A new PET for Data Collection via Forms with Data Minimization, Full Accuracy and Informed Consent

Nicolas Anciaux
Inria, U. Paris-Saclay (UVSQ)
Versailles, France
nicolas.anciaux@inria.fr

Sabine Frittella
INSA Centre Val de Loire, U. Orléans
Bourges, France
sabine.fritella@insa-cvl.fr

Baptiste Joffroy
INSA Centre-Val de Loire, U. Orléans
Bourges, France
baptiste.joffroy@insa-cvl.fr

Benjamin Nguyen
INSA Centre-Val de Loire, U. Orléans
Bourges, France
benjamin.nguyen@insa-cvl.fr

Guillaume Scerri
U. Paris-Saclay, ENS Paris-Saclay
Orsay, France
scerri@ens-paris-saclay.fr

ABSTRACT

The advent of privacy laws and principles such as data minimization and informed consent are supposed to protect citizens from over-collection of personal data. Nevertheless, current processes, mainly through filling forms are still based on practices that lead to over-collection. Indeed, any citizen wishing to apply for a benefit (or service) will transmit all their personal data involved in the evaluation of the eligibility criteria. The resulting problem of over-collection affects millions of individuals, with considerable volumes of information collected. If this problem of compliance concerns both public and private organizations (e.g., social services, banks, insurance companies), it is because it faces non-trivial issues, which hinder the implementation of data minimization by developers.

In this paper, we propose a new modeling approach that enables data minimization and informed choices for the users, for any decision problem modeled using classical logic, which covers a wide range of practical cases. Our data minimization solution uses game theoretic notions to explain and quantify the privacy payoff for the user. We show how our algorithms can be applied to practical cases study as a new PET for minimal, fully accurate (all due services must be preserved) and informed data collection.

1 INTRODUCTION

Long seen as the resource to fuel the frantic growth of the digital economy, personal data is now considered in many parts of the world as an asset that must be regulated [16, 19] or that citizens should strive to protect [11, 44]. Data collection, as a first step to data exploitation, is carried out through various mechanisms depending on the nature of the data and purpose for which it is collected. Excessive data collection results in oversized personal data sets to be processed and stored, with an unnecessary risk to individual privacy and new form of digital pollution [7].

To avoid these pitfalls, various privacy laws worldwide enforce the principles of data minimization, storage limitation and informed consent. To cite a few examples, the European GDPR [19] states in Article 5 that personal data shall be "*relevant and limited to what is necessary in relation to the purposes for which they are processed*". The Californian CPRA [16] specifically bars companies from collecting more personal data than "*reasonably necessary and proportionate to achieve the purposes*" for which the data was collected. The Australian CDR [6] enacts that data holders "*must not collect more data than is reasonably needed in order to provide the requested goods or services*".

In this paper, we focus on the simplest and most common form of personal data collection: direct collection from users through digital fill-in forms. Indeed, many services collect personal data from users through forms, in order to automate the allocation process for various benefits based on individuals' information. For example, public administrations and government services ask citizens for personal information in order to calibrate social aid and tax benefits. Banks and insurance companies ask applicants to fill in forms in order to calibrate the terms of loans and insurance contracts, etc.

The volume of personal data thus collected through forms is massive and this data collection concerns millions of individuals. For example, hundreds of different types of aids are offered in France at the local or national level, to meet different social needs or provide incentives, with a very large number of beneficiaries. This includes aid for family, education, savings, access to housing, health, transport, mobility, energy saving, legal aid, etc. Governmental portals offer simulators[1] so that citizens can identify the aids they can apply for. Each proposed aid, in order to be obtained, requires the beneficiary to fill out a form to collect their personal data showing that they meet the eligibility criteria and calibrating the aid. The allocation decision is made based on the collected data, which leads to the processing of millions of forms per year[2], with several weeks of instruction by the administration.

Given these staggering numbers, an effective solution to minimize data collected via forms, while properly informing the individual and obtaining their informed consent to the collection process, may lead to significant resource savings and more rigorous compliance with legal principles [16, 19]. This is the objective of our paper. More precisely, our goal is to formalize a data collection process using forms which: (i) preserves *full accuracy*, meaning individuals must obtain the full benefits due to them, (ii) minimizes the data to be filled in the form prior to

[1] For example, aide-sociale.fr offers a rather holistic simulator (more than 1000 aids); the governmental portal mesdroitssociaux.gouv.fr is targeted on the most solicited aids (58 welfare aids) and also offers a simulator for young adults (500 aids). Private initiatives like mes-allocs.fr offer administrative paperwork assistance such that users receive their aid (estimated at 267€ for the average user per month) directly into their account thanks to their experts for a 29.90€ fee per quarter.

[2] For example, the public data set provided by the French government indicates that 6 million forms per year were processed on average over the period 2016-2020 for only 10 family welfare benefits. As another example, the only social aid related to complementary health coverage (which is studied in Section 5) concerns 7.19 million beneficiaries in 2022 –see annual report of the complementary health insurance on p.10–, all of whom had to send in their corresponding form to activate and subsequently renew the aid each year (see same report at bottom of p.8.)

processing and archiving, and (iii) informs the individual of the data collected and of the impact of data minimization choices on inferred data in order to obtain their informed consent.

Data minimization prior to processing has long been considered an unsolvable problem [37], with an overly negative influence on the quality of the processing [15], sometimes in contradiction with other legal obligations such as discrimination prevention [23], fairness [30] or fraud detection [22]. Implementation is deemed too complicated for developers [1, 34], which has led to an overall limited adoption and compliance with this principle.

Some barriers were removed by the introduction of a formalism clarifying the general objectives of a data minimizer [4]. Indeed, recent contributions show that such a formalism can be applied in the context of using machine learning to predict recommendations/personalization through performance metrics [8, 35], while minimizing the data taken as input and still producing a useful model as output, even though its quality is slightly reduced. Such a trade-off between data minimization (or privacy) and performance (utility of the processing) does not apply in our context of data collection through forms, where the full range of benefits due to individuals must be offered by virtue of law: any metric that reduces the legitimate benefits due to a user in exchange for better data minimization would not be acceptable. Moreover bringing information to the individual in order to obtain their informed consent on the collection process is required, and is not tackled by [4]. Nevertheless, this demonstrates the possibility of practical data minimization solutions if a precise mathematical formalism can be defined and transposed to the targeted application context.

We propose a practical solution for personal data collection via forms, where data collected is used to grant benefits to individuals, as practiced by administrations (e.g., social assistance, taxes advantages), banks (loan calibration) or insurance companies (health and insurance contracts conditions). More precisely, we make the following contributions:

- we analyze the data collection problem using forms and formulate the main goals (Section 2);
- we propose a new data minimization model for this problem and an algorithm to resolve it (Section 3);
- we introduce a notion of privacy payoff to quantify the privacy of the individuals and inform them about the personal data collected, and propose an algorithm to decide best subset of personal data items to provide (Section 4);
- we show, through a proof of concept case study, the practical implementation of the proposal as a new PET in real-world scenarios, specifically in the context of typical welfare allocation applications (Section 5).

2 PROBLEM ANALYSIS

This section introduces the problem of over-collection of personal data using forms, it exposes the data minimization and informed consent legal requirements, and formulates our problem.

2.1 Data Collection via Forms

Depending on the service, forms can request anything from a limited amount of information (e.g. name and address when sending a parcel), to vast quantities, particularly when a complex decision process is involved subsequently as is the case in administrative applications (e.g., allocation of social benefits or tax deductions) that express complicated laws, or commercial applications (e.g. bank services or clauses in insurance contracts) whose decisions are based on particularly fine-grained criteria.

The general process illustrating current practices is as follows. First, a decision process is formulated (by the service provider or organization, by the legislator, etc.) determining eligibility criteria for the different benefits of a service offer. This first step is performed only once. Then, for each user applying for the service offer, the process is as follows: (1) applicants retrieve the application form from the service provider; (2) they fill out the form using their personal information (including potentially certified data items retrieved from trusted data providers, e.g., health records) and return it; (3) this personal information passes through the service providers information system, is examined and processed by employees/software to verify the validity of the data items contained in the forms (e.g., by checking certificates or crossing information with external databases) and to evaluate the benefits that the applicant can claim. The processing information is typically (4) stored in a database, possibly for several years, as legal proof of the process and/or transaction, or simply to be used for internal audit or quality control purposes for later use or recourse (e.g., social services must keep data in case of audit).

2.2 Data Collection Problem Example

In practice, the application form is obtained by constructing the union of all data items possibly considered by the decision making process to evaluate the application and build the service proposal. It hence collects all possible data items which *may* impact the final decision, yet maybe for a given user only a small subset will effectively impact that decision. The following running example illustrates this problem[3].

Running Example: *District council benefits scenario.*

> A district council offers the following benefits to its constituents: (i) one can obtain a subsidized public transportation card if one is younger than 25 years old or if one is unemployed and lives in the suburbs; (ii) one can obtain local tax reduction if one is younger than 25 years old and is working (i.e., is not unemployed); (iii) one can obtain a free parking card if one is younger than 25 years old and lives in the town center (i.e., does not live in the suburbs).

For an individual with data values v_1 = [age = 28, unemployed = *true*, suburbs = *true*] the minimum data set would be [unemployed = *true*, suburbs = *true*]. For an individual with v_2 = [age = 20, unemployed = *true*, suburbs = *true*] it would be [age = 20]. Hence, the district council issuing the form cannot specify *a priori* a minimum set of attributes needed since this decision depends on looking at the values of all attributes available.

2.3 Requirements Imposed by Law

To avoid the problem of over-collection and comply with privacy laws regarding data collection, three simple requirements must be met (see Section 6 for more details on privacy laws):

R1: Full Accuracy. The accurate benefits of data collection should be fully delivered to the applicants and the service provider, i.e., applicants receive all the benefits they are eligible for and the service provider can provide the service as intended for the expected duration and comply with the legal necessity, e.g., keep audit records.

[3] This is a simple toy example built for pedagogical purposes only; in a real application, more attributes would be requested, many of which would probably not be needed to take the right decision and thus the privacy gain would be greater. We detail in Section 5 two real use cases from French welfare.

R2: Minimality. As little data as possible should be collected and stored by the service provider. This requirement stems from the data minimization principle enacted in the GDPR [19] or the concepts of constraint of collection and minimization of data in the CPRA [16].

