bases (or equivalently, in a single base with no relationships between columns). If done naively, this only allows some descriptive statistics — getting the proportion of people unhappy about a particular element — and qualitative feedback. However, it does not allow the study of any between-groups differences — such as whether one gender has different work experiences. While one can compute arbitrary correlations when one has full answer sheets, a minimalist decorrelated approach makes it impossible.

This means that one requires two types of columns: those containing raw data (such as the answers to the question "are you in a management position") and those containing more complex information. For example, if one wants to study the correlation between being a manager and being overworked, there can be the two raw data columns, plus one additional column corresponding to the correlation between the two: a list of data pairs, each of which can be computed on the participant's side. From a technical perspective, the system can be implemented by filling an answer sheet on the user's side and, upon submission of the filled answer sheet, computing all the correlations on the user's side and sending each as a separate update of one of the database's columns. Correlations between more than 2 variables can also be recorded but each additional variable makes deanonymisation easier. Even with only 2-variable correlations, care should also be taken to avoid correlation chains when designing the questions — while keeping the context in mind and how a single question could deanonymise certain persons if the sample set is small enough.

As any correlation must be planned in advance, this approach forces a reflection on the part of survey designers, who cannot simply compile a list of questions and then explore the data for links. The minimalist approach requires starting with a clear set of objectives and constraints. It also means that, once an initial set of questions is considered, a necessary step is to compute whether — if response rates are reasonable — these questions will give actionable information. It also makes transparency less costly (as the costs are paid upfront). Indeed, to facilitate trust, the organisers can choose to publish in advance the full list of questions and correlations — and optionally even open a consultation on whether some questions should be added/removed/reworded. This comes at no cost to the organisers as it does not hinder their future analyses any more than the system already does.

This is also a first step to ensure privacy-by-default as it strongly restricts the type of information that can be obtained from any single answer. Fully preventing correlation chains is generally not possible from within the system, as doing so requires contextual information, such as the number of people from a certain demographic who occupy a specific position. It can sometimes be possible to ensure that a correlation chain is impossible no matter the context, but this requires larger survey populations — and is the context where differential privacy is often explored [8, 9]. In the case of small survey populations (i.e., around 100), we thankfully have multiple ways to address deanonymisation shown below, still following a minimalist approach.

Nominative information

The easiest way to deanonymise is to obtain nominative data, such as the participant's email or IP. This is not always stored as part of the answer sheet, but a unique identifier or password is commonly sent to users to prevent spamming and limit answers to one per person. Naturally, any privacy-oriented system should prevent the storage of directly identifiable data (or its access by the organisers when storing it is unavoidable). If the password is sent by email, it creates the opportunity for organisers to directly attribute answers to known email addresses, and participants can then have legitimate privacy concerns as they cannot know whether organisers are able to track their answers.

One way to address this is to use simple passwords or passphrases — for example, two common words — and to tell users that they are free to exchange them with colleagues (although each code can only be used once). Moreover, those passwords should be checked once when the survey data is submitted (or when access is granted), but should not be stored as part of the same database to avoid it being correlated with any other information. Another option (especially with emails if the potential list is of sufficient size) is to compute an expensive hash (e.g., with Argon 2 [3]) and check if the hash is already present in the database, with the hash being made costly enough to dissuade brute-force attacks.

Some other elements can also contain nominative information. For example, open-ended questions (such as requests for "potential improvements to the workspace") allow participants to refer to arbitrary information and can make it very easy to

generally be a way to anonymously report harassment, observing victimisation rates in systematic surveys can also play a role.

identify them. As correlating such answers with any others seldom gives usable information, the natural option is to prevent any correlation with such open-ended questions. It is also worth stating that this is the case in the survey's welcome page to encourage participants to give detailed answers.

Decorrelating answers

Once the nominative data is removed, the next step to avoid deanonymisation is not just to avoid storing full answer sheets, but to make sure that such answer sheets cannot be obtained from the database even with full (*a posteriori*) access. Thus, not only should each question be stored separately, the n-th line of each column should not just consist of the n-th participant's answers. Thus, each time data is stored for a question (that isn't a counter), the column should be reordered randomly. Instead of reordering the whole array, it is thankfully enough to only permute the last element added with another (including itself) uniformly, akin to a reversed Fisher-Yates shuffle [4].

Avoiding partial results

There is one way to deanonymise participants even if the system uses the previous elements. By observing the results at multiple points in time (ideally between each participant), it becomes possible to infer the full answer sheets. A way to prevent this is to only make the results available once the survey is finished.

If, as above, we don't consider attackers with physical access to the server, a solution is to host the survey externally, or at least on a server administered by someone with no direct links to the participants — as external hosting is not always legally allowed for sensitive data. We can then differentiate between the person with complete server access — who has arbitrary power over the survey but no motive[4] — and the organisers. The latter should only be authorised to input the questions as well as the list of participants' emails (to access the survey), end the survey, and publish or download the results.

