SoK: A Classification for AI-driven Personalized Privacy Assistants

Victor Morel
Chalmers University of Technology
and University of Gothenburg
Gothenburg, Sweden
morelv@chalmers.se

Leonardo Iwaya
Karlstad University
Karlstad, Sweden
Leonardo.Iwaya@kau.se

Simone Fischer-Hübner
Karlstad University, Chalmers
University of Technology, and
University of Gothenburg
Karlstad, Sweden
simone.fischer-huebner@kau.se

ABSTRACT

To help users make privacy-related decisions, personalized privacy assistants based on AI technology have been developed in recent years. These AI-driven Personalized Privacy Assistants (AI-driven PPAs) can reap significant benefits for users, who may otherwise struggle to make decisions regarding their personal data in environments saturated with privacy-related decision requests. However, no study systematically inquired about the features of these AI-driven PPAs, their underlying technologies, or the accuracy of their decisions. To fill this gap, we present a Systematization of Knowledge (SoK) to map the existing solutions found in the scientific literature. We screened 1697 unique research papers over the last decade (2013-2023), constructing a classification from 39 included papers. As a result, this SoK reviews several aspects of existing research on AI-driven PPAs in terms of types of publications, contributions, methodological quality, and other quantitative insights. Furthermore, we provide a comprehensive classification for AI-driven PPAs, delving into their architectural choices, system contexts, types of AI used, data sources, types of decisions, and control over decisions, among other facets. Based on our SoK, we further underline the research gaps and challenges and formulate recommendations for the design and development of AI-driven PPAs as well as avenues for future research.

KEYWORDS

privacy assistant, privacy, data protection, artificial intelligence, machine learning, systematic review

1 INTRODUCTION

As the world becomes increasingly digitalized, people are faced with a higher number of decisions related to their privacy. We surround ourselves daily with several apps and websites, and the number of smart gadgets and Internet of Things (IoT) devices continues to grow [1]. Furthermore, to enforce the individuals' rights to informational self-determination and comply with privacy laws such as the General Data Protection Regulation (GDPR) [23], software systems regularly require us to make privacy-related decisions regarding our personal data: *Do you grant this permission? Do you want to accept the cookies? Should this sensor be left on when you host friends?* Consequently, the cognitive burden increases, leaving users in disarray, tired, and unable to decide in their best interests [17].

During the last decade, researchers have been building privacy assistants to alleviate this burden and support users in their decisions (see the patent on Personalized Privacy Assistant registered in 2023 in the US by Sadeh et al. [54]). With the progress made in Artificial Intelligence (AI), it is no surprise that some of these assistants leverage this technology, notably enabling more personalized support. However, the extent to which AI drives these Personalized Privacy Assistants (AI-driven PPAs), their efficiency, privacy-friendliness, functioning, and eventual addressing of legal requirements remains unclear. In fact, to the best of our knowledge, there have been no surveys or systematic reviews on the topic of AI-driven PPAs. This lack of systematization of knowledge prevents other researchers from identifying existing gaps in the field and efficiently addressing the challenges.

To address this lack of coherence, provide a common vocabulary, and better compare and categorize the different AI-driven PPA solutions, we propose a Systematization of Knowledge (SoK) of the last decade of research. In doing so, we aim to draw insights and lessons for future assistants and to formulate better recommendations for research, design, and development of AI-driven PPAs. Formulated otherwise, we tackle the following Research Questions (RQs):

- **RQ1:** *What is the current state of the literature on AI-driven PPAs for automated support of end-users privacy decisions in IT systems?*
- **RQ2:** *What are the key attributes and properties of the proposed AI-driven PPAs in the literature?*

Here, we understand agents and assistants in a broad sense (any logical entity able to support users, including unimplemented theoretical models, see our selection criteria in Table 1); AI in a generic sense as well (see Section 2.3); and privacy decisions as individual decisions regarding one's personal information management (see Section 2.2).

To address our RQs, we performed a Systematic Literature Review (SLR) on research papers that provided technical solutions, published between 2013 and 2023 in peer-reviewed venues, and a further snowballing process until early May 2024. We screened 1697 unique papers from IEEE, ACM, Scopus, and Web of Science, resulting in 39 selected papers after several rounds of snowballing. We extensively read and analyzed all the included papers, and the information extracted forms the basis of our work.

As a result of our SLR, we make the following contributions:

Proceedings on Privacy Enhancing Technologies YYYY(X), 1–17

https://doi.org/XXXXXXX.XXXXXXX

A Classification for AI-driven PPAs - We propose the first classification for AI-driven PPAs, providing a common vocabulary for designers of such systems.
Data Charting & Quantification - We charted and quantified several aspects of AI-driven PPAs based on the aforementioned classification.
Research Gaps & Challenges - We underline the current gaps in the state of the art and highlight challenges for designing AI-driven PPAs based on our data.
Recommendations & Research Avenues - We formulate recommendations for improving AI-driven PPAs, and propose several avenues for future research.

2 BACKGROUND AND RELATED WORK

As a background, this section first introduces legal concepts around privacy, then provides an overview of different types of privacy decisions for which individuals could receive support from AI-driven PPAs. Lastly, it presents AI and Machine Learning technologies that can provide the technical foundations for AI-driven PPAs.

2.1 Legal Background

2.1.1 Roles and Obligations According to the GDPR and AI Act. AI-driven PPAs are using AI techniques for processing data, including personal data, to assist users with making personalized privacy-related decisions. For the discussion of any legal requirements regarding the use of personalized privacy assistants, the question of who the data controller is for any personal data processing by the assistants will be of relevance.

If AI-driven PPAs are installed and run by users on their own devices or other servers under their control, the users will likely act as data controllers or joint controllers with other service providers. The so-called household exemption (Art. 2(2)(c) GDPR) can take effect, meaning that the GDPR [23] will not apply if the user is using the assistant for private purposes on their private devices or servers under their control. If the AI-driven PPA is run not only for purely private purposes on the user's devices or controlled servers, the data controller may be another entity different from the user (e.g., the user's employer). Legal obligations need to be fulfilled by the controllers regarding data protection by design and default (Art. 25) of the assistants, security of data processed (Art. 32), implementing data subject rights, including the data subject's rights to transparency (Art. 13–15), their rights to object to profiling (Art. 21), and the right not to be subject to a decision based solely on automated processing, including profiling (Art. 22).

In addition, legal obligations according to the EU AI Act [24] may also have to be considered for the producer and also by the deployers of AI-driven privacy assistants, including requirements for risk management (Art. 9), transparency (Art. 13, 50), robustness, security, accuracy (Art. 15). These obligations however mostly apply if AI-driven PPAs could be classified as "high-risk" AI systems. This should, however, seldom be the case, especially as AI-driven PPAs are typically used for users' own privacy management, which should typically not interfere with the fundamental rights of others. Exceptions could, however, be AI-driven PPAs that are, for example, used for setting permissions for safety-critical applications impacting the safety of the users or others.

2.1.2 Legal Requirements for Transparency. In cases where the data controllers of the AI-driven PPAs are not the data subjects themselves, the controllers should provide the data subjects with privacy policy information *ex-ante* at the time when data are obtained from them according to Art. 13 GDPR, and *ex-post* through the right to access granted in Art. 15 GDPR. This also should include information about purposes of processing, data categories concerned, but also information about the logic involved and significance, and envisioned consequences of automated decision-making and profiling performed by the AI-driven PPAs.

The AI Act also includes obligations for transparency for the producers and deployers of limited-risk and high-risk AI systems (Art. 50). While the providers of limited-risk AI systems have to mainly ensure that humans are informed that AI systems are used, high-risk AI systems require that further clear, comprehensible and adequate information is given to the deployer (Art. 13), traceability of results via logging (Art. 12) and appropriate human oversight (Art. 14).