R3: Informed consent. Complete information about their personal data that will be revealed to the service provider through data collection must be provided to applicants so that their decision to provide such data is fully informed (again, compliance with privacy laws such as GDPR or CPRA requires informed consent in most cases of direct data collection).

These three requirements have different impacts on the problem we study in this paper. *R1* is a constraint on the problem we need to solve. *R2* emphasizes the fact that the problem is an optimization problem hence the need to propose a good objective function to quantify the information collected by the service provider. *R3* imposes the quantification of data exposed by the user, so that the user can take an informed decision, when presented with different alternatives.

2.4 Decision Model and Problem Formulation

Decision model. In many traditional automated decision-making processes, especially those that comply with the GDPR's *explicability* [32] regime introduced in GDPR Article 22, the decision procedures must be explained by the service provider to users. These procedures, especially if they have a legal origin, are most often described in natural language by listing the cases in which users are eligible/allowed to access a service. They can thus be encoded into a Classical Propositional Logic (CPL) formula (Definition 3.1) for each service. Every formula of CPL being equivalent to a DNF formula, without loss of generality, we restrict ourselves to Disjunctive Normal Form (DNF) formulas (Definition 3.2 and the real examples in Section 5 with direct transcription of eligibility criteria into logical rules).

We therefore consider that the form to be filled in by a user will be formally described using a set of propositional variables. A completed form thus corresponds to a specific valuation of these variables (Definition 3.3).

We aim at providing a framework that helps us to characterize decision processes (requirement *R3*) and compute them accurately (*R1*) using a minimum amount of information (*R2*). Understanding what is a *minimum* amount of information to communicate to the service provider is linked to two aspects: (1) the information communicated to the service provider by the user must provide a proof (i.e., a set of attribute/value pairs) that the user is allowed to benefit from the service, and as many such proofs may exist, (2) which proof reveals less private information about the user. These proofs are encoded in our framework by so-called *accurate subvaluations* (Definition 3.13). Note that in our context, we are (only) interested in proofs that allow the individuals to obtain all their benefits and the service providers to allocate all possible benefits for efficiency reasons (e.g., in the context of social services, limiting non-take-up of benefits improves the impact of measures [40]).

The main difficulty relies in understanding which proof reveals the least private information about the user. We call this problem *an exposure problem* (Definition 3.11). We characterize solutions to a given exposure problem as *minimal* accurate subvaluations (Definition 3.13). We introduce the notion of *privacy payoff* (Definition 4.1) to quantify the privacy the user has when they only partially fill in the form, along with the necessary information for informed consent to the best possible choice.

Problem formulation. The problem is to find the *minimal accurate subvaluations* (MAS) and to inform users about their choice and privacy. This first challenge is to study a novel distributed optimization problem answering requirements *R1* and *R2*. This is difficult because we are interested in minimizing the knowledge an attacker (with full knowledge of (i) the form, (ii) the decision process of the service provider and (iii) the data minimization algorithm) can deduce about the user from the information the user sends to the service provider. The second challenge concerns how to present the results of the optimization problem to the user, since we will show that the result depends on the strategies of others, thus answering requirement *R3*.

3 RULE-BASED DATA MINIMIZATION

We present here a rule-based model to formalize the data minimization problem in light of the requirements stated above.

3.1 Introducing the Exposure Problem

Preliminaries. Let us recall here some standard definitions and notations of classical propositional logic.

*Definition 3.1 (*CPL*).* Let AtProp be a countable set of propositional variables. The language of *Classical Propositional Logic* (CPL) over the set of propositional variables AtProp, denoted $\mathcal{L}(\text{AtProp})$, is the set of formulas defined by induction as follows:

$A := 0 \mid 1 \mid p \mid \neg A \mid A \vee A \mid A \wedge A \mid A \rightarrow A$, with $p \in \text{AtProp}$.

*Definition 3.2 (*DNF*).* Let AtProp be a countable set of propositional variables. A *literal* is either a propositional variable or the negation of a propositional variable. A formula is in *disjunctive normal form* (DNF) if it is the disjunction of one or more conjunctions of one or more literals.

Application form and benefits. In order to collect (and subsequently process) applicant's data, an application form X_p is built and sent to applicants. In the physical world, this is a document containing blank entries to be filled. In our formalism it is a set of predicates. An applicant (user) will subsequently fill in this application form, by giving values to each predicate.

Definition 3.3 (Ω-(partial)valuations). Let $\Omega \subseteq \text{AtProp}$ represent a set of propositional variables. An Ω-*valuation* is a function from Ω to $\{0, 1\}$. An Ω−*partial-valuation* is a partial function from Ω to $\{0, 1\}$.

Remark that an Ω-valuation is also an Ω-partial-valuation. When it is irrelevant of being a valuation or partial valuation, we write (partial)valuation.

Notation 3.4 ((partial)valuations). We note Val^{Ω} the set of Ω-valuations. Let $\Omega' \subseteq \Omega$, we note $\text{SVal}^{\Omega}_{\Omega'}$ the set of Ω-partial-valuations with domain Ω'. We note $\text{Dom}(v)$ the domain of the (partial)valuation v and $p(v)$ the value of the predicate p for the (partial)valuation v. Note that $\text{SVal}^{\Omega}_{\Omega} = \text{Val}^{\Omega}$.

Definition 3.5 (Subvaluation of). Let v be an Ω-(partial)valuation, we say that w is a *subvaluation of* v if $\text{Dom}(w) \subseteq \text{Dom}(v)$ and $v(x) = w(x)$ for all $x \in \text{Dom}(w)$.

Notation 3.6 (Subvaluation of). We note $w \preccurlyeq v$ if w is a subvaluation of v (note that $\preccurlyeq$ is a partial order).

Definition 3.7 (Application Form). An application *form* is encoded as a finite set X_p of propositional variables representing the items to be filled. A *fully filled form* for a particular individual is encoded as an X_p-valuation v. A *partially filled form* for a particular individual is encoded as an X_p-partial-valuation w with $w \preccurlyeq v$.

Definition 3.8 (Benefits). A *set of benefits* offered by a service provider is encoded as a finite set X_b of propositional variables. We assume that $X_p \cap X_b = \emptyset$. A *fully filled processed form* is encoded as an $\{X_p \cup X_b\}$-valuation to account for the outcome of the decision process.

(Running Example, see Section 2.2) Form and benefits:

> We need to list the benefits X_b offered by the district council (b_1: "subsidized public transportation card", b_2: "local tax reduction" and b_3: "free parking card") and the elements used to decide whether to offer benefits to a given applicant with a fully filled form X_p (p_1: "age $\leq$ 25", p_2: "unemployed" and p_3: "suburbs"). Therefore we build the following sets of propositional variables $X_p = \{p_1, p_2, p_3\}$ and $X_b = \{b_1, b_2, b_3\}$ thus $X_p \cup X_b = \{p_1, p_2, p_3, b_1, b_2, b_3\}$.

General data collection is based on forms filled by the user e.g. giving the precise value of their *age*. This is typically not compliant with minimalization. We use these values to compute the truth value of the predicates used in our application form, e.g. if a user give the value $age = 18$, this will mean $p_1 = true$. The exact value of *age* can thus be deleted.

Decision process. The service provider decision process (or "rules") is formalized by a set of CPL equivalence formulas, the left part of the formula containing only predicates that the applicant can provide (such as "age $\leq$ 25" in the running example), and the right part of the formula is a single predicate representing a benefit (such as "local tax reduction"). If the formula is $true$, it means the applicant must receive the benefit. If the formula is $false$, the applicant must not receive the benefit. In what follows R_{DP} represents such a set of rules encoding a decision process.

Definition 3.9 (Decision process rules). Given a form X_p, a set of benefits X_b, we call set of *decision process rules* R_{DP} a set of CPL equivalence formulas where the left hand side is a DNF formula on predicates in X_p and the right hand side is a single predicate in X_b.

In order to capture additional constraints between predicates, business rules, or the fact that an attribute can be derived from another[4] (such as e.g., "age $\leq$ 25" $\rightarrow$ "age $\leq$ 40", or $\neg(payIncomeTax = true) \rightarrow (annualSalary \leq 15K)$, see Section 5), it is possible to build an additional set of CPL formulas, that we note R_{ADD}. We note $R = R_{DP} \cup R_{ADD}$ this set of decision rules and constraints. In our setting, R is assumed given, but it would also possible to generate these rules from a more expressive logical model and reasoning procedure (e.g. RDF/OWL).

(Running Example) Decision process:

> In the running example, for simplicity's sake we assume that $R_{ADD} = \emptyset$ thus $R = R_{DP}$. We formalize the decision procedure of the district council by translating the textual rules presented in Section 2.2 into the following decision rules R_{DP}:
>
> $$(p_1 \vee (p_2 \wedge p_3)) \leftrightarrow b_1 \quad (1)$$
> $$(p_1 \wedge \neg p_2) \leftrightarrow b_2 \quad (2)$$
> $$(p_1 \wedge \neg p_3) \leftrightarrow b_3 \quad (3)$$
>
> An applicant has to fill in the form $X_p = \{p_1, p_2, p_3\}$ by answering the following questions: p_1: "Are you less than 25 years old?", p_2: "Are you unemployed?" and p_3: "Do you live in the suburbs?". For an applicant with values [age = 28, unemployed = $true$, suburbs = $true$], filling the form means building the following X_p-valuation $v_1 \in \mathrm{Val}^{X_p}$:
>
> $$\begin{aligned} v_1 \ : \ \{p_1, p_2, p_3\} &\rightarrow \{0, 1\} \\ p_1 &\mapsto 0 \\ p_2 &\mapsto 1 \\ p_3 &\mapsto 1. \end{aligned}$$
>
> The associated $\{X_p \cup X_b\}$-valuation, that we note d as in *decision* would be $d = \{p_1 = 0, p_2 = 1, p_3 = 1, b_1 = 1, b_2 = 0, b_3 = 0\}$ with $v_1 \preccurlyeq d$. Here the minimum data set (in terms of set inclusion) to communicate to the district council is [unemployed, suburbs]. Indeed, there is no point for the applicant to send the truth value of $[age \leq 25]$, since regardless of this value, the same benefits are provided (only b_1 in this case). The data set sent to the service provider is represented in our formalism by a X_p-partial-valuation w_1 defined on the subset of attributes $\{p_2, p_3\} \subset X_p$ as:
>
> $$\begin{aligned} w_1 \ : \ \{p_2, p_3\} &\rightarrow \{0, 1\} \\ p_2 &\mapsto 1 \\ p_3 &\mapsto 1. \end{aligned}$$
>
> It is clear that w_1 is a subvaluation of v_1 ($w_1 \preccurlyeq v_1$).