Post-survey correlation chain elimination and question twinning

If one is given preliminary contextual data (such as demographic information), it can become possible to perform additional correlation checks at the end of the survey. For example, let's suppose a sensitive question is correlated to a few demographic categories including gender and age, and let's suppose it is known that only three 60 year-old men work for the company and no 60-year old women do. If no women answer yes to the question but the three men do, they can be deanonymised (thanks to the contextual knowledge). The correlation by itself is not necessarily at fault as the uncertainty would remain if one of them did not answer (or answered differently).

These considerations should come into account when designing the data requested. For example, instead of asking for age, having age brackets tailored to the expected data limits the data accuracy but also prevents many forms of deanonymisation.

Another potential method is to automatically detect potential deanonymisation cases, either by feeding the system some contextual information initially or by guaranteeing that everyone in the population participated in the survey, giving it total demographic information.

In any case, simply removing the deanonymising question from the survey is not always a solution. Indeed, as sometimes only one answer set can be deanonymising, removing the question can be just as deanonymising as leaving it. One option is then to analyse *a priori* which questions could lead to such cases and to twin them: if at least one of them is removed, then the other also is. This eliminates the risk by creating an ambiguity, although at the expense of additional data loss.

USABILITY AND INCLUSION

Enforcing the anonymity of participants using the methods discussed above creates both new opportunities and new hurdles in terms of usability. Some common constraints or improvements to generic surveys can either be made impossible by this or can negate the design's advantages. This section will then go over three such questions: participant's self-identification, cookie-handling and data modification or deletion.

Self-identification

As social norms evolve, certain demographics questions can become both more complex and more politically loaded and can create some friction depending on which categories are available. For example, asking for the participants' gender/sex runs into multiple issues, beyond the simple possibility of the participant refusing to answer:

- If only two options (e.g., "man" and "woman") are available, this excludes some participants (e.g., the ones who are queer, intersex, non-binary, etc.) [24].
- If more than two options are available, their very presence can make it a political issue and cause strife as some participants can react negatively to these options being presented[5]. Moreover, this does not solve the question of which options to show (as just adding a field marked "other" is — literally — othering [2]).
- Both of these alternatives can create legal issues as increasingly many jurisdictions implement legal recognition of more than two gender options [6]. Some policies can also come into potential conflict, such as local or national policies taking one stance (e.g., mandating or forbidding having 3 options) while supranational or federal rules mandate the opposite in evolving political landscapes[14].

One potential solution to this issue is to leave an open field for all participants (and not just for those who'd choose "other"). This allows for self-identification without pushing the inclusivity to the forefront (and thus limiting the political strife it creates).

[4] Of course, it might be possible to bribe or coerce the administrator. However, this means that the bribing party exposes themself if the administrator reveals the attempt. Keeping the administrator's identity private to most would also limit the exposure — and reduce the set of potential guilty parties if a bribe is attempted.

[5] Although offering multiple options can be a step for inclusivity, it can also in turn create risks for the affected people as it can lead to adverse reactions, up to harassment in the workplace of both participants and survey organisers [15]

This method has two main drawbacks. First, it reduces usability by converting an “easy” question with a radio button into an open text field. This increases the chance of users not understanding the question, declining to answer it or accidentally skipping it.

A second issue is that it complexifies the analysis by offering a number of categories unknown in advance. Thus, it requires finding a way to handle the correlations when one has an open field to correlate.

If one keeps the correlation with the open text field, it can allow anyone with the full dataset to reconstruct some answer sheets. These can easily help identify someone if they’re the only non-binary employee, but also for anyone else with a unique answer (up to white-space) as it allows multi-variable correlations. It keeps the highest level of detail at the cost of privacy, and can make future analyses more complex (depending on how the data is eventually clustered into a reduced number of categories[6]).

Another option is to parse the data immediately on the client’s side(with an extensive but non-exhaustive initial list) into a few categories, for example “Woman”, “Man”, “Other” and “Did not respond” (the latter two can be combined). This is somewhat more inclusive than just having an “Other” option within the survey (as participants aren’t directly facing it) and facilitates the correlation analyses. However, it adds noise depending on the parser’s accuracy, as some participants will be put into the “Other” category despite indicating clearly (albeit with a rare formulation) that they belong in the first two. This noise can thankfully have one positive impact, as it helps improve the privacy of the people who give less common answers (who can also be clustered with those who refuse to answer).

Cookies

Due to the survey not saving full user sheets, any problem in the database could corrupt the data in a way that cannot be handled by simply simulating the inputs on the server side. More importantly, if a correlation that was meant to be measured failed, it is impossible to get it back from the available data — unlike with the standard database structure where the organisers can choose what to analyse *a posteriori*.