2.1.3 Legal Requirements for Consent. Art. 4 (11) of the GDPR defines 'Consent' of data subjects as any freely given, specific, informed, and unambiguous indication of the data subject's wishes by which they, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to them. A valid consent has thus to fulfill several conditions. Namely, it needs to be:

- *Freely given*, i.e., the data subject needs to have free choices – this is usually not the case if there is an imbalance of power in the relation between the data subject and the data controller. Furthermore, there should be no negative consequences if consent is not given. Moreover, consent may not be bundled as a non-negotiable part of terms and conditions.
- *Specific*, which means that consent must be given for one or more specific purposes and that the data subject must have a choice in relation to them – i.e., separate opt-in is needed for each purpose.
- *Informed*, which means that the data subject has to be informed about certain elements that are crucial to making a choice. This includes information about the controller's identity, the data processing purposes, the type of data, the right to withdraw consent, any use for decisions based solely on automated processing, and risks of data transfers to third countries.
- Moreover, a confirming statement or *affirmative action* is needed for a valid consent and requires that the data subject has taken a deliberate action to consent. Therefore, silence, pre-ticked boxes, or inactivity should not constitute consent (Recital 32 GDPR). It also means that consent cannot be fully automated.

2.2 Privacy Decisions

Among the most notable definitions, Westin [65] has defined privacy as the right to informational self-determination, meaning that individuals should have the *right to decide* for themselves when, how, and what information about them is communicated to others. As mentioned, in the EU, the GDPR emphasizes that individuals should have control of their personal data (Recital 7), and thus should be empowered to make decisions about their data as one prerequisite for exercising such control. Delving deeper into this notion of *privacy decisions*, we further elaborate on this concept in the following subsections.

2.2.1 Individual Privacy Decisions Regulated by Privacy Laws. Some privacy decisions individuals can make to exercise control over their data are regulated under the GDPR and other privacy laws. These decisions notably include, but are not limited to, the *decisions to grant or to withdraw consent* to data collection and processing.

Indeed, the GDPR and most other privacy laws regulate *decisions to exercise data subject rights* granted by the respective laws. For instance, according to Art. 15-22 GDPR, data subjects have the rights to access data, request rectification or deletion of data, export data, and object to direct marketing and profiling. Data subjects can also object in cases where the legal ground for the processing is public interest or legitimate interest, or exercise their right not to be subject to automated decision-making.

2.2.2 Further Types of Privacy Decisions. Further types of privacy decisions concerning users' choices regarding the use of their data by others, which are not directly mentioned or regulated by the GDPR, include *decisions of individuals to publish or share data on their own initiative*, e.g., in social networks. In these cases, data sharing has typically not been formally triggered by a consent request to allow data sharing with another party.

Moreover, privacy decisions encompass *privacy permission* (or access control rights) settings, which grant others certain rights for using their data and are, for instance, typically used for permission systems of mobile phone operating systems, such as Android or iOS. Setting privacy permissions on mobile operating systems often requires consent at installation or during runtime. However, instead of consent, other legal grounds – such as a contract (Art. 6 (1)(b) GDPR) –, can be used, e.g., for a banking app to forward account information when transferring money [7]. Let us also note the peculiar case of Global Privacy Control (GPC), a unary signal that permits or prohibits third-party tracking on the browser [31]. Due to its enforceability under the California Consumer Privacy Act (CCPA), it is regulated by a privacy law but is technically more akin to a privacy permission.

Additionally, some privacy-enhancing technologies and protocols allow users to decide and set *privacy preferences*, which are simply indications of the users' privacy wishes of how their data should be used without actually granting any rights to others, and thus without legal mandate. Privacy preferences have, for instance, been used earlier by the Platform for Privacy Preferences (P3P) [18] or Do Not Track (DNT) as an example for signals that can be set manually in browser settings for allowing users to specify their privacy choices.

2.3 AI for Decision-making

AI is a generic term for various strategies and techniques enabling computers and machines to simulate human intelligence and problem-solving capabilities [53]. Machine learning (ML) is a field of AI (we subsume the former under the latter in the rest of the document) that develops and studies statistical algorithms and models, draws inferences from patterns in data, and learns and adapts without following explicit instructions. AI-powered tools can particularly lighten the user's cognitive load and thereby improve their decision-making, e.g., by decision support, augmentation, or automation.

While there are different ways to categorize AI systems, we refer in the present work to the survey paper on eXplainable AI (XAI) by Arrieta et al. [6]. They distinguish between transparent models and those requiring post-hoc explainability (non-inherently transparent). We leverage this reference because AI-supported decision-making must be explained under specific circumstances according to the GDPR and the AI Act [51].

In their words: "A model is considered to be transparent if by itself it is understandable." [6] Such models include linear regression, decision trees, k-nearest neighbors, rule-based learning, general additive models, and Bayesian models. Nonetheless, models that are not deemed intrinsically transparent can be made explainable through the use of *post-hoc* techniques. Neural networks (especially deep and convoluted) and Support Vector Machines (SVM) typically fall under this category, as well as reinforcement learning [52].

3 METHODOLOGY

This SoK study adopts the widely known methodology for systematic literature reviews (SLRs) proposed by Kitchenham [36]. The SLR methodology offers us a well-defined and rigorous sequence of methodological steps consisting of three main phases: (1) planning, (2) conducting, and (3) reporting the review. A SLR Protocol that describes the entire research process has been written for this study (a summary version of which can be found in Appendix A.1). Furthermore, we make our research data openly available in an anonymised GitHub repository [1] for reproducibility. Our material comprises the citation files of each query, the Data Extraction Forms (DEFs) of the selected papers, and the charting spreadsheet used to compile all our data. Thus, due to page constraints, we refer readers to these documents for methodological details.

3.1 Planning the Review

During the planning phase, our first activity was determining the need for this SLR. Several databases were searched to verify if any surveys or reviews had been conducted on AI-driven PPAs. Search terms such as privacy, data protection, assistant, agent, artificial intelligence, and machine learning were used. However, we could not identify any survey or systematic reviews on the topic, reassuring the need for an SLR.

The research questions, presented in Section 1, guided the remaining phases of this SLR with respect to the search process, selection criteria, and data synthesis.

3.2 Conducting the Review

3.2.1 Search Strategy. Based on our RQs and previous preliminary searches when designing the SLR Protocol, we identified a list of nine relevant keywords, i.e., *privacy, data protection, assistant, agent, artificial intelligence, machine learning, intelligent, automatic*, and *personalized.* These keywords were used to construct the following search query:

[1] https://anonymous.4open.science/r/SoK_AI_PPA-E29F

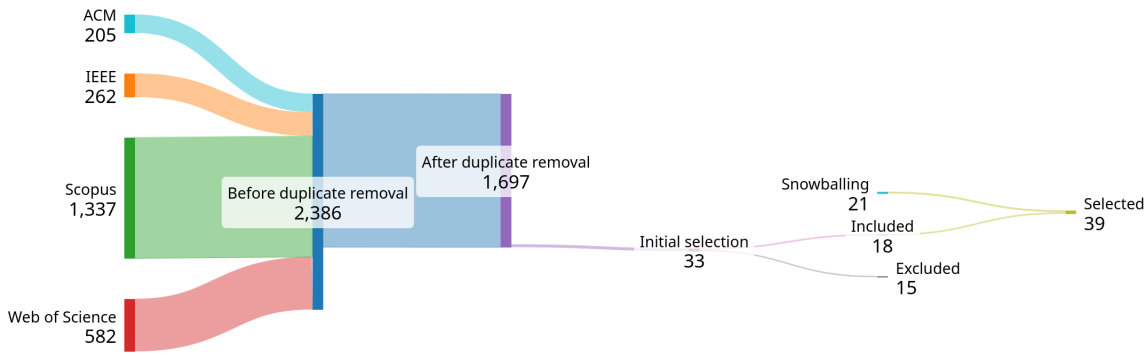

Figure 1: Sankey chart of the selection process.

```
(privacy OR "data protection") AND (assistant* OR agent*)
AND ("artificial intelligence" OR "machine*learning" OR
intelligent OR automat* OR personali*ed)
```

As such, the search query targets papers working on three joint topics: 1) privacy (or data protection), using either 2) an assistant or an agent, and leveraging 3) artificial intelligence or personalization.