Proof of eligibility. Using this formalism, the objective for an applicant is to produce a partially filled form (X_p-partial-valuation), which will grant them the same set of benefits $F \subseteq X_b$ (as granted per decision rules and constraints R) as their fully filled form (X_p-valuation). If this is the case, then we say that this X_p-partial-valuation is a proof of benefits F under R.

Notation 3.10 (Proof of F by w under R). Let Ω be a set of predicates. Let v be a Ω-valuation and F a set of CPL formulas on Ω. We write $v \models F$ if every formula in F is true under the Ω-valuation v. We write $v \not\models F$ if there is a formula in F that is false under the Ω-valuation v. Let X_p represent an application form and X_b a set of benefits. Let w be a X_p-partial-valuation and R, F sets of formulas on $\{X_p \cup X_b\}$. The notation $w, R \models F$ (proof of the set of benefits F by X_p-partial-valuation w under the constraint R, or simply proof of F by w under R) means that for every $\{X_p \cup X_b\}$-valuation d, we have ($w \preccurlyeq d$ and $d \models R$) implies $d \models F$. The notation $w, R \not\models F$ means that there exists an $\{X_p \cup X_b\}$-valuation d such that: ($w \preccurlyeq d$ and $d \models R$) and $d \not\models F$.

(Running Example) Proofs of eligibility:

> Consider w_1 as defined earlier. It is easy to see that $w_1, R \models b_1$, since for any value of p_1, and $\{X_p \cup X_b\}$-valuation d such that $w_1 \preccurlyeq d$ respects $d \models b_1$: the only two models for d are $[p_1 = 0, p_2 = 1, p_3 = 1, b_1 = 1, b_2 = 0, b_3 = 0]$ and $[p_1 = 1, p_2 = 1, p_3 = 1, b_1 = 1, b_2 = 0, b_3 = 0]$. Thus an applicant will lose no benefit by sending the partially filled form (X_p-partial-valuation) w_1 instead of the fully filled form (X_p-valuation) v_1.

[4] We leave for future work deriving attributes probabilistically using probabilistic logic.

Now consider $w_2 \preccurlyeq w_1 \preccurlyeq v_1$ defined as follows on the subset of attributes $\{p_2\} \subset X_p$:

$$w_2 \ : \ \{p_2\} \rightarrow \{0,1\}$$
$$p_2 \mapsto 1$$

and consider the following $\{Xp \cup X_b\}$-valuation $d_2 = [p_0 = 0, p_1 = 1, p_2 = 0, b_1 = 0, b_2 = 0, b_3 = 0]$. It is clear that $w_2 \preccurlyeq d_2$, and $d_2 \not\models b_1$. Thus we have $w_2, R \not\models b_1$. In other words, an applicant cannot send w_2 instead of v_1 since no proof would be provided for benefit b_1.

Exposure problem and solution. Informally, computing minimal (in the sense of set inclusion of domains) X_p-partial-valuations that prove a given set of benefits F under constraints R, means solving an exposure problem. Solutions, when they exist, are not necessarily unique, and thus are incomparable.

(Running Example) Exposure problem:

Once an applicant has filled in a form with (partial)valuation v_1 to apply to the decision procedure of the district council, represented by the set of formulas $R_{DP} = \{(p_1 \vee (p_2 \wedge p_3)) \leftrightarrow b_1,\ (p_1 \wedge \neg p_2) \leftrightarrow b_2,\ (p_1 \wedge \neg p_3) \leftrightarrow b_3\}$, one can easily compute the unique $\{X_p \cup X_b\}$-valuation $d_1 \ : \ \{X_p \cup X_b\} \rightarrow \{0,1\}$ such that $v_1 \preccurlyeq d_1$ and $d_1 \models R_{DP}$. The $\{X_p \cup X_b\}$-valuation d_1 contains the information from the form of the applicant (because $v_1 \preccurlyeq d_1$) and the decision of the district council (because the valuation of variables b_1, b_2 and b_3 is fully determined by the valuation of variables p_1, p_2 and p_3 and the formulas in R_{DP}). Here, one aims at minimizing the personal data revealed to the district council by the applicant to prove eligibility to subsidized public transportation card (b_1). In this example, this means providing the partial valuation $w_1 \preccurlyeq v_1 \preccurlyeq d_1$ that is such that $w_1, R \models \{b_1\}$.

We define an exposure problem as follows:

Definition 3.11 (Exposure problem). Given a form X_p, a set of benefits X_b, a set of decision process rules and constraints encoded as a finite set R of formulas in the language $\mathcal{L}(X_p \cup X_b)$, we call *exposure problem* E the triplet $E = (R, X_p, X_b)$.

Intuitively, individuals must provide a proof of eligibility for the benefits they can obtain, but no more. What is considered an acceptable proof is captured as follows:

Definition 3.12 ($\preccurlyeq$-minimal element). Let $S \subseteq \mathsf{SVal}^{\Omega}$ and $v \in S$, we say that v is a $\preccurlyeq$-*minimal* element of S if $w \preccurlyeq v$ implies $v = w$ for all $w \in S$.

Definition 3.13 (Minimal and accurate subvaluations (MAS)). Let v be a X_p-(partial)valuation, let w be a X_p-partial-valuation, with $w \preccurlyeq v$. We say that the subvaluation w of v is *accurate* w.r.t. $E = (R, X_p, X_b)$, if for all $x \in X_b$, $v, R \models x$ is equivalent to $w, R \models x$. We say that an accurate subvaluation w of v w.r.t. $E = (R, X_p, X_b)$ is *minimal* if it is $\preccurlyeq$-minimal in the set of accurate subvaluations of v.

Notation 3.14 (MAS). As it does not lead to any confusion, we also use the term MAS to refer to the X_p-partial-valuation that is a MAS.

(Running Example) Solutions to the exposure problem:

This exposure problem is encoded by the triplet $E = (R, X_p, X_b)$ with $R = \{(p_1 \vee (p_2 \wedge p_3)) \leftrightarrow b_1,\ (p_1 \wedge \neg p_2) \leftrightarrow b_2,\ (p_1 \wedge \neg p_3) \leftrightarrow b_3\}$, $X_p = \{p_1, p_2, p_3\}$ and $X_b = \{b_1, b_2, b_3\}$. Using the previous values of v_1 and w_1, w_1 is an accurate subvaluation of v_1 entitled to (only) benefit b_1. This is formalized as:

$$v_1, R \models b_1, \qquad v_1, R \not\models b_2 \qquad \text{and} \qquad v_1, R \not\models b_3.$$

X_p-partial-valuation w_1 fixes only the values of p_2 and p_3 to *true* which implies that the variable b_1 must be *true* in any model d that satisfies every formula in R and such that p_2 and p_3 at *true*, which we note $w_1, R \models b_1$. Hence, we have $v_1, R \models x$ implies $w_1, R \models x$ for every $x \in X_b$. To verify that w_1 is a MAS of v_1, it is enough to show that there is no partial-valuation $w_1' \preccurlyeq w_1$ such that $w_1', R \models b_1$.

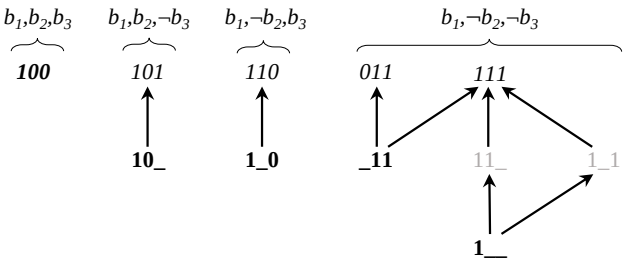

Figure 1: Minimal Accurate Subvaluations (MAS) for the Running Example. *X_p-(partial)valuations corresponding to MAS are identified in bold font; X_p-valuations are in italic; X_p-partial-valuations corresponding to non-minimal accurate subvaluations are in gray; the set of benefits for which each X_p-(partial)valuation is a proof of eligibility is indicated at the top, e.g.,* 110 *and* 1_0 *are eligible to benefits $\{b_1 = 1, b_2 = 0, b_3 = 1\}$; X_p-valuations* 000, 001, *and* 010 *providing no benefit are not represented.*

The formal statement of the minimization problem that we address in this paper is hence as follows:

Minimization problem. Given an exposure problem $E = (R, X_p, X_b)$ and an X_p-valuation v representing a fully filled form of an applicant, compute F the set of benefits triggered by v, then find $W = \{w | w \text{ is a MAS of } v \text{ w.r.t. } E\}$.

In practice, each applicant may decide to chose any $w \in W$ to apply for benefits. W being composed of MAS, this complies with accuracy (R1) and minimality (R2) requirements. Information (R3) imposes to describe the exposure problem.

3.2 Describing an Exposure Problem

An exposure problem can be represented as a directed graph as shown in Figure 1, representing the partially ordered set (in the sense of domain inclusion) of accurate subvaluations on the running example.

Definition 3.15 (Blank attribute values). Given a X_p-partial-valuation w we call *blank values* the attribute values left unset.

Figure 1 is read as follows. It is a directed graph whose nodes are X_p-(partial)valuations and whose edges indicate that a X_p-(partial)valuation is an accurate subvaluation of another one (i.e., providing the same benefits). For instance, 1__ is the X_p-partial-valuation that sets the propositional variable p_1 to *true* and does not set the values of propositional variables p_2 and p_3 (p_2 and p_3 are blank); 11_ is the X_p-partial-valuation that sets propositional variables p_1 and p_2 to *true* and does not set the value of the propositional variable p_3; 111 is the X_p-valuation that sets propositional variables p_1, p_2 and p_3 to *true*. The arrow between the X_p–partial-valuations 1__ and 11_ encodes the fact that 1__ is an accurate subvaluation of 11_. Hence, 1__, 11_, 1_1 and _11 are accurate subvaluations of 111, but only _11 and 1__ are MAS of 111. _11 is both MAS of 011 and 111. Note that 100

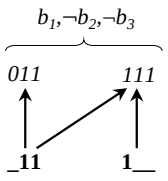

Figure 2: Choices of user u_{111}.

has no accurate subvaluations other than itself, which means that such an applicant will need to send a fully filled form.