One of the solutions is to ask the participants to retake the whole survey, but that has a high user cost and compounds with the dropout risk. Another option is to store all the participants’ data on the client’s side (as a full sheet in a cookie). In case of a server-side issue, it then becomes possible to update the code then ask users to go back to the survey page and resubmit their original data — plus eventual new correlations computed on the client’s side.

This does create some security and privacy risk depending on the exact context. It can also have a small usability cost as the cookie storage requires compliance with various regulations such as GDPR[7] [25]. The cookie information being available in cleartext on the client machine is also a non-negligible privacy risk, which can be mitigated by encrypting the cookie data. This can be done either in an asymmetric way — in which case the server is asked to decrypt the data upon a second login — or using symmetric encryption without storing the key on the client’s machine except during the session. If there is a risk of a participant stealing another participant’s cookies, the password should be different for all users (and could be partially based on the user’s password).

Data modification and deletion

Another issue with avoiding user sheets is that it becomes not directly possible to remove or modify one user’s data. However, if one is using cookies as above, then an option becomes available — with an additional security risk if anyone has full access to the server, although as stated above this renders most points moot. In addition to the decorrelated user data, the server can store a hash[8] of each sheet (including the password). Then if a user tries to login back into the system, they can go into a special modification mode where the client keeps the old cookie with the old answers, and sends one message to the server with the initial list of answers, the corresponding hash (which the server checks before deleting each answer from the corresponding database), and the corrected list of answers. The list of hashes should in any case not be public and should be deleted when the final results are computed to prevent future bruteforce attempts.

CONCLUSION

This paper proposes a minimalist approach to workplace surveys that focuses on anonymity by following a privacy-by-default approach. This is not just a theoretical contribution, as the ideas above were used in a prototype system that was tested in a French university in 2022 with plans to eventually make it open-source after further rounds of testing. Some proposals — such as question twinning and data modification through hashes — were left out because they seemed too complex for the task at hand.

Although we cannot give details on the test that was performed, we can still report some feedback. First, the response rate was substantially higher than similar surveys in similar contexts we are aware of, although this can come from a variety of causes, including the extended discussions about the system within the community and the frequent reminders to participate. Some of the discussions were focused on privacy (as it was a “selling point” of the system), which then became more of a public issue, with multiple users contacting the organisers to express doubt and mistrust in the system. Finally, the open text field for gender did lead to non-negligible noise (more than 10%), partially because of a much higher non-response rate compared to other demographics question, and partially because the range of answers went far beyond the parser’s planned list.

[6] If the data is not meant to be clustered at all, then one can question whether it should be correlated at all.

[7] One option is to have at the end of the survey a checkbox with the option to store the data locally if the user wants to, in which case a GDPR cookie warning on the front page wouldn’t be necessary.

[8] It would be preferable to use a costly hash to resist bruteforce (e.g., Argon2 [3]), especially if the survey is short.

We also found multiple questions, social and technical, for which we have no good answers and which could be investigated in the future. Here are the main ones:

- To what extent would pre-registering the correlations and letting the participants see those in advance affect their responses?
- As transparency is good for trust, we advocate for publishing in advance the question and correlation list. However, the question remains of whether to automatically make public (to the participants) the results of the study once it is completed, and the corresponding trade-offs deserve an analysis.
- The approaches shown above can be made compliant with legislation such as GDPR (and were made so in the case of the deployed prototype). Indeed, as the data truly is anonymised and not pseudonymised, it follows different regulatory constraints (especially regarding the "right to be forgotten"). However, doing so is non-trivial: it requires fully separate data handling for the email part (which is not anonymised, even if it is only used to send an initial invitation to the survey) and for the survey's responses. As such, it might be too costly for administrators and organisers, and more work is needed on the cost-benefit analysis, including for the participants who might prefer having the possibility of deleting their data to just having stronger privacy guarantees that no-one can know that it is their data.
- Our model assumes that an attacker does not have direct access to the server (as it is generally beyond the technical purview of the survey organisers to prevent such attacks). For situations where higher security is required, this vulnerability needs to be addressed. A simple method is to have a double system where two machines in different locations are in continuous contact with each other and the Internet. If contact is broken at any point or if someone tries to access one of the machines physically, a public alert is sent (for example by email, on Twitter or a blockchain). This is already a better system, but it is extremely prone to false alarms (and to denial-of-service attacks). Using systems such as proactive secret sharing [10] as well as TPMs, could a distributed encrypted system avoid this issue while being able to recover when one machine fails?

REFERENCES

[1] Hervé Abdi. 2010. *Holm's Sequential Bonferroni Procedure*. SAGE Publications.