Four scientific databases were selected, i.e., Scopus, Web of Science, IEEE Xplore, and ACM Digital Library, due to their high relevance to the areas of computer science and engineering, comprising the vast majority of published research in the field. We also specified inclusion and exclusion criteria (see Table 1) used during the screening of publications retrieved from the databases. Before starting the search process, two authors piloted the searches on all databases and ran a *calibration exercise* to verify the consistency of the inclusion criteria. For that, the authors independently screened 10% of the results and discussed their decisions. The conflicts were all discussed and solved, sometimes with the help of a third author. This process was repeated a second time, screening another 10% of the papers at a point that the authors fully agreed with the consistency of the selection process.

Inclusion Criteria
- Provides a technical solution (implemented or theoretical) to help end-users automate personal (and personalized) privacy decisions with an assistant (or artificial agent) in IT systems. - Papers from 2013 onward to concentrate on the state-of-the-art. - The concept of AI needs to be explicitly stated in the papers.

Exclusion Criteria
- Papers with solutions that are purely theoretical without substantial explanations on how they could be implemented in practice. - Papers with solutions that solely automate the analysis of privacy policies but without any type of personalization. - Papers with poor scientific quality (e.g., lack objectives or research questions, the methodology is not described, the solution is insufficiently/vaguely described, etc.).

Table 1: Criteria for the inclusion and exclusion of studies.

3.2.2 Selection Process. Figure 1 presents an overview of the selection process. The querying of the databases mentioned above on October 19, 2023, yielded 2386 papers and 1697 unique entries after removing duplicates. The screening phase lasted until November 23, 2023, and resulted in the selection of 33 papers. Two authors then read 10% of these 33 papers (3) and adjusted the DEF based on mutual feedback. This step helped us add new important fields to the DEF and consistently extract data from the papers.

The first data extraction phase, consisting of a full reading of each of the 33 papers, was performed over weeks 4 to 7 (included) in 2024. Fifteen papers were excluded after full reading for different reasons: they were duplicates (i.e., same work published in different venues); they did not provide any technical solution; the automated decisions were not personalized to an end user; AI was not used for automating decisions; or they are of poor scientific quality (see our criteria in Table 1); one paper was not available for download, we could not access it even after reaching out the authors.

We then proceeded to several snowballing phases [68], during which we checked the abstracts of all seemingly relevant[2] papers cited (backward snowballing), and screened citing papers (forward snowballing). The snowballing process lasted from week 8 to week 19 of 2024 and resulted in 21 additional papers after exclusion, for a total of 39 papers (33-15+21).

3.2.3 Data Extraction and Analysis. The data extracted in the DEFs was compiled and further organized in spreadsheets during weeks 20-21. This process also included the initial aggregation of data and the creation of frequency charts across several data categories (e.g., studies per year, types of publications, authors and affiliations, etc.)

Although we attempted to extract as much data from the studies as possible using a DEF, we found during the data analysis process the need to further categorize studies across other *facets*. For instance, additional information was compiled in the spreadsheets, such as a high-level categorization of certain fields (i.e., the type of AI used) or a critical appraisal of the user studies presented in the selected papers. This collection of facets created during the study design and data analysis processes forms the basis of the work's final classification scheme, which is presented as part of the main results. All authors were involved in the data analysis process and the definition of facets that further classify studies on the topic.

[2]We only assessed papers cited in relevant sections such as the related work.

By definition, an AI-driven PPA leverages AI techniques. We therefore collected information about the *Type of AI used*. AI models rely on data for training and decision-making. As such, we extracted the *source of data*. During the adjustment of the DEF, we observed that AI-driven PPAs are usually designed for a specific *system context*, and for one or several *types of decision*. Connected to the system context, we extracted the *choice of architecture* of the implementation (if any) to analyze the trust implications. We also collected the methods for an *empirical assessment*, presence and quality of user studies, or the means used to measure the accuracy, to gain insights on eventual benchmarks of AI-driven PPAs. Studies can be classified as evaluation or validation research, as proposed by Wieringa et al. [66]. An evaluation works in real-world practice and is implementing/deploying the solution or testing in an actual project with real test users, such as real case studies and realistic user testing of prototypes/systems. A validation is a limited illustrative or hypothetical "case study" or "use case" performed as a lab experiment. In general practice, prototypes are often validated by cross-sectional studies. Finally, initially guided by legal requirements for consent and the exercise of data subject rights under the GDPR (eventually, no paper considers consent), we extracted what became *user control over decisions*.

3.3 Reporting the Review

Based on the data analysis, a whole coherent narrative was written by the research team, i.e., the SoK conveying all the results, our interpretation of the main findings, and identifying research gaps. This synthesis of this SLR on AI-driven PPAs is thus reported in the following sections of this paper.

4 SUMMARY OF DATA CHARTING RESULTS

This section provides a brief overview of the quantitative insights generated through the data charting process (e.g., publications per year, citations, types of decisions). Due to page constraints, further descriptions of the data charting results are provided in the appendices. Most classification features are reported in Table 2, except for the *Empirical assessment* which, along with the results of our critical appraisal and the type of contribution, are reported in Table 6 due to space constraints.

Among the 39 papers surveyed, we tallied 15 different countries for the authors' affiliations (see Table 4 in Appendix A.2), with the USA and UK leading in numbers. About 54% ($n = 21$) of the selected publications were published from 2019 to 2023, with the year 2021 being the most productive with 8 publications (see Table 5 in the Appendix A.2). At the time of writing, papers were cited between 0 and 275 times with an average of 38.56 citations, a median of 14, and a standard deviation of 59, indicating a power law distribution of the citation count. The most cited papers are Liu et al. [41] ($n = 275$), Yu et al. [70] ($n = 199$), and Squicciarini et al. [61] ($n = 130$).

Regarding the sources of data used by the AI-driven PPAs, context data, attitudinal data, and metadata were the most prevalent. We observed a relatively balanced distribution when it comes to the types of decisions (between 12 and 15 for each type) and the system contexts (between 11 and 13, with one outlier for *Cloud*). For the types of AI systems that we were able to classify, most models were deemed non-intrinsically transparent (NIT, $n = 13$), followed by transparent (T, $n = 6$) and partially transparent (PT, $n = 4$) models. Note also that Das et al. [21] did not specify the type of AI used in their paper, we were therefore unable to categorize their solution in that respect (under *Type of AI used*).

The publications were also classified by their types of contributions, according to the categories proposed by Kuhrmann et al. [39] and Shaw [59] (see Appendix A.1). We observed a prevalence of models ($n = 23$) and tools ($n = 19$), followed by frameworks ($n = 9$), as shown in Table 6.