Without loss of important information, an exposure problem can be represented by keeping only X_p-partial-valuation corresponding to the MAS of at least one X_p-valuation (in bold font in Figure 1) and X_p-valuations (in italic). In this case, the exposure problem becomes a bipartite graph linking X_p-valuations with their MAS (after removing X_p-valuations with no other MAS than themselves like 100 in Figure 1).

To describe an exposure problem for a user having a specific valuation, let us consider a user u_{111} with valuation 111. In order to achieve requirement *R3*, when deciding which minimal set of predicates to publish in order to receive the benefits they are entitled to, the partial subgraph representing the possible alternatives as indicated in Figure 2, with all the predecessors of the MAS, gives all the information. This is exactly the connected component of the subgraph of the exposure problem containing the X_p-valuation of u_{111}. In reference to game theory, we call this bipartite graph the *choices* of u_{111}. Note that other users, such as u_{011}, may also appear in this graph. The *choices* of u_{111} are important in order to let u_{111} make an informed choice of which MAS to publish: _11 or 1__.

3.3 Informed Data Minimization Algorithm

Our data minimization algorithm constructs the bipartite graph between the MAS and X_p-valuations, for a user with a fixed X_p-valuation and for a given rule set. The result is used (see Section 4) to propose an optimal choice based on privacy payoffs.

Data Minimization algorithm (Algorithm 1). Data minimization is achieved in three phases. First, the set of candidate MAS is constructed (lines 5-13) from the Cartesian product of conjunctions leading to each benefit obtained by the input valuation v; the instances of the Cartesian product potentially giving rise to other benefits not obtained with v are ignored (lines 12-13). Second, the subvaluations of other MAS which are not minimal are removed from this set of candidate MAS (lines 14-17). Third, the set of candidate X_p-valuations for these candidate MAS as well as the X_p-valuation/MAS edges of the bipartite graph are generated (lines 18-29). The candidate X_p-valuations from the MAS (line 19) are obtained by enumerating the unfixed variables $p_i \in X_p$, discarding those that does not correspond to the expected benefits (lines 21-23) and inserting them in the bipartite graph (lines 25-28).

The optimality of a data minimizer is difficult to state in absolute terms. [4] introduce the concept of "best" minimizer as follows (we have adapted the definition to our context):

Definition 3.16 (Best minimizer, adapted from Def.6 in [4]). A minimizer $\mathcal{M}$ for a particular program (in our example, the allocation of benefits to users) is said to be a *"best" minimizer* if there exists no other minimizer $\mathcal{M}'$ able to produce for a given input (here, an X_p−valuation v), an output (a MAS for v) which has as a set of antecedents by $\mathcal{M}'$ a strict superset of the one it has with $\mathcal{M}$.

Having a best minimizer allows, on the one hand, to maximize the average number of predecessors per output (number of X_p-valuations per MAS), and thus to better protect the users (their X_p-valuation is hidden in a larger set). On the other hand, when users have a X_p-valuation allowing a choice between several MAS, it allows them to exercise their choice with the guarantee that no other minimizer algorithm exists to make this choice with a better protection.

Theorem 3.17 (Optimality of Algorithm 1). *Algorithm 1 is a "best" minimizer, given the definition by [4].*

Proof. (Sketch) We can easily show that Algorithm 1 is *by construction* a "best" minimizer, because (1) for any X_p−valuation v, any X_p-partial-valuation $w \preccurlyeq v$ is a subvaluation of one of the MAS, and any set of predecessor $X - p$-valuations of w is therefore a subset of the predecessors of one of the MAS, and (2) the set of predecessors X_p-valuations of a MAS cannot be contained entirely in the set of predecessors of another MAS, because either the MAS would be a subvaluation of the other MAS and thus not minimal, or the MAS would not be accurate, which is proscribed by construction. □

4 GAME THEORETIC USER INFORMATION

In this section, we focus on a given user u_v with X_p-valuation v and aim to answer the question: how can u_v decide which MAS to publish ? As we can see in the running example, u_{111} has the choice between publishing either _11 or 1__. To support the user in making an informed choice (requirement R3), we intend to quantify the privacy gain of each choice. We approach the problem using a game theoretic approach, considering that all users will want to (i) evaluate the "payoff" they obtain with different choices (the higher the payoff, the better the choice in terms of privacy) and (ii) make the "move" (or choice) that maximizes their privacy payoff, while maintaining full accuracy.

4.1 Game Theory Concepts

The need for game theory is linked to the fact that the value of many privacy payoff functions will depend on the different possibilities and subsequent choices (*moves*) of other players.

Users vs. players. In reference to game theory, we consider *players* in a specific *game*. Note that *players* differ from real users. A *user* is a person filling out a form with a valuation[5]. In order to not depend on the set of users, in the game we consider, a *player* is *exactly one* valuation. We show how to quantify the privacy protection obtained by player/valuation v when choosing a MAS $w \preccurlyeq v$. For simplicity's sake, we consider that all valuations are *realistic* (i.e. could correspond to real *users*), but if necessary, it would suffice to discard non realistic valuations.

The game. The objective for each player is to maximize the value of their privacy payoff function (noted PO_f). Informally, a payoff function computes the payoff for a player playing with a specific strategy, given that other players are also playing with their own strategy. The game considered here is a *non-cooperative, symmetric, non-zero-sum, simultaneous* game, played by n players, where n is the total number of possible valuations which produce at least one benefit for a given exposure problem, thus $n \leq 2^{X_p}$.

[5]In the rest, we write *valuation* as a shorthand for X_p-valuation, and *partial-valuation* as shorthand for X_p-partial-valuation.

Algorithm 1: Informed Data Minimization

Input: v: X_p−valuation for a user; R: rule set.
Output: (V, M, E): bipartite graph with vertices V X_p−valuations and M minimal accurate subvaluations, and with edges E.

$V \leftarrow \emptyset$ // empty set of X_p−valuations
$M \leftarrow \emptyset$ // empty set of MAS
$E \leftarrow \emptyset$ // empty set of edges between the two
$C_R \leftarrow \emptyset$ // empty multiset of conjunctions
// Build MAS candidates:
$F_v \leftarrow \{b, \text{benefit in } R \text{ with } v\}$ // benefits of v
for $b \in F_v$ **do** // for each benefit
  $C_b \leftarrow \{c, \text{conjunctions of benefit } b\}$
  $C_R \leftarrow C_R \bigtimes C_b$ // cartesian product
for $C \in C_R$ **do** // for each set of conjunctions
  $w \leftarrow buildSV(C)$ // $w \preccurlyeq v$ with predicates in C
  $F_w \leftarrow \{b, \text{benefit of in } R \text{ with } w\}$
  if $F_w = F_v$ **then** // same benefits
    $M \leftarrow M + w$ // add w to candidate MAS
// Filter out non minimal MAS candidates:
for $m \in M$ **do** // for each candidate MAS
  for $m' \in M$ **do** // for each candidate MAS
    if $m \neq m'$ & $m' \preccurlyeq m$ **then** // if m subval.
      $M \leftarrow M - m$ // filter it out
// Build candidate valuations and edges:
for $m \in M$ **do** // for each MAS m
  $V_m \leftarrow \{v' \ X_p\text{−valuation}, v' \succcurlyeq m\}$ // candidates v'
  for $v' \in V_m$ **do** // for each candidate v'
    $F_{v'} \leftarrow \{b, \text{benefit of } v' \text{ in } R\}$ // its benefits
    if $F_{v'} \neq F_v$ **then** // not same benefits as m
      $V_m \leftarrow V_m - v'$ // remove candidate v'
    else // Build edges for v':
      $E \leftarrow E + (v', m)$ // add edge (v', m)
      // Look at other edges for v':
      for $m' \in M$ **do** // for each MAS m'
        if $v' \preccurlyeq m'$ **then** // v' subval. of m'
          $E \leftarrow E + (v', m')$ // add edge
  $V \leftarrow V + V_m$ // add candidate X_p-valuations
return (V, M, E)

The game is composed of only one move by each player: each player publishes a MAS of their valuation.

Privacy payoff functions. The privacy payoff function (see Definition 4.1) captures the privacy gain, when taking into account knowledge of the exposure problem (hence the possibility of considering every logical consequence using R_{ADD}), and the strategy used by all players (see Section 4.3). We note a game $\mathcal{G}_f = (E, PO_f)$ where E represents an exposure problem $E = (R, X_p, X_b)$ and PO_f represents a payoff function. In what follows, we will consider two different payoff functions, PO_{blank} and PO_{SM}, each modeling a different privacy protection: PO_{blank} evaluates the number of unknown predicates (akin to *plausible deniability* [9]), PO_{SM} counts the number of similar players (akin to *anonymity set size* [42]). These functions present the advantage of leading to the existence of (Nash) equilibrium strategies.

Definition 4.1 (Privacy Payoff). A (privacy) payoff function is a function of domain $(\mathsf{Val}^{X_p} \times \mathsf{SVal}^{X_p}) \times (\mathsf{Val}^{X_p} \times \mathsf{SVal}^{X_p})^{(n-1)}$ and co-domain $\mathbb{R}^+$, representing the privacy guarantees that a player v will have, when publishing an accurate subvaluation w of v following a specific strategy (the first $(\mathsf{Val}^{X_p} \times \mathsf{SVal}^{X_p})$), and when all other players also follow a specific strategy, hence the $n-1$ other $(\mathsf{Val}^{X_p} \times \mathsf{SVal}^{X_p})$ couples. The higher the payoff, the better the minimization.

Nash equilibrium. Equilibrium strategies (Definition 4.2) are optimal strategies in the sense that no player in a given game has a better strategy if all the other players are following this optimal strategy. We propose in Section 4.3 equilibrium strategies linked to our proposed payoff functions.