[2] Greta Bauer. 2012. Making Sure Everyone Counts: Considerations for Inclusion, Identification and Analysis of Transgender and Transsexual Participants in Health Surveys. (2012). `https://open.library.ubc.ca/cIRcle/collections/facultyresearchandpublications/52383/items/1.0132676`

[3] Alex Biryukov, Daniel Dinu, and Dmitry Khovratovich. 2016. Argon2: new generation of memory-hard functions for password hashing and other applications. In *IEEE European Symposium on Security and Privacy – EuroS&P*. IEEE, 292–302.

[4] Paul E. Black. 2019. Fisher-Yates shuffle. *Dictionary of Algorithms and Data Structures* 19 (2019).

[5] Denny Borsboom. 2006. The attack of the psychometricians. *Psychometrika* 71, 3 (2006), 425.

[6] Dylan Amy Davis. 2017. The normativity of recognition: Non-binary gender markers in Australian law and policy. In *Gender panic, gender policy*. Vol. 24. Emerald Publishing Limited, 227–250.

[7] Don A. Dillman. 2020. *Towards Survey Response Rate Theories That No Longer Pass Each Other Like Strangers in the Night*. Springer International Publishing, Cham, 15–44. DOI: `http://dx.doi.org/10.1007/978-3-030-47256-6_2`

[8] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.

[9] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. 2019. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2468–2479.

[10] Yair Frankel, Peter Gemmell, Philip D MacKenzie, and Moti Yung. 1997. Optimal-resilience proactive public-key cryptosystems. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*. IEEE, 384–393.

[11] Andrew Gelman and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University* 348 (2013).

[12] Nitza Geri and Yariv Geri. 2011. The Information Age Measurement Paradox: Collecting Too Much Data. *Informing Sci. Int. J. an Emerg. Transdiscipl.* 14 (2011), 47–59.

[13] Dan Goodin. 2021. Trusted platform module security defeated in 30 minutes, no soldering required. Ars Technica. (2021). `http://web.archive.org/web/20220523092016/https://arstechnica.com/gadgets/2021/08/how-to-go-from-stolen-pc-to-network-intrusion-in-30-minutes/`

[14] Haute Autorité de Santé. 2020. *Sexe, genre et santé*. Technical Report. Haute Autorité de Santé. `https://www.has-sante.fr/upload/docs/application/pdf/2020-12/rapport_analyse_prospective_2020.pdf`

[15] S. Jaroszewski, D. Lottridge, O. L. Haimson, and K. Quehl. 2018. "Genderfluid" or "Attack Helicopter": Responsible HCI Practice with Non-Binary Gender Variation in Online Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. DOI: `http://dx.doi.org/10.1145/3173574.3173881`

[16] Os Keyes and Jeanie Austin. 2022. Feeling fixes: Mess and emotion in algorithmic audits. *Big Data & Society* 9, 2 (2022), 20539517221113772.

[17] Marc Langheinrich. 2001. Privacy by design—principles of privacy-aware ubiquitous systems. In *International conference on ubiquitous computing*. Springer, 273–291.

[18] Alec Levenson and Alexis Fink. 2017. Human capital analytics: too much data and analysis, not enough models and business insights. *Journal of Organizational Effectiveness: People and Performance* (2017).

[19] Jennifer M. Logg and Charles A. Dorison. 2021. Pre-registration: Weighing costs and benefits for researchers. *Organizational Behavior and Human Decision Processes* 167 (2021), 18–27. DOI: `http://dx.doi.org/https://doi.org/10.1016/j.obhdp.2021.05.006`

[20] Daniel Moghimi, Berk Sunar, Thomas Eisenbarth, and Nadia Heninger. 2020. {TPM-FAIL}:{TPM} meets Timing and Lattice Attacks. In *29th USENIX Security Symposium (USENIX Security 20)*. 2057–2073.

[21] RENATER. 2020. Conditions d'utilisation du service d'enquête. (2020). `http://web.archive.org/web/20201203200601/https://services.renater.fr/groupware/enquetes/conditions`

[22] Paul M. Sanchez. 2007. The employee survey: More than asking questions. *Journal of Business Strategy* (2007).

[23] Sarah Spiekermann. 2012. The challenges of privacy by design. *Commun. ACM* 55, 7 (2012), 38–40.

[24] Mathieu Trachman, Tania Lejbowicz, and Katharine Throssell. 2018. Putting LGBT and non-binary people in boxes. Statistical categorization and criticism of gender and sexuality assignations in a study on violence. *Revue française de sociologie* 59, 4 (2018), 677–705.

[25] Razieh Nokhbeh Zaeem and K Suzanne Barber. 2020. The effect of the GDPR on privacy policies: Recent progress and future promise. *ACM Transactions on Management Information Systems (TMIS)* 12, 1 (2020), 1–20.