Lastly, we also performed a critical appraisal on the user studies presented, although only when those user studies were used to evaluate the AI-driven PPA, and not when they were used for data collection for dataset building. We used the CAT (Critically Appraised Topic) Manager App of CEBMA (the Center for Evidence-Based Management) [16], which provides a practical yet rigorous approach to evaluate studies based on objective criteria. The result of our critical appraisal can be found in Table 6 in the Appendix A.2. Out of the 39 publications, only 16 presented a user study, and in terms of quality, they mostly scored "low" or "very low" ($n = 12$) according to the CEBMA checklist. Exceptionally, only the studies of Liu et al. [41] and Baarslag et al. [9] were appraised as of high quality.

5 CLASSIFICATION FOR AI-DRIVEN PPAS

We provide in this section a classification for AI-driven PPAs as the main contribution of this SoK. The classification comprises several dimensions, i.e., features typically considered in the design of such an assistant. These dimensions are the *type of decision* (Section 5.1), the *type of AI* (Section 5.2) and the *source of data* (Section 5.3) used in the decision, the *system context* (Section 5.4), the *choice architecture* of its eventual implementation (Section 5.5), the *empirical assessment* (Section 5.6), and the extent to which *users have control over the decisions* (Section 5.7).

The classification and its dimensions are **data-driven** in the sense that they were derived based on what is described in the papers, reflecting the current state of the literature. For example, considering the category of system contexts, more dimensions could be envisioned, but we limited it to the four dimensions (i.e., mobile apps, social media, IoT, and cloud) that were found in the papers. Each feature will be explored in more detail in this section, and substantiated with non-exhaustive examples for each possible option, while an overview is provided in Figure 2.

Note that not all dimensions are necessary for composing an AI-driven PPA. The dimensions for the type of AI, source of data, type of decision, and system context are *"mandatory,"* consisting of essential requisites that an AI-driven PPA needs to consider (solid boxes in Figure 2). Other dimensions such as the empirical assessment, choice architecture, and user control over decisions are *"optional"* since not all the identified AI-driven PPAs were evaluated, some do not have an implementation (and therefore an architecture), and some (regrettably) do not empower users with much control for various reasons (dashed boxes in Figure 2).

Furthermore, the papers address each of these dimensions to different extents, and their options are often non-exclusive. For instance, all articles surveyed discuss the type of decision, but this is not the case for the choice architecture; and while most solutions

		Type of decision			Type of AI used					Type of source of data						System context				Architecture			User control over decision			
Year	Publication	Permissions	Preferences	Data sharing	Classification	Clustering	Rule-based	Logic-based	Reinforcement	Context	Attitudinal data	Metadata	Type of data	Content of data	Behavioral data	Mobile apps	Social media	IoT	Cloud	Local	Remote	Federated	Informed	Semi-automated	Specific	Revoke
2014	Xie et al. [69]		•		• (NIT)					•										–	–	–	No	Yes	Yes [3]	No
2015	Apolinarski et al. [5]	•			• (NIT)					•						•				•			D	Yes	Yes	No
2015	Hirschprung et al. [29]	•				•					•								•	–	–	–	D	No [4]	Yes	No
2015	Squicciarini et al. [61]			•			•			•		•		•						–	–	–	D	Yes	Yes	No
2016	Liu et al. [41]	•			• (T)	•					•					•				?	?		D, P	Yes	Yes	Yes
2016	Albertini et al. [2]			•			•				•						•				•		D	Yes	No	No
2016	Dong et al. [22]			•	• (T)					•				•			•			–	–	–	–	–	–	–
2017	Baarslag et al. [9]	•						•			•		•			•				•	•		Unclear	Yes	Yes	No
2017	Fogues and Murukannaiah [27]			•	• (PT)					•	•						•				•		No	No	No	No
2017	Zhong et al. [72]			•	• (NIT)						•		•				•			–	–	–	–	–	–	–
2017	Misra and Such [44]			•	• (NIT)					•				•			•				•		D	Yes	Yes	No
2017	Nakamura et al. [46]		•		• (NIT)								•		•					–	–	–	–	–	–	–
2017	Olejnik et al. [49]	•			• (T)					•	•					•				•			No	Yes	Yes	No
2018	Das et al. [21]		•								•							•			•		Yes	It depends	Yes	No
2018	Tan et al. [62]	•			• (T)							•				•					•		No	No [5]	Yes	No
2018	Wijesekera et al. [67]	•			• (NIT)					•		•	•			•				•	•		D, C	Yes	Yes	Yes
2018	Yu et al. [70]			•	• (NIT)					•				•			•			–	–	–	–	–	–	–
2018	Bahirat et al. [10]		•		• (T)						•							•		–	–	–	D, P [6]	It depends	It depends	No
2019	Klingensmith et al. [37]	•			• (NIT)							•			•			•		•	•		D	Not always	Yes	No
2019	Barbosa et al. [11]		•		• (PT)						•	•						•		–	–	–	–	–	–	–
2019	Alom et al. [4]		•		• (PT)					•	•									–	–	–	–	–	–	–
2019	Alom et al. [3]		•		• (NIT)						•				•			•		–	–	–	–	–	–	–
2020	Kasaraneni and Thomas [33]			•	• (T)	•						•					•				•		D	Yes	Yes	No
2020	Kaur et al. [34]		•		• (NIT)					•		•				•		•		–	–	–	–	–	–	–
2020	Botti-Cebriá et al. [14]			•	• (PT)									•			•				•		D	Yes	Yes	No
2020	Kokciyan and Yolum [38]	•					•			•		•						•		–	–	–	–	–	–	–
2020	Sanchez et al. [55]	•				•					•							•		–	–	–	Unclear	Yes	Yes	No
2021	Kaur et al. [35]	•							•	•		•				•		•		–	–	–	–	–	–	–
2021	Lobner et al. [42]		•		• (T)							•	•		•		•			–	–	–	–	–	–	–
2022	Filipczuk et al. [26]	•						•			•		•			•				•	•		D	Yes	Yes	No
2022	Hirschprung and Alkoby [30]			•				•		•							•			–	–	–	–	–	–	–
2022	Kökciyan and Yolum [40]		•					•		•					•			•		–	–	–	No	It depends	Yes	No
2022	Ulusoy and Yolum [63]			•					•	•					•		•			–	–	–	–	–	–	–
2022	Zhan et al. [71]	•						•	•		•							•		–	–	–	–	–	–	–
2022	Brandão et al. [15]		•		• (NIT)	•				•						•						•	–	–	–	–
2022	Mendes et al. [43]	•			• (NIT)					•						•					•		No	No	No	No
2022	Shanmugarasa et al. [58]	•				•				•	•	•	•					•		•			No	Yes	Yes	No
2023	Ayci et al. [8]			•	• (NIT)						•	•					•				•		No	Yes	Yes	No
2023	Serramia et al. [57]		•				•				•	•						•			•		No	Yes	No	No

[3] Only location

[4] Not necessarily, depends on what they call the Configuration Options

[5] Not by default, they have a sort of 'user settings' for expert users

[6] Not consistently

Table 2: Summary table of our classification. For user control over decisions, we specify the elements present to inform users under *Informed*. An empty field signifies that the solution does not exhibit the characteristic (e.g., does not consider Y type of decision). Under Architecture, we denote with "–" when the criterion is not applicable (no implementation/tool is presented) and when the solution presents an implementation, but the paper did not specify enough information to infer its architecture. For the type of AI used, we specified whether the classification model is Transparent (T), Not-Inherently Transparent (NIT), or Partially Transparent (PT) because several models are used.

are composed of different sources of data and combine different AI models, the system context is often exclusive in the sense that solutions are often designed for a specific system context.