Definition 4.2 (Nash Equilibrium (adapted from [28])). Let $U = \{u_i\}$ be a set of players, and $V = \{v_i\}$ the set of associated valuations. Let $E = (R, X_p, X_b)$ represent an exposure problem. Let $\mathcal{G}_f = (E, PO_f)$ represent a game using payoff function PO_f. Let S_i represent the set of all possible strategies for player u_i (i.e., their possible moves). We note $S^* = (S_i^*)$ a set of strategies where $\forall i, S_i^* \in S_i$ (one strategy per player). We note $S_{\neg i}^* = S^* \setminus S_i^*$ the set of all these strategies, except the strategy S_i^*. S^* is said to be a *Nash Equilibrium* if $\forall i, \forall s \in S_i$, we have $PO_f(S_i^*, S_{\neg i}^*) \geq PO_f(s, S_{\neg i}^*)$

Attack model. Although not part of game theory concepts, we conclude this section by explaining what type of attacker we consider. Let v represent a player (valuation) of an exposure problem $E = (R, X_p, X_b)$. Player v publishes a partial-valuation (i.e., plays a move) $move(v)$, in principle a MAS, with regards to a game $\mathcal{G} = (E, PO_f)$, and a strategy $\mathcal{S}$. We consider an honest-but-curious attacker $\mathbb{A}$, which is an entity with the knowledge of $\mathcal{G}, \mathcal{S}$ and $move(v)$, who seeks to obtain maximal information about v. Our goal is to (i) correctly quantify this maximal information leak (requirement *R3*) and (ii) minimize this maximal information leak (requirement *R2*).

Why a game-theoretic approach ? When deciding which MAS to publish, a player will take their decision based on the decisions of other players, since these decisions will impact the privacy score of their MAS. The fact that many players are concurrently taking decisions which have an influence on each other is exactly the context where game theory is useful. Indeed, Th.4.6 (see Sec.4.3), showing that there exists an equilibrium strategy means that it is possible and makes sense to compute a correct estimation of the score of each possible MAS, as we are able to anticipate an *optimal* strategy for all players, which they will play if they are *rational*, i.e. want to maximize their privacy payoff.

4.2 Privacy Payoff Functions

To simplify the notations of Definition 4.1, we write the payoff function as $\mathsf{Val}^{X_p} \times \mathsf{SVal}^{X_p} \rightarrow \mathbb{R}^+$, where $\mathsf{Val}^{X_p} \times \mathsf{SVal}^{X_p}$ represents the strategy (leading to play a specific move) of the player, and drop the notation on the other players' strategies. We show that an equilibrium can be reached when all players adopt the same strategy. Note that all payoff functions proposed can be evaluated locally using a valuation and the rules, which means it is simple and efficient to evaluate them.

(Example) Informed choice:

> If a player decides to publish their complete form, this means publishing their valuation and the privacy payoff is 0 (i.e., any entity that gets hold of the form knows all the values of the player's predicates). In the general case, a player publishes a MAS and the privacy payoff function must correctly capture what information is transmitted, so that a player who can publish several MAS can make an informed choice.

PO_{blank}: the number of blank predicates. A straightforward idea, with the benefit of being easy to understand, is to consider

the number of hidden predicates. This is a classical metric, proposed by [3], which evaluates the quality of form minimization as a function of the number predicates published. However, the minimization algorithm used in [3] does not take into account reasoning (e.g. some blank values may in fact be deduced by looking at the rules, see example "Players and choices" below) and hence returns overestimated values. Therefore, the exposure function must be adapted to consider only the MAS as follows.

Definition 4.3 (PO_{blank}). Given a player v and a MAS $w \preccurlyeq$ v, $PO_{blank}(v, w) = \{k |\ k$ is the number of blank values in w that an attacker cannot deduce$\}$.

It is important to understand that it is *not* sufficient to simply count the number of blank values in w, since an attacker can consider all the players in the game.

(Running Example) Players and choices:

> Consider valuation 111 in Figure 2. This player is the only one with a choice to make between MAS and thus whose strategy will have an impact. All other players can only play one MAS, thus they only have one strategy. Assume that this player decides to play 1__, which contains 2 blank values. The attacker knows that only the value of p_1 is used to compute benefits, thus $\{b_1\}$ is the only benefit triggered. Indeed, triggering benefits b_2 or b_3 would mean publishing $\neg p_2$ or $\neg p_3$. This simple reasoning shows this player has values $p_2 = 1$ and $p_3 = 1$. In this case, the number of blank values is (not 2 but) $PO_{blank}(111, 1\_\_) = 0$.
>
> On the other hand, consider the same player plays _11, which contains only 1 blank values (p_1). In this case, the attacker has no deduction possible to determine if $p_1 = 0$ or $p_1 = 1$, since both players u_{011} and u_{111} trigger the same benefit. Thus $PO_{blank}(111, \_11) = 1$ and $PO_{blank}(011, \_11) = 1$.
>
> Similarly, we have $PO_{blank}(100, 100) = 0$, $PO_{blank}(101, 10\_) = 0$ and $PO_{blank}(110, 1\_0) = 0$. Not shown on the schema, $PO_{blank}(000, \_\_\_) = 2$ as is the case for $PO_{blank}(001, \_\_\_) = 2$ and $PO_{blank}(010, \_\_\_) = 2$. Note that in practice, players obtaining no benefits will not send any information, but our approach is able to model that an attacker would be able to deduce that $p_1 = 0$ for these players.

PO_{blank} **privacy protection.** The PO_{blank} payoff function expresses for a player the number of attributes for which plausible deniability can be claimed, i.e., there exists at least another valuation with a different value for all these attributes.

Proposition 4.4 (Computing PO_{blank}). *Computing $PO_{blank}(v, w)$ means computing the set of all valuations $V = \{v' | w \preccurlyeq v'$ and $w \preccurlyeq v\}$, and counting the number of predicates p for which there exists at least two players v_1, v_2 with $v_1, v_2 \in V$ such that $p(v_1) = \neg p(v_2)$ and who will decide to play this move (based on their strategy).*

Proof. (Sketch) A MAS is included in all other X_p-valuations and X_p-partial-valuations leading to this MAS. The set of possible X_p-valuations that lead to this MAS (noted V) can be computed by following all directed paths. An attacker that knows that all players are playing this move is not able to deduce anything more than this set of possible X_p-valuations for this given MAS, thus PO_{blank} can be evaluated by looking at the number of conflicting predicate values in V (see the example on Fig. 1). □

Extending PO_{blank} with weights. In some cases, the sensitivity (in the sense of privacy) of all attributes is not the same. However, users largely agree on the degree of sensitivity of attributes in a given context, as shown e.g. by [38] in the employment context. Thus to support this feature, we only need to slightly modify PO_{blank} by adding a weight for each predicate, and summing the weights of blank nodes, instead of counting them. For simplicity, we consider the unweighted version in the rest of the article[6].

PO_{SM}: the number of players with the same move. As seen when studying the PO_{blank} payoff function, an alternative (and maybe simpler) measure of privacy is to count how many different players will be playing the same move. As previously, this will depend on the strategy chosen by the players.

Definition 4.5 (PO_{SM}). Given a player v and a MAS $w \preccurlyeq$ v, $PO_{SM}(v, w) = \{k - 1 |\ k$ is the number of players playing $w\}$.

PO_{SM} **privacy protection.** The PO_{SM} payoff function expresses the number of different players playing the same move. This can be seen as hiding in a crowd, a bit like k-anonymity. PO_{SM} can also be seen as a measure of the entropy of the move.

We show next that both the PO_{blank} and PO_{SM} privacy payoff functions can be used to build an equilibrium strategy.

4.3 Equilibrium Strategy

We propose an equilibrium strategy for the two non-collaborative games $\mathcal{G}_{blank} = (E, PO_{blank})$ and $\mathcal{G}_{SM} = (E, PO_{SM})$. This strategy, for which the payoff function is simply a parameter, is described in Algorithm 2. The intuition behind this strategy is for each player to know exactly which move all the other players are going to make, and thus take the optimal decision in terms of PO_{SM} or PO_{blank}. When players have a seemingly equivalent choice to make, their decision is taken based on a predefined ordering of the moves (e.g., lexicographical order can be defined as the canonical order on words on the ordered alphabet _, 0, 1).

Strategy selection algorithm (Algorithm 2). The strategy applied for each player with valuation v_i is as follows: (1) if the player v_i has only one possible move w_i (i.e., v_i has a single MAS w_i), they play it (lines 1-3); (2) if v_i can play a move that is the best of all possible moves of all other players (i.e., one of the MAS of v_i has the highest payoff), they play this move (lines 4-6), or they play the first move in lexicographical order when several best moves exist (lines 7-10); otherwise (lines 11-18) (3) if other players have better moves assume all players play their best move in succession, and each time, recompute the values of the privacy payoff function. Wait until the payoff of best move dominates all other to play it – back to case (2).

Correctness of the strategy. Each time the strategy is recursively called, either the strategy returns a move, or it recursively calls itself after having decided which player is going to play the best move, which means in the worst case $n - 1$ recursive calls (where $n \leq 2^{|X_p|}$ is the number of players with at least one benefit). Thus $\mathcal{G}_f$_strategy defines a strategy for a player.

(Running Example) Applying the strategy:

> On the simple example of Figure 1, for each player, we run the $\mathcal{G}_f$_strategy (Algorithm 2):
>
> - Players 100, 101 and 110 have only one possible move, with zero privacy payoff: $PO_{blank}(100, 100) = 0$,

[6] If we assume that we have obtained such weights (e.g. from the user, or by asking a few general questions to the user to try to evaluate which topics are sensitive for them), then one could either compute a personalized payoff for the MAS's or present the results of the algorithm in a user-friendly manner (e.g. if medical information is sensitive to the user, then one could highlight that one MAS reveals more medical information than the other).

$PO_{blank}(101, 11\_) = 0$, and $PO_{blank}(110, 1\_0) = 0$, or $PO_{SM}(100, 100) = 0$, $PO_{SM}(101, 11\_) = 0$, and $PO_{SM}(110, 1\_0) = 0$;
- Player 011 has only one possible move: _11, with positive privacy payoffs $PO_{blank}(011, \_11) = 1$ and $PO_{SM}(011, \_11) = 1$, based on the fact that player 111 also plays the same move, which is its best move;
- Player 111 can play either $w_1 = 1\_\_$ or $w_2 = \_11$. Regardless of the choices of the other players, its best move is w_2, since $PO_{blank}(111, \_11) = 1$ and $PO_{blank}(111, 1\_\_) = 0$ (resp. $PO_{SM}(111, \_11) = 1$ and $PO_{SM}(111, 1\_\_) = 0$).

Equilibrium strategy for G_f. We show that this strategy is an equilibrium strategy for PO_{blank} and PO_{SM}.