5.1 Type of Decision

Decisions taken by an AI-driven PPA can be of different types, and it is essential to distinguish them to assess the possibilities they offer. Indeed, some decisions – such as permissions – have a binding character, i.e., constraining the system to act according to the user's choice, while others do not, such as preferences. Note that it may not always be possible to distinguish between each type of decision clearly (as discussed in Section 2.2.2). Other types of decisions with different implications regarding their enforcement can be envisioned by an AI-driven PPA (such as consent or deletion requests, see Section 2.2).

5.1.1 Permission Settings. The first type of decisions that many AI-driven PPAs assist the users with is *permission settings*, which, as discussed in Section 2.2.2, correspond to access control settings. Permissions are system-specific and binding, as the underlying operating system should enforce them.

We typically find mobile app permissions (e.g., in Baarslag et al. [9], mobile apps are addressed in 11 papers), but they are not restricted to the mobile environment. AI-driven PPAs can deal with permissions in IoT environments (see, e.g., [37], IoT is covered by 13 papers) or in the cloud [29].

5.1.2 Preference Settings. The second type of decision covered by the literature is *preference settings*, which, unlike permission settings, should be understood as expressions of will. Several works refer to preferences while they actually deal with permissions [26, 29, 41, 58, 67]. It is indeed common to talk about preferences imprecisely, but they should not be confused with permissions that have a binding property.

5.1.3 Data Sharing. Data sharing is the third type of privacy decision of AI-driven PPAs encountered in the reviewed literature, for which the binding character is uncertain for users (for instance, assessing whether a limitation in the audience is enforced is not always possible from a user point of view, because the underlying

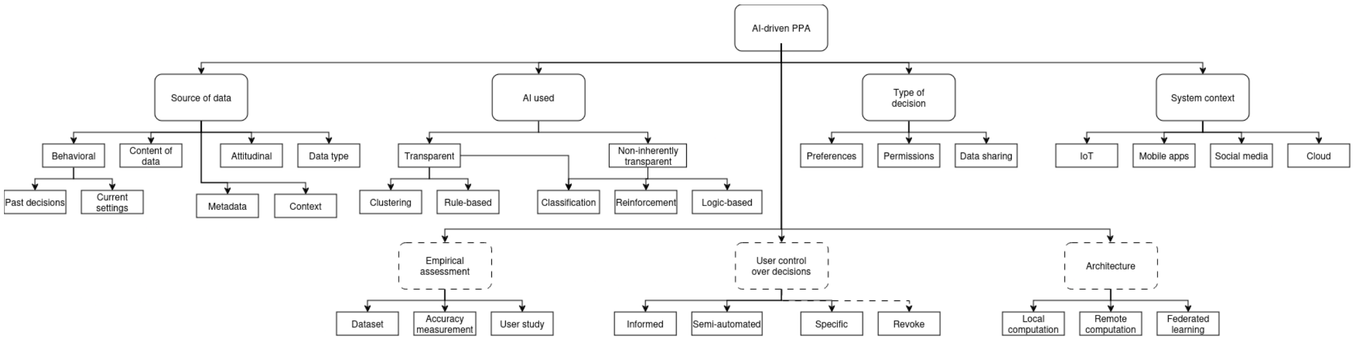

Figure 2: A schematic representation of the classification presented in Section 5. Each facet is represented as a rounded box, solid for the mandatory features and dashed for the optional ones. For User control over decisions (see Section 5.7), we distinguish between qualities of control (solid arrows) and instruments of control (dashed arrows).

technical system is inaccessible to them, see for instance Ulusoy and Yolum [63]). Typically, it can be difficult or even impossible to assess whether most social media platforms strictly account for the user's privacy decisions, or merely welcome them as recommendations to be applied only if possible. Papers classified under this type of decision usually do not mention the binding character of their solution (or the lack thereof).

5.2 AI Technology Used

Another significant characteristic of AI-driven PPAs is the type of AI used. Many solutions are based on machine learning models, such as supervised ML (classification), non-supervised ML (clustering), and reinforcement learning, sometimes combined. It is, however, also possible to find older AI techniques grouped under the umbrella of expert or rule-based systems.

We also classified the different AI technologies used by the AI-driven PPAs reviewed regarding their explainability, or their inherent transparency. However, XAI is only explicitly addressed by one work [42]; the other models are therefore categorized based on Arrieta et al. [6]'s taxonomy. We annotated T for Transparent in Table 2, NIT for Not-Inherently Transparent, and PT for Partially Transparent when the solution relies on models with different levels of transparency.

5.2.1 *Transparent.*

Classification. Supervised machine learning, also called classification models, is a common set of techniques deployed in AI-driven PPAs. In this context, a model is trained to classify an object of decision into a choice tailored to the users' desires.

Transparent classification models [6] (used in 7 papers) are composed of decision trees (used for instance in Bahirat et al. [10]), k-nearest neighbors (leveraged in Botti-Cebriá et al. [14]), and Bayesian models (see Olejnik et al. [49]).

Clustering. Several works use clustering techniques for their AI-driven PPA. In this context, clustering is classically used to create a set of *privacy profiles*, i.e., an archetypal ensemble of default parameters (for preferences or permissions) to which a user is then assigned. Clustering algorithms (leveraged in 6 papers) used are hierarchical clustering (Liu et al. [41]), k-means (Brandão et al. [15]), k-modes (Shanmugarasa et al. [58]), although several papers did not disclose the exact method used (Hirschprung et al. [29] for instance).

Rule-based. AI-driven PPAs can be powered by non-machine-learning algorithms, based instead on rules (Albertini et al. [2] implement association rules). This comprises theoretical as well as practical works, with two of them out of four providing a tool (Albertini et al. [2] and Serramia et al. [57]).

5.2.2 *Not-inherently Transparent.*

Classification. Non-transparent classification models (found in 13 papers) typically encompass classic neural networks (as in Klingensmith et al. [37]) and deep neural networks (see for instance Yu et al. [70]); random forests ([44]), Ada Boost [43] and Support Vector Machines (used in Wijesekera et al. [67]) complete the picture. Post-hoc explanations must complement these models, as they are not easily understandable by themselves.

Reinforcement. Reinforcement learning is the least used family of machine-learning techniques in AI-driven PPAs. It is implemented in Kaur et al. [35] and Ulusoy and Yolum [63], both used to adapt users' feedback to their preferences, and in Zhan et al. [71]. The first paper uses it to disclose information (using permissions), while the second uses it to learn bidding preferences in a negotiation context.

Logic-based. AI-driven PPAs can be based on logic (5 papers), for instance, expert systems (Kökciyan and Yolum [40] uses an agent-based model) or game theory (such as Hirschprung and Alkoby [30]). These works, albeit few, span various system contexts and types of decisions.

5.3 Source of Data

An AI-driven PPA can rely on various *sources of data* when using AI to help with a privacy decision. These data sources are very often combined, and a careful choice is necessary to fully exploit the potential of the models described in the previous section.

5.3.1 *Context.*

Context is an often-used data source, yet not always well-defined. However, when it is defined, it is composed of the location [69], the time, relationships with other individuals [27], or the activity performed [4].

External data provided by third parties or other unrelated entities is sometimes used to predict privacy decisions, and this external data can arguably be considered context. For instance, under this term, we find risk factors [8] or information related to other applications in the background [15].

Context is usually a crucial component for an effective AI-driven PPA because, as has been argued under theory of privacy as contextual integrity [48], context is paramount to design appropriate information flows and to respect privacy norms.

5.3.2 Attitudinal Data. A few AI-driven PPAs ask users questions to elicit so-called *attitudinal data* about stated practices or preferences regarding privacy recommendations to avoid the so-called cold-start problem, which arises when no past data is available to provide a recommendation. For example, Nakamura et al. [46] focuses on asking a minimal set of questions while keeping accuracy as high as possible, or Alom et al. [3] asks "a reasonable number of questions (50) to the users."

5.3.3 Behavioral Data. Another common source of data is *behavioral data*. Behavioral data has the advantage of reflecting the *actual* privacy decisions of users to predict the next ones, as it does not simply rely on stated practices (unlike attitudinal data). While it can be a powerful tool, it can also create a feedback loop, reinforcing the same decisions.

Behavioral data can encompass past decisions, such as in Zhan et al. [71], which leverage past choices to fill a knowledge base, then used them to predict privacy decisions. It can also comprise current settings or preferences on a specific type of data to infer a decision for another type [29]. The system can also use these preferences to match users to a particular privacy profile, such as using clustering techniques (see Section 5.2.1).

5.3.4 Metadata. Metadata is data that provides information about other data, for example, the name of an application used [67], network requests [62], the purpose associated with processing [11], the usage frequency of certain permissions (such as location) by an app [34], or tags associated with images [61]. To some extent, metadata can overlap with context, for instance, when considering time or location. However, the articles surveyed more often refer to the time and location of collection *of a certain data point* for metadata, and to the *current time and location* when a decision has to be made for context. Metadata can provide peripheral information to make decisions, although it is rarely used as a sole source of data (only 3 papers out of 13 [33, 37, 62] rely only on metadata).

5.3.5 Data Type. The *data type* refers to the category of data concerned by the decision, such as whether it is an image to share on social media [72], the location requested by an app [26], or various sensor data by an IoT device [58]. The type of data can provide accurate information about the sensitiveness of a decision (location data can, for instance, provide sensitive information regarding the users' context, e.g., from location data that reveals that a user visits a clinic or church, medical, or religious information could be inferred), yet only a relatively low number of solutions rely on the data type to build an AI-driven PPA [9, 26, 42, 46, 58, 67, 72].

5.3.6 Content of Data. The *content of data* refers to the specific content of a data point, as the name indicates. However, we also include data that can be directly inferred from the content of data under this category. For example, Botti-Cebriá et al. [14] and Dong et al. [22] estimate the sensitivity of the content of the information to be shared to help make a decision. Indeed, content can be leveraged to tailor decisions: a picture deemed private should not receive the same treatment as one deemed public, and a geolocation trace that may providentially allow inferring religious practice should be dealt with cautiously.

5.4 System Context

Most AI-driven PPAs target a specific *system context*, that is, a set of technologies with distinct characteristics. Indeed, each system context has specific requirements that one must consider when designing an AI-driven PPA. System contexts differ by the availability of an **interface**, **computational power**, and control over the **architecture**.

5.4.1 Mobile Apps. Several works focus on mobile applications, and often on Android [5, 67]. Mobile ecosystems have the advantage of being well-defined ecosystems, enabling the possibility to strictly enforce privacy decisions (i.e., it is often addressed with permissions, see Section 5.1.1).

Mobile phones also possess reasonable computational power (in the sense that they can run an AI-driven PPA) and a screen enabling direct user interactions. Hence, an AI-driven PPA can be implemented directly on a smartphone (see Baarslag et al. [9]), and it can interact with and even regulate mobile apps, all of which make mobile ecosystems suitable candidates for AI-driven PPAs under the users' control.

5.4.2 IoT. Another widely used system context for AI-driven PPAs is the Internet of Things (IoT). We understand IoT as a network of devices, including sensors, mechanical and digital machines, as well as consumer devices, all connected to the Internet. In practice, AI-driven PPAs have been developed for smart homes [11, 58], on campuses [21], or for wearables such as fitness devices [55] for instance.

Most IoT devices are usually not equipped with proper interfaces and lack computational power. These characteristics make it challenging to build AI-driven PPAs assisting with permission settings, yet not impossible (see Klingensmith et al. [37] for instance, who manage to do so with an AI-driven PPA located on end devices).

5.4.3 Social Media. According to our classification of the literature, the third major system context is social media, for which several AI-driven PPAs have been designed to help make privacy decisions. In this case, neither the interface nor the computational power are usually limiting factors. However, the design and implementation of social media platforms (that are usually not published openly) make it difficult to assess the binding character of privacy decisions supported by AI-driven PPAs running on social media platforms. AI-driven PPA solutions are rather designed to support data sharing, i.e., whether a specific post should be shared on social media and with whom, than focusing on assisting users with privacy decision-making.

5.4.4 *Cloud.*

Another system context is cloud environments, even though only one of the reviewed articles proposes an AI-driven PPA for the cloud [29]. Their solution offers a method to simplify information disclosure in cloud environments such as Google Drive. However, this work is thus a lone example and contrasts with the otherwise balanced distribution of works among other system contexts.

5.5 Architecture

By architecture, we refer here to where the computation happens, i.e., the decision-making, and not necessarily the pre-processing steps such as building privacy profiles. Directly connected to the architecture is the trust model of the AI-driven PPA. While this term is usually reserved for security-oriented research, describing whether one has to trust the different entities or not provides relevant information for understanding the privacy boundaries.

Note that the location of the computation is only relevant for implemented AI-driven PPAs, and not for theoretical models. Similarly, most solutions surveyed do not explicitly describe a threat or trust model in their paper. Nonetheless, it is possible to infer that trusted parties are required in some solutions. For instance, Tan et al. [62] describes an architecture comprising a remote classifier, in which one has to place trust, yet no trust model is described.

5.5.1 *Local Computation.*

The processing can happen locally on the user device, such as on a smartphone (see, e.g., [49]), but this device can also be a home pod in an IoT context (see, e.g., [58]).

Creating and processing user profiles, using local AI models, and locally deriving privacy decisions have the advantage that the user can keep control over the locally processed data, including their profiles and AI models, which usually can include sensitive information about the user's preferences or behavior. However, local data processing also puts more responsibilities on the user to secure the devices properly against malware or other attacks.

5.5.2 *Remote Computation.*

The AI-driven PPA could also be based on remote (according to the user's point of view) data processing, involving a central server that processes personal privacy decisions, contextual data including, e.g., location data or another type of data. Remote computation raises the question of the trust placed in the party performing this computation to protect the data properly, to enforce the data subject's rights (e.g., to access or to delete their data and computed profiles or models), and not to use the data for any unintended purposes [62].

Several solutions rely on a remote third party that has to be trusted, e.g., Baarslag et al. [9] or Tan et al. [62]'s solution that places trust on their own remote classifier. In contrast, others only require trusting the operating system (OS) on which the AI-driven PPA is implemented [49], or require trusting both the OS and mobile applications [5].

5.5.3 *Federated Learning.*

Only one article, by Brandão et al. [15], presented an AI-driven PPA based on federated learning. In this work, the processing of user data for the computation of locally trained neural network models happens on the user devices that share only the neural network weights with a central server, which will, in turn, average all the local weights and send back the results to the clients, which can use these new weights to continue the training process. Federated learning is a privacy-enhancing approach for processing the users' raw data only locally, which can achieve a performance comparable to the centralized approach (remote computation). Nonetheless, federated learning could still be attacked, e.g., with membership inference attacks for leaking personal data from locally trained models [60].

5.6 Empirical Assessment

AI-driven PPAs' performance can be measured in terms of accuracy, but because several solutions are meant to be usable tools, assessing an AI-driven PPA encompasses more than a mere measurement of how well a privacy decision is predicted.

As mentioned in Section 3.2.3, an empirical assessment can be an evaluation (see e.g. [41]) or a validation (e.g., [30]).