Theorem 4.6 (Equilibrium strategy for G_f). *Let $E = (R, X_p, X_b)$ represent an exposure problem. Let $\mathcal{G}_{blank} = (E, PO_{blank})$ (respectively $\mathcal{G}_{SM} = (E, PO_{SM})$ represent a game played by $n \leq 2^{|X_p|}$ players. $\mathcal{G}_f$_strategy with $f = blank$ (resp. $f = SM$) is an equilibrium strategy for $\mathcal{G}_{blank}$ (resp. $\mathcal{G}_{SM}$).*

Proof. (Sketch) Lines 3 and 6 represent the fact that if the player has a move that dominates all other moves, regardless of what the other players play, then this move is played. This player has no incentive to change their strategy (move) in this case. Lines 10 and 15 indicate that if the player has several moves with the same payoff, then they decide to play the first one in lexicographical order. Should a player v_i decide to follow a different strategy, e.g. playing in reverse lexicographical order w'_i instead of the lexicographical order move w_i, this will never be beneficial for them, since the other players will be anticipating that they will be playing w_i, and thus will not play w'_i, the same move as them. In consequence, $PO_f(v_i, w_i) \leq PO_f v_i, w'i$. □

$\mathcal{G}_f$_strategy as a solution to data minimization problem. As using $\mathcal{G}_f$_strategy as a strategy for all players is an equilibrium for $\mathcal{G}_f$, this means that we can anticipate how players will behave when faced with a specific exposure problem. Hence, the payoff functions can be used to correctly assess and **inform** a user (associated with a player) of (i) the MAS to publish and (ii) the protection received, thus answering requirements R1, R2 and R3.

5 PROOF OF CONCEPT PET CASE STUDY

We illustrate how a new Privacy Enhancing Technology (PET) based on the concepts introduced would work on two real cases studies, to assist users with the minimization process (R2), produce information necessary for informed consent (R3), while the organization receiving the minimized form can still provide all the due benefits (R1). The first scenario concerns applications for *complementary health coverage* (noted *H-cov* in this section) in France. The eligibility criteria are simple enough to allow an end-to-end process to serve as a proof of concept and demonstrate the feasibility of the proposal. The second scenario concerns the application form for the *active solidarity income*[7] (noted *RSA*) in France, which is a more complex form leading to more complex rules[8]. It is a widely allocated social aid in France allocated to more than two million households[9]. We use this second case to confirm the possible impact of our proposal.

[7] "The Active Solidarity Income", Evaluation of a public policy, Jan. 2022 (link).
[8] Eligibility criteria and benefits are outlined on the French Family Allowance Fund.
[9] See the Court of Auditors website.

Algorithm 2: $\mathcal{G}_f$_strategy: strategy selection

Input: v_i: a Player; $S^*_{\neg_i}$: the set of all possible moves for players other than v_i; $otherMoves$: a set of moves from other players initialized at $\emptyset$; PO_f: a non-collaborative privacy payoff function.
Output: w_{best}: the best move for v_i.

```
// Current player v_i has a single move:
1  if there exists only one MAS w_i such w_i ≼ v_i then
2      w_best ← w_i
3      return w_best // Play this move
   // Player v_i has best moves wrt the other players:
4  if there exists moves w_i such that
     ∀j, PO_f(v_i, w_i) > PO_f(S*_¬i \ otherMoves) then
5      if this move is unique then
6          return w_i // Play this move
7      else // this move is not unique
8          order the moves {w_i} by lexicographical order
9          w_best ← the first move in lexicographical order
10         return w_best// play the first in lex. order
   // Build set of players with the top score move:
11 V_top ← all players v_k such that ∃w_k ≼ v_k such that
     PO_f(v_k, w_k) > PO_f(S*_¬i \ otherMoves ∪ {v_k})
12 v_top ← first player of V_top in lexicographic order
13 w_top ← first best move of player v_top in lexicographic order
14 if v_i = v_top then // Current player is a top player
15     w_best ← w_top
16     return w_best // Play this top move
17 S*_temp ← S*_¬i \ all moves of {v_top}
18 otherMoves ← otherMoves ∪ {v_top, w_alpha}
19 return G_f_strategy(v_i, otherMoves, S*_temp, PO_f)
```

Health coverage (H-cov) scenario. We focus on real eligibility criteria and form[10] used to apply for complementary health coverage to the French social security system. This aid is one of the 19 main social aids proposed in France at the national level (many other aids are proposed at local levels) for which the government proposes an official simulator[11]. We choose this particular aid for our case study because it concerns 7.19 million beneficiaries with a total budget of 2.8 billion euros[12] and is well documented with a precise textual description of the eligibility criteria. A two-step methodology is used to provide a new PET implementing data minimization and informed consent based on our proposal:

Step 1: Derive the predicates and set of rules. This first step is performed by the organization (here, the French welfare system) for each form used, so that the resulting rule set can be fed as input to the data minimization and informed user consent algorithms. Below, the quoted text is a direct -translated- transcription of the eligibility criteria announced on the French social security website[5], with the numbered predicates used to form the logical rules added in brackets:

> *"The applicants are under 16 years old (p_1) and are under the jurisdiction of the child welfare system (p_2); the applicants are minors over 16 years old (p_3) and have broken off their family ties (p_4); the applicants are adults below 25 years old (p_5), no longer*

[10] The eligibility criteria can be found on the French social security website Ameli. In French: "Une demande individuelle est possible (...)" of which we give an English translation in this section– and the corresponding form is available here.
[11] The simulator is accessible here and informs the French citizen about 58 social aids among which 19 national ones.
[12] See p.10 of the 2022 French social security report available here.

Predicates X_p	Decision Process Rules R_{DP}
p_1: "age below 16"	$(p_1 \wedge p_2)$
p_2: "child welfare"	$\vee(p_3 \wedge p_4)$
p_3: "minor over 16"	$\vee(p_5 \wedge p_6 \wedge p_7 \wedge \neg p_8)$
p_4: "broken family tie"	$\vee(p_5 \wedge \neg p_6 \wedge p_9)$
p_5: "adult below 25"	$\vee(p_6 \wedge p_{10} \wedge p_{11})$
p_6: "not same roof"	$\vee p_{12}$
p_7: "separate tax return"	$\leftrightarrow b_1$
p_8: "receive alimony"	Additional consitency rules R_{ADD}
p_9: "with child"	$p_1 \rightarrow \neg p_3 \wedge \neg p_5$
p_{10}: "student"	$p_3 \rightarrow \neg p_1 \wedge \neg p_5$
p_{11}: "emergency aid"	$p_5 \rightarrow \neg p_1 \wedge \neg p_3$
p_{12}: "separated"	$p_{12} \rightarrow \neg p_1$

Table 1: Predicates and rules for H-cov.

Characteristics	H-cov	RSA
Number of MAS	6	24
Number of valuations	1560	1296
Number of predicates per MAS	2 to 6	9 to 13
Number of valuations with a single MAS	1272	368
Number of valuations with 2 MAS	280	526
Number of valuations with 3 MAS	8	144
Number of valuations with 4 MAS		172
Number of valuations with 6 MAS		66
Number of valuations with 8 MAS		14
Number of valuations with 12 MAS		6

Table 2: MAS eligible in H-cov and RSA.

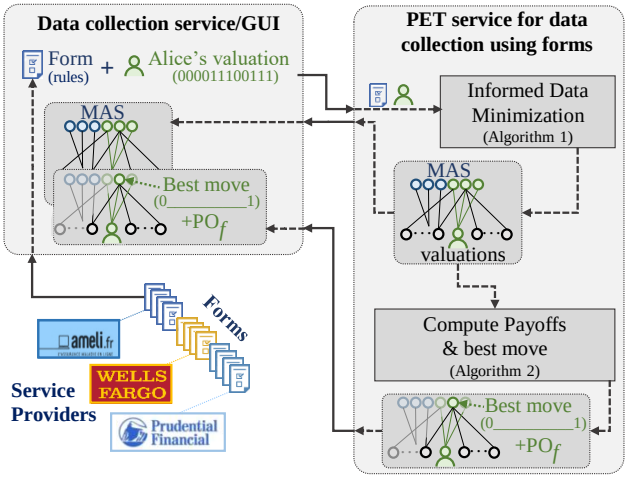

Figure 3: PET for Informed Data Minimization in forms.

MAS	players	PO_{SM}	PO_{blank}
0_____1	1024	1024 (744,1024)	10 (10,10)
0_0__1___11_	128	64 (56,128)	6 (6,7)
0_0_10__1___	128	64 (64,128)	6 (6,7)
0_0_1110____	64	24 (24,64)	5 (5,6)
0_110_____	256	128 (128,256)	7 (7,8)
110_0_____	256	256 (256,256)	8 (8,8)

Table 3: The payoffs for the selected MAS (H-cov).

> *live under the same roof as their parents (p_6), file a separate tax return (p_7) and do not receive alimony ($\neg p_8$); the applicants are adults below 25 years old (p_5), live under the same roof as their parents ($\neg p_6$) and are themselves parents (p_9); the applicants are isolated (p_6) students (p_{10}) and receive annual emergency aid (p_{11}); or the applicants are separated from their spouse (p_{12})."*

This leads to the predicates and rules summarized in Table 1, with only one rule with a single benefit noted b_1 (i.e., be eligible for health coverage) and some obvious consistency rules between predicates (additional constraints R_{ADD}, see Section 3.1).

Step 2: Data minimization, informed choice and consent. Our PET takes the form of a service for developers and service providers who collect personal data through forms, in order to retrieve all the information necessary to inform users (e.g., by means of a user interface) and collect their consent to data collection (see Figure 3). It takes as input the set of rules presented above and asks the user to indicate the set of predicates they validate (i.e., their valuation). Using Algorithm 1, the GUI can retrieve (1) all minimal subsets of predicates sufficient to demonstrate the user's eligibility for supplemental health coverage (i.e., MAS), and (2) all valuations corresponding to other potential users (i.e., players in Section 4.1) leading to these same MAS. This constitutes all the information conveyed by the predicates that users can provide (and do not provide) and is available to the GUI to inform the user. Algorithm 2 is used to help the users select the set of predicates to be provided (or not) within their form, by computing the privacy payoffs of their MAS.