5.6.1 *User Study.*

A classical way to validate a tool or a method is to conduct a user study, and we found 16 papers reporting a user study to validate usability. A user study can have various interpretations, ranging from a simple questionnaire to rate satisfaction (such as Alom et al. [4]) to a large-scale randomized controlled study (see for instance Liu et al. [41]) – the former being more akin to a mere validation, the latter a full-fledged evaluation.

Note that several works elicited data to build a dataset through a user study, which was therefore not meant as a means of assessment (annotated as α in Table 6).

5.6.2 *Purely Statistical (Dataset).*

Several works provide a validation without a user study, that is, only based on a purely statistical analysis based on a dataset (such as Kasaraneni and Thomas [33], Botti-Cebriá et al. [14], Bahirat et al. [10], Kokciyan and Yolum [38], Ayci et al. [8], Zhan et al. [71], Kökciyan and Yolum [40], Botti-Cebriá et al. [14]). Such a measure, although potentially subject to a higher degree of statistical rigor, cannot necessarily capture users' expectations and may even fall into the pitfall of Goodhart's law.[7]

5.6.3 *Accuracy Measurement.*

Accuracy can measure the capacity of an AI-driven PPA to predict a privacy decision, but not all papers measure the same type of accuracy. Tan et al. [62] measure a privacy leakage detection, Botti-Cebriá et al. [14] whether the correct category of data is predicted or not, Serramia et al. [57] the acceptability rate, Barbosa et al. [11] the Area Under the Curve (AUC) of a binary allow/deny for a given scenario, etc.

Other works, while they do measure the accuracy of their solution to predict a privacy decision, present their work with limited rigor or precision. For example, Fogues and Murukannaiah [27] only present their results in plots. In contrast, others, such as Olejnik et al. [49], dedicate an entire subsection to explaining accuracy measurements.

5.7 User Control Over Decisions

Finally, AI-driven PPAs should not only assist users with making privacy decisions, but should at the same time also empower users with various options to improve *control over their decisions.* These options span over **qualities** of control (solid arrows in Figure 2) and **instruments** of control (dashed arrows). The former denotes adjectives that can be appended to control over a decision (akin to

[7] According to which "When a measure becomes a target, it ceases to be a good measure."

non-functional requirements in software engineering [50]), and the latter denotes concrete possibilities or actions for users (similar to functional requirements).

These options are partly related to GDPR requirements for consent (introduced in Section 2.1), which are thus relevant for privacy decisions that constitute consent. Note, however, that only a handful of papers specifically refer to legal considerations. Filipczuk et al. [26] refer to the GDPR, Mendes et al. [43] acknowledge that an automated response to a permission request might not constitute legal consent, Lobner et al. [42] base the rationale of explainability on legal requirements, and Sanchez et al. [55] even claim GDPR compliance. Nonetheless, decisions for setting permissions for mobile operating systems, for instance, still require consent at installation or run time. Thus, legal requirements for consent remain relevant for these types of decisions.

5.7.1 Ex-ante Transparency. Under Art 13 GDPR, data subjects should receive information if data are collected from them, and informing users is also an integral part of the dominant transparency paradigm in the US (the *notice* of the notice and choice approach). Informing data subjects with intelligible notices arguably improves their control over decisions. Several AI-driven PPAs only inform about the type of data concerned by the privacy decision [2], some inform in addition about the controller [66] or of the purpose of processing [10]. Meeting this criterion should not be difficult in theory, although providing intelligible notices requires significant expertise in practice (as illustrated in [56]).

5.7.2 Semi-automated. The semi-automated character of a decision refers to the inclusion of an affirmative action of the user to confirm the decision, which is therefore not fully automated [45]. Most solutions provide a semi-automated decision process, although not systematically (e.g., Das et al. [21] mention that only opt-out is possible for facial recognition), or not always (e.g., Klingensmith et al. [37] offers different types of "privacy profiles", one of which – *Laissez-Faire* – enables full automation). Tan et al. [62] do not leave users in the loop by default, but the system allows the possibility to change the settings for "experienced users," while it depends on the Configuration Option for Hirschprung et al. [29].

5.7.3 Specific. The specificity of a decision refers to the presentation and the possibility for users to decide on the granularity of each data type, purpose and controller separately. For an AI-driven PPA, it means having a fine-grained selection process, during which users should not be presented with bundled decisions. For instance, Shanmugarasa et al. [58]'s solution works per "situational context:" who (is requesting data), data type, purpose, and resharability (to third parties); while Xie et al. [69] only works for one type of data (location), therefore only meeting this option in a restricted sense.

5.7.4 Revoke. Finally, we observed that some AI-driven PPAs enable users to withdraw decisions. Here, rather than denying a decision or a recommendation, revoking operates after a given decision to withdraw it. This feature has rarely been observed in practice – at least explicitly – although Liu et al. [41] allows revoking previously granted decisions. Revoking previously made decisions, such as sharing data on social media, can be challenging to enforce. Also, note that certain operating systems – such as mobile OSes – will still allow users to revoke their decisions manually, although we stress that this action is performed outside the AI-driven PPA.

6 DISCUSSION

This systematic survey provides unique insights into how state-of-the-art research has designed AI-driven PPAs during the last decade. For instance, IoT became a system context of interest only in 2018, and we observe a similar late adoption trend for reinforcement learning after 2021.[8] However, AI techniques have been used in every system context for all types of decisions throughout the years without any apparent pattern. While this lack of a clear pattern is not the most informative in itself, we ought to look instead at the **gaps** this survey highlights, the **challenges** AI-driven PPAs raise, then to inform better **design and development recommendations** based on these analyses.

This section discusses the issue of *Evaluating AI-driven PPAs* in Section 6.1, AI-driven PPAs not sufficiently addressing *Privacy-by-Design* in Section 6.2, the (lack of) *explanations and explainability* in Section 6.3, the relationship with *legal considerations* in Section 6.4, the concerns surrounding *system contexts* in Section 6.5, the challenges in leveraging different *sources of data* in Section 6.6, to finally pave *research avenues* in Section 6.7.

6.1 Evaluating AI-driven PPAs

The problem of the evaluation of AI-driven PPAs is two-fold. First, we observe that **the evaluation of AI-driven PPAs are not based on the same or on comparable accuracy metrics or measurements**. Indeed, as presented in Section 5.6, accuracy is measured regarding the privacy decision, but also regarding privacy leakage, acceptability rate, etc. Second, **our data shows a lack of user study evaluation**, and our critical appraisal shows a trend toward "low" or "very low" scores for these user studies. Only 16 out of 39 papers mentioned that they performed a user study to evaluate their solution[9], only four of which are above (or equal to) 70% based on the CEBMA critical appraisal we performed. We acknowledge that such user studies might not be in the scope of theoretical papers (e.g., models or frameworks without prototype implementation). Yet, we contend that the validation offered by these theoretical papers, often cross-validation on a dataset, is far from being able to reflect reality.

Recommendation: *Based on the current lack of empirical studies, we propose that the usability of an AI-driven PPA should be validated with a user study following a high standard of practice, and such evaluation should notably encompass the accuracy of the privacy decision taken.*

6.2 Lack of Privacy by Design

We identified a gap regarding following a privacy-by-design approach for AI-driven PPAs since hardly any of the papers we surveyed focus on or mention how the AI-driven PPAs themselves can be designed in a privacy-preserving manner. More specifically,

[8]Some papers may have been published on the topic earlier than in 2013, the year from which we started to include papers in our survey.
[9]Some papers include a user study for collecting data, which is not the focus of the present argument.

among papers describing a technical architecture[10], **only one uses federated learning as a privacy-enhancing approach,[11] and many require trust in a central server** where the AI-driven PPAs' data processing is performed. However, Wijesekera et al. [67] provides an insightful analysis of the trade-off of having either a local (offline) or a remote computation, concluding that offline learning still performs well (almost 95% accuracy). Also note that the privacy threat models are rarely described, making it difficult to evaluate security and privacy assumptions critically.