Results on H-cov scenario. We ran Algorithm 1 to obtain the bipartite graph between all valuations and MAS. As shown in Table 2, there are 6 possible MAS and 1560 possible valuations providing the benefit b_1. Out of these valuations, 1272 are associated to a single MAS (users with only one choice to fill in the form), 280 lead to a choice between 2 MAS and 8 valuations offer a choice between 3 MAS. A first conclusion is that the data minimization rate is important, as users have to provide only 2 to 6 predicates instead of 12 to be granted b_1. Note that the execution time is acceptable, a few minutes on a desktop[13].

Second, we run Algorithm 2 on each valuation to compute the payoffs PO_{blank} and PO_{SM} (execution time is a few seconds). The obtained values are reported in Table 3. The first value listed in the PO_{SM} (resp. PO_{blank}) column corresponds to the final value, and between brackets, the left value is the number of players who have no other choice than to play this MAS (resp. real number of hidden predicates), and the right value is the maximum number of players who can play this MAS (resp. maximum number of hidden predicates). Consider MAS 0_____1 and PO_{SM}. There are 744 players who can only play this MAS and 280 players who can also play another MAS (0_0__1___11_, 0_0_10__1___, 0_0_1110____ or 0_110_____). As 0_____1 is the best case all these players play this MAS. Consider MAS 0_0__1___11_ and PO_{SM}. 56 players can play only this MAS, 64 have already played 0_____1 and 8 can also play 0_0_1110____. Concerning PO_{blank}, consider users disclosing 0_0_1110____. 6 predicates are supposed to be left "blank" (i.e., p_2, p_4, p_{9-12}), but only 5 are actually unrevealed. Indeed, the predicate p_{12} is revealed because the 32 valuations whose choice of MAS is 0_0_1110____ rather than 0_____1 all necessarily have the same value $p_{12} = 0$. This is why the value of PO_{blank} is 5 and not 6.

[13]The code is written in Java, run with Java SE version 15 on a DELL Precision Intel Xeon W2225 4.6GHz with 64 GB RAM.

MAS	players	PO_{SM}	PO_{blank}
_1001__0_11___111	128	32 (28, 128)	5 (5, 7)
_100_1___11___111	256	128 (70, 256)	8 (7, 8)
_100__1__11___111	256	256 (70, 256)	7 (7, 8)
_1000__1111___111	64	16 (14, 64)	4 (4, 6)
_1001__0_1_111111	32	4 (4, 32)	2 (2, 5)
_100_1___1_111111	64	16 (10, 64)	5 (4, 6)
_100__1__1_111111	64	32 (10, 64)	4 (4, 6)
_1000__111_111111	16	2 (2, 16)	1 (1, 4)
0_1_1__0_11___111	256	128 (28, 256)	8 (5, 8)
0_1_11___11___111	256	16 (14, 256)	7 (4, 8)
0_1_1_1__11___111	256	48 (14, 256)	6 (4, 8)
0_1_1__1111___111	128	96 (14, 128)	5 (4, 7)
0_1_1__0_1_111111	64	16 (4, 64)	5 (2, 6)
0_1_11___1_111111	64	2 (2, 64)	4 (1, 6)
0_1_1_1__1_111111	64	6 (2, 64)	3 (1, 6)
0_1_1__111_111111	32	12 (2, 32)	2 (1, 5)
0__11__0_11___111	256	256 (28, 256)	7 (5, 8)
0__111___11___111	256	128 (14, 256)	6 (4, 8)
0__11_1__11___111	256	16 (14, 256)	5 (4, 8)
0__11__1111___111	128	32 (14, 128)	4 (4, 7)
0__11__0_1_111111	64	32 (4, 64)	4 (2, 6)
0__111___1_111111	64	16 (2, 64)	3 (1, 6)
0__11_1__1_111111	64	2 (2, 64)	2 (1, 6)
0__11__111_111111	32	4 (2, 32)	1 (1, 5)

Table 4: The payoffs for the selected MAS (RSA).

Application to real users (H-cov). Consider two users, Alice and Bob, wishing to apply for medical coverage for which they are eligible. Alice is 24 years old. She lives separated from her spouse and parents and files a separate tax return. She has resumed her studies and receives annual emergency aid (the valuation of Alice is 000011100111). To fill in the form, Algorithm 1 offers her 3 choices : 0_____1, 0_0__1___11_ or 0_0_1110____. Algorithm 2 suggests making the first choice, which reveals that she lives separated and is older than 16 years old, but preserves her privacy concerning the 10 other predicates.

Bob is a 20 years old father, financially independent from his parents, who lives with his daughter and her mother (Bob's value is 000011100000). To request the medical coverage, Algorithm 1 offers only one solution to Bob : 0_0_1110___. By considering only the MAS, Bob may believe he can hide that he is living with the mother of their daughter. However, the GUI informs Bob that predicate p_{12}, not included in his response, is nevertheless disclosed, thus he is living with someone. Indeed, had this not been the case, Bob could have sent the blank form 0_____1.

Results on RSA scenario. The RSA scenario has 17 predicates with benefits granted incrementally (eligible users receive at least the basic financial income, plus in some cases, additional income). Table 2 (right column) and Table 4 summarize the results. For this form, we get 24 MAS and 1296 valuations with at least one benefit. Many users have choices (with up to 12 different MAS) to get the benefits due. The data minimization rate is important, as up to 8 predicates out of 17 can be omitted.

Conclusion. The statutory requirements are met as follows:

R1: Accuracy. The forms are filled out in such a way that the user receives all the benefits due to them. This is ensured by Algorithm 1, which offers the user only partial-valuations that will trigger the full set of achievable benefits.

R2: Minimality. Algorithm 1 minimizes the data from the user's valuation so that as little data as possible is collected through the form and then processed and stored by the service provider, in compliance with the data minimization principles of privacy laws such as GDPR [19] or CPRA [16]. Minimization ratio can be significant, as over 70% for H-cov and 30% for RSA of the predicates are removed for a population of valuations taken uniformly across all eligible valuations.

R3: Informed consent. Algorithm 2 allows developers to have all the information to explain to users the collection choices made, and what is exactly revealed by their form.

6 RELATED WORK

Informed consent. For consent to be valid under regulations such as the GDPR, [12] studies different channels that must be used (sms, email or voice message) in different situations and identifies the characteristics required for consent to be considered informed in a general framework: the subject must be able to understand "*what the consent is for, what the implications are, and what risk the processing of his or her data would entail.*" Other work addresses the question of the right semantics to represent and inform consent[25] and enable the design of consent management platforms [29]. The importance of not just requiring consent, but providing clear requirements on how to ensure that users can make an informed choice is stressed in [39]. Other works propose to minimize the number of subjects from whom to seek input in an open system given the difficulty of obtaining consents [17] or maximize data reuse in accordance with consent when consent has already been obtained [10]. All this only reinforces the need for the notion of informed consent in our context.

Legal principles of data minimization. Data minimization is enacted in the EU GDPR Article 5(1c) [19], stipulating that personal data shall be "*adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimization').*" It is completed by storage limitation (which can be viewed as applying data minimization before storage) in Article 5(1e): "*kept (...) for no longer than is necessary for the purposes for which the personal data are processed (...) ('storage limitation').*" In addition to data minimization and storage limitation, GDPR defines informed consent in Article 6 and gives in Article 13 individuals an additional right to be informed when personal data is directly obtained from the data subject (as in our context of form-based data collection), which means providing all necessary information to understand which data is collected and for which reasons. Similar privacy principles are included in privacy laws worldwide like in the US [16] and in Australia [6].

Their implementation in practice. The principle of storage limitation can be seen as a simple form of data minimization, intended to delete data that are no longer needed, in order to reduce storage costs and avoid unexpected personal data leakage. Solutions have been proposed in the database context [2] using automata triggering progressive generalization and destruction of collected data when purposes are fulfilled. Similar principles are applied before off-line data storage [24] and integrated in continuous business process [43]. These simple solutions do not address the broader context of data minimization and informed consent in the case of personal data collected via forms as addressed in this paper. Data minimization is more complex and still remains difficult today to be apprehended in practice by developers as shown by many recent field studies [1, 26, 33].

General formalism. The obstacles identified to the adoption of data minimization techniques are the lack of clear mathematical definitions allowing their interpretation [4] and the incompatibility with other legal principles such as auditing [5]. A formal definition of a data minimizer is introduced in [4] as a pre-processor that reduces the information given to a program (as input) without compromising its functionality (as output). [5] proposes to encode key articles of the GDPR, including Article 5(1c) on data minimization, in logical form for automatic and auditable processing. These proposals give a general framework but more effort is needed for an application of data minimization in each specific concrete context.

Application of data minimization. Some proposals build on such general theoretical framework and propose data minimizers for particular cases. A solution for Trigger-Action Platforms proposed in [14] (TAP) relies on the semantics of user-created TAP rules regulating automatic exchanges of personal data between applications (e.g., from Gmail to Slack), to establish the data minimizer. Several solutions target data minimization for fairer machine learning programs [41], in particular prediction models [8, 36]. [8] shows that data minimization can be implemented in the context of prediction models through performance (model accuracy) metrics. The data minimizer ensures that the produced prediction model is obtained with good accuracy using a minimum numbers of features. This approach is extended in [31] to the case of automatic auditing, to allow the verification in a black-box setting of the necessity of the various model inputs. Obviously, this work does not apply to the particular issue of data minimization in forms, as full accuracy (see requirement R3 in Section 2.3) would not be achieved. In addition, unlike our proposal, these works do not propose a solution to minimize the data provided by a given user (i.e., every individual provides all the requested features, with no other possible choice) nor inform users why a given feature should not be provided or what it brings with respect to their own privacy.

Previous on minimization of data collected via forms [3] propose to rely on logical rules (as our proposal), but do not take into account the deductions that the attacker can make. It is therefore not a data minimizer in the sense of [4]. As a consequence, the user is only weakly protected, and does not get the necessary information allowing them to correctly apprehend their level of protection. To the best of our knowledge we are the first to propose a solution implementing data minimization for forms, thus there are no concurrent proposals we could evaluate ours against. We leave the efficiency questions of our approach to future work but we stress that as shown in our online demonstration, our prototype is capable of efficiently assessing real use cases.

General data protection approaches. There are many studies on privacy-preserving data publishing (PPDP)[13, 21], and solutions proposed, such as differential privacy [18] or GDPR concepts such as unlinkability. However, our context involves forms with personal information, and our goal cannot be to publish anonymized versions of these forms. Instead, we aim to minimize the amount of data they contain, as an additional measure to prevent data leakage. Thus, our research focuses on data minimization as a additional defense mechanism. Despite the fact the concept of data minimization is well understood in regulations, to the best of our knowledge, very little (technical) work has been done to implement it in a less than trivial approach concerning forms: a recent survey [27] states: "only such personal data will be collected that is appropriate for the intended context of the AI services used (data minimization)" but cites only a single reference on the topic of data minimization.

ZKP and anonymization. As stated above, approaches as anonymization [18] or anonymous credentials using e.g. Zero-knowledge proofs of knowledge (ZKP)[20] can be used to break the link between an individual and a credential, however this can not be used in our use case since we need to be able to trace back to the individual to give them the benefit. ZKPs actually illustrates the interest of our approach : a user publishing a ZKP of benefits is exactly publishing the list of benefits. As the set of rules is public, a lot of knowledge can be deduced (and could be calculated by using our framework). Thus similarly, a user should be informed of the data leakage of a ZKP.

7 CONCLUSION, LIMITS AND NEXT STEPS

In this paper, we propose a new modelization of the problem of data minimization and user informed choices to guide what personal data should be provided in a form used to evaluate decision procedures encoded by DNF rules. We propose a theoretical solution to the data minimization problem and an algorithm to implement this solution. We also propose a new way to inform the users of their choices using game theory with a new notion of privacy payoff and a corresponding algorithm where a Nash equilibrium can be obtained. Finally, we show how these algorithms can be used as a new PET on real cases.

Discussion and limitations. The model chosen (i.e. Classical Propositional Logic) is deliberately simple, and does indeed not capture all types of decision making processes (e.g. neural networks). However, our goal is to show the feasibility of the approach in a real world setting. Indeed, despite its simplicity, CPL is expressive enough to offer a usable solution in large scale uses cases (RSA concerns tens of millions of annual requests). We also assume a lexicographical order on the predicates. In order for the approach to work, there needs to be a central authority which will define this order. This can trivially be done by the service provider on form download.

Future work. The primary perspective of this work is to develop a practical version of the proposed PET and to evaluate its adoptability in the field. Interesting perspectives at the technical level include the support of specific payoff function for players (rather than considering a single one for all players). Considering the probabilistic case would also be valuable as a probabilistic minimization strategy would potentially allow an increase in the privacy gains with plausible deniability-based metrics because the number of potential valuations predecessors of each MAS would naturally increase. Integrating a probabilistic part in the strategy is non-trivial as it means considering mixed strategies equilibrium for Algorithm2. Another interesting direction is to explore "solidarity" privacy metrics and strategies, i.e., able to take into account users' choices while allowing to increase the minimum value of privacy payoff when making choices. Indeed, it is sometimes enough that very few users change their choice, with potentially a very small penalty, for a large number of other users to benefit from a privacy increase. For instance, in the concrete example of Section 5 assigning a complementary health insurance, 24 players are forced to make the least favorable choice (MAS 0_0_1110____) with the lowest privacy payoff ($PO_{blank} = 5$). Only one more player (chosen appropriately) is needed to increase the gain to 6 for these 24 players.

This work was supported by grants ANR JCJC 2019, projects PRELAP (ANR-19-CE48-0006) and iPoP PEPR (ANR-22-PECY-0002).

REFERENCES

[1] A. Alhazmi and N. A. G. Arachchilage. I'm all ears! listening to software developers on putting gdpr principles into software development practice. *Personal and Ubiquitous Computing*, 25(5):879–892, 2021.

[2] N. Anciaux, L. Bouganim, H. Van Heerde, P. Pucheral, and P. M. Apers. Instantdb: Enforcing timely degradation of sensitive data. In *2008 IEEE 24th International Conference on Data Engineering*, pages 1373–1375. IEEE, 2008.

[3] N. Anciaux, B. Nguyen, and M. Vazirgiannis. Limiting data collection in application forms: A real-case application of a founding privacy principle. In N. Cuppens-Boulahia, P. Fong, J. García-Alfaro, S. Marsh, and J. Steghöfer, editors, *Tenth Annual International Conference on Privacy, Security and Trust, PST 2012, Paris, France, July 16-18, 2012*, pages 59–66. IEEE Computer Society, 2012.

[4] T. Antignac, D. Sands, and G. Schneider. Data minimisation: a language-based approach. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, pages 442–456. Springer, 2017.

[5] E. Arfelt, D. Basin, and S. Debois. Monitoring the GDPR. In *European Symposium on Research in Computer Security*, pages 681–699. Springer, 2019.

[6] Australia. Chapter 3: Privacy Safeguard 3 — Seeking to collect CDR data from CDR participants. *CDR Privacy Safeguard Guideline*, 2021.

[7] O. Ben-Shahar. Data pollution. *Journal of Legal Analysis*, 11:104–159, 2019.

[8] A. J. Biega, P. Potash, H. Daumé, F. Diaz, and M. 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.

[9] V. Bindschaedler, R. Shokri, and C. A. Gunter. Plausible deniability for privacy-preserving data synthesis. *arXiv preprint arXiv:1708.07975*, 2017.

[10] M. Brahem, G. Scerri, N. Anciaux, and V. Issarny. Consent-driven data use in crowdsensing platforms: When data reuse meets privacy-preservation. In *2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10. IEEE, 2021.

[11] F. Brunton and H. Nissenbaum. *Obfuscation: A user's guide for privacy and protest.* Mit Press, 2015.

[12] A. C. Carvalho, R. Martins, and L. Antunes. How-to express explicit and auditable consent. In *2018 16th Annual Conference on Privacy, Security and Trust (PST)*, pages 1–5. IEEE, 2018.

[13] B. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala. Privacy-preserving data publishing. *Found. Trends Databases*, 2(1-2):1–167, 2009.

[14] Y. Chen, M. Alhanahnah, A. Sabelfeld, R. Chatterjee, and E. Fernandes. Practical data access minimization in {Trigger-Action} platforms. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2929–2945, 2022.

[15] B. Custers and H. Uršič. Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection. *International data privacy law*, 6(1):4–15, 2016.

[16] L. Determann and J. Tam. The california privacy rights act of 2020: A broad and complex data processing regulation that applies to businesses worldwide. *Journal of Data Protection & Privacy*, 4(1):7–21, 2020.

[17] O. Drien, A. Amarilli, and Y. Amsterdamer. Managing consent for data access in shared databases. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 1949–1954. IEEE, 2021.

[18] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[19] European Council. Regulation EU 2016/679 of the European Parliament and of the Council. *Official Journal of the European Union (OJ)*, 59(1-88):294, 2016.

[20] U. Feige, A. Fiat, and A. Shamir. Zero-knowledge proofs of identity. *Journal of Cryptology*, 1(2):77–94, 1988.

[21] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, 2010.

[22] L. Găbudeanu, I. Brici, C. Mare, I. C. Mihai, and M. C. Şcheau. Privacy intrusiveness in financial-banking fraud detection. *Risks*, 9(6):104, 2021.

[23] G. Galdon Clavell, M. Martín Zamorano, C. Castillo, O. Smith, and A. Matic. Auditing algorithms: On lessons learned and the risks of data minimization. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 265–271, 2020.

[24] M. Kangwa, C. S. Lubobya, and J. Phiri. Prevention of personally identifiable information leakage in e-commerce via offline data minimisation and pseudonymisation. *Int. J. Innov. Sci. Res. Technol*, 6(1):209–212, 2021.

[25] G. Konstantinidis, J. Holt, and A. Chapman. Enabling personal consent in databases. *Proceedings of the VLDB Endowment*, 15(2):375–387, 2021.

[26] A. Mazeli. A framework to support software developers in implementing privacy features. In *2022 IEEE/ACM 44th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, pages 245–247. IEEE, 2022.

[27] C. Meurisch and M. Mühlhäuser. Data protection in ai services: a survey. *ACM Computing Surveys*, 54(2):1—-38, 2021.

[28] J. F. Nash. Non-cooperative games, 1951.

[29] M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. 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*, pages 1–13, 2020.

[30] Q. Ramadan, D. Strüber, M. Salnitri, J. Jürjens, V. Riediger, and S. Staab. A semi-automated bpmn-based framework for detecting conflicts between security, data-minimization, and fairness requirements. *Software and Systems Modeling*, 19(5):1191–1227, 2020.

[31] B. Rastegarpanah, K. Gummadi, and M. Crovella. Auditing black-box prediction models for data minimization compliance. *Advances in Neural Information Processing Systems*, 34:20621–20632, 2021.

[32] A. Selbst and J. Powles. "meaningful information" and the right to explanation. In *Conference on Fairness, Accountability and Transparency*, pages 48–48. PMLR, 2018.

[33] A. Senarath and N. A. G. Arachchilage. Understanding software developers' approach towards implementing data minimization. *CoRR*, abs/1808.01479, 2018.

[34] A. R. Senarath and N. A. G. Arachchilage. Understanding user privacy expectations: A software developer's perspective. *Telematics Informatics*, 35(7):1845–1862, 2018.

[35] S. Shabanian, D. Shanmugam, F. Diaz, M. Finck, and A. Biega. Learning to limit data collection via scaling laws: Data minimization compliance in practice. arXiv, July 2021.

[36] D. Shanmugam, F. Diaz, S. Shabanian, M. Finck, and A. 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.

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

[38] J. Tolsdorf, D. Reinhardt, and L. L. Iacono. Employees' privacy perceptions: exploring the dimensionality and antecedents of personal data sensitivity and willingness to disclose. *Proceedings on Privacy Enhancing Technologies*, 2022(2):68–94, 2022.

[39] C. Utz, M. Degeling, S. Fahl, F. Schaub, and T. Holz. (un) informed consent: Studying gdpr consent notices in the field. In *Proceedings of the 2019 acm sigsac conference on computer and communications security*, pages 973–990, 2019.

[40] W. Van Oorschot. Failing selectivity: On the extent and causes of non-take-up of social security benefits. In *Empirical poverty research in a comparative perspective*, pages 101–132. Routledge, 2019.

[41] M. Veale and R. Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

[42] I. Wagner and D. Eckhoff. Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)*, 51(3):1–38, 2018.

[43] R. Zaman and M. Hassani. On enabling GDPR compliance in business processes through data-driven solutions. *SN Computer Science*, 1(4):1–15, 2020.

[44] S. Zuboff. *The age of surveillance capitalism: The fight for a human future at the new frontier of power: Barack Obama's books of 2019.* Profile books, 2019.