Recommendation: *We contend that AI-driven PPAs must embrace stronger privacy-by-design principles, including better design strategies but also better integration of Privacy Enhancing Technologies for achieving data minimization, such as federated learning combined with differential privacy, or the use of privacy-preserving data analytics based on multi-party computation homomorphic or functionally encrypted data (see also [13]).*

6.3 Unexplainable AI

Another pitfall identified is the lack of explanations provided by most AI-driven PPAs, combined with the lack of explainability offered by the AI models used. **Only one of the surveyed papers explicitly addresses explainability of decisions** [42], and only 16 use a transparent model (see Section 5.2) to make a prediction.

The growing trend to use deep learning architectures may not facilitate the explainability of decisions, but this challenge is not insurmountable. It is indeed possible to devise *post hoc* explanations and take inspiration from other existing work on usable explanations for AI-made decisions. Note, however, that inherent transparency can come at the expense of the quality of decisions (deep neural networks tend to perform better than their simpler counterparts, although this statement does not seem to generalize to all kinds of decisions, such as decisions made in highly unpredictable settings like social predictions [47]).

As we discuss above in Section 2.1.1, transparency of AI can, in some specific use cases related to AI-driven PPAs, be a legal requirement for the data controller according to the GDPR, or for the provider or deployer according to the AI Act, even though this will not apply for the majority of AI-driven PPAs and use cases.

Recommendation: *Based on this analysis, we recommend 1) the use of inherently explainable AI models such as decision trees for the classification, whenever possible in terms of potential acceptable quality loss implications, or 2) the integration of ad-hoc explanations otherwise, e.g., for neural networks and SVMs. Indeed, explainability can also foster trust in technology, improving the uptake of such systems.*

6.4 Few Legal Considerations

As some privacy decisions made by AI-driven PPAs have legal privacy implications or issues, legal requirements, e.g., under the GDPR or the AI Act, should be discussed and considered for the design and use of AI-driven PPAs. Although partially explainable by 1) the geographic distribution of the solutions surveyed (only 12 papers have EU affiliations, in contrast with both the GDPR and the AI Act being EU regulations), and by 2) the timing of publications (17 papers were published before the GDPR was enacted, and none before the AI Act came into force), it can still be surprising to find **only 4 papers mentioning (but not even addressing) legal considerations**.

Recommendation: *We recommend a deeper consideration of legal requirements for the design of AI-driven PPAs. Such efforts should particularly amount to 1) meeting consent requirements when assisting on decisions related to consent, such as permission settings; 2) the introduction of AI-driven PPAs assisting and enabling users to exercise their data subject rights; and 3) the incorporation of usable explanations for the logic behind the AI-based proposed decisions, and an assessment of risks and consequences addressing related requirements by the GDPR for automated decision-making or the AI Act.*

In the more general case, we contend that even when legal privacy principles in a certain use case or context do not apply, **they can still provide valuable guidelines for the design of AI-driven PPAs**. For instance, assisting users with making informed, unambiguous, and explicit privacy decisions (as required for consent) may foster trust in AI-driven PPAs even when the privacy decision to be made does not formally constitute consent. Also, usable explanations of the risks and implications when using an AI-driven PPA can in general help raise awareness among users.

6.5 Missing System Context

Our SLR covered four system contexts: mobile applications, social media, IoT, and the cloud. We are, however, surprised **not to find other contexts, such as web browsers or Trigger-Action Platforms (TAPs)**. The former because cookie notices are notoriously a "hassle" for users in modern web experience; we therefore expected to encounter solutions tackling this issue. The latter refers to platforms offering applications for connecting otherwise unconnected devices and services using simple recipes, such as "Every morning at 7 am, send a Slack message with the first meeting of the day from Google Calendar." Trigger-action programming has gained a lot of traction in the last years (IFTTT, the most important TAP, boasts over 27 million users [32]), yet no AI-driven PPAs specifically addressed this environment.

Both these system contexts possess their specific features: many controllers with non-standard interfaces for cookie notices, and numerous actors mediated through a single centralizing entity for TAPs. They therefore require targeted efforts from designers to offer adequate technological solutions to manage privacy decisions.

Recommendation: *We encourage researchers and developers of AI-driven PPAs to expand their efforts into a broader range of system contexts, encompassing also but not limited to web browsers and TAPs.*

[10] Recall that theoretical papers are excluded from this analysis.
[11] However, without discussing that federated learning is still vulnerable to privacy attacks.

6.6 Use of Varied Sources of Data, Accounting for Both Context and Personal Preferences

Last, our study yielded that AI-driven PPAs leverage various sources of data (context, attitudinal data, behavioral data, type of data, content of data, and metadata), but not necessarily within the same solution. However, utilizing several of these data sources can be a challenge in itself, as it requires careful curation of the datasets and adequate use of the AI models. The benefits harnessed can be high, resulting in higher prediction accuracy.

We also acknowledge the difficulty of determining certain sources of data – such as context –, or the sensitivity of data. As mentioned in Section 5.3.1, context is rarely defined. It is, however, possible to draw inspiration for a rigorous definition from the seminal paper by Barth et al. [12] on the formalization in a logical framework of the concept of contextual integrity coined by Nissenbaum [48]. As for the sensitivity of data, it is notably incumbent on context when, for instance, the same type of data (e.g., location) can be deemed non-sensitive in one context (e.g., at a workplace in the middle of the week), but sensitive in another (e.g., Sunday morning near a church, thereby disclosing potential religious beliefs).

Recommendation: *Based on the relative singularity of data sources, we advocate for a plurality of data sources, encompassing context as much as personal preferences.*

6.7 Research Avenues

In this final section of our Discussion, we investigate prospective paths for research on AI-driven PPAs, informed by the results of our study and the current social, technical, and legal landscape.

6.7.1 The Future of AI-driven PPAs and Generative AI. While the uptake of generative AI, such as Large Language Models (LLMs), is undeniable, their application to privacy decisions is not addressed yet in the literature. It is however already possible to find privacy assistants powered by this technology, for instance Hamid et al. [28] who provide a benchmark for evaluating Generative AI-based Privacy Assistants, [12] although we did not find papers captured by our selection criteria. **We are only one step away from having AI-driven PPAs powered by LLMs**, which is not without raising interrogations.

Indeed, these models are intrinsically challenging to explain. Also, because of their inherent tendency to glitch (or "hallucinate"), **they come with privacy challenges regarding compliance with the data accuracy principle of the GDPR,[13] and should therefore be incorporated in AI-driven PPAs with caution**. Still, future research should address opportunities and challenges of designing and using LLM-based Personalized Privacy Assistants, as well as technical and legal requirements for involving LLMs in assisting users with privacy decisions.

6.7.2 Designing Genuinely Privacy-friendly AI-driven PPAs. A promising yet critical avenue for future research remains **to design a genuinely privacy-friendly AI-driven PPA, with the right amount of notice** to empower users and avoid the so-called "consent fatigue." This right amount of notice can be a difficult balance to achieve – some users favor more notice than others –but it is a crucial step for the uptake of such assistants.

The design should naturally be informed by the latest results in the academic literature [25]; it should carefully consider the number of notices, their content, their timing, etc. However, it should also be complemented by usability studies conducted in the early stages of the prototype, as iterations over the design of the assistant are likely to be required in order to fine-tune it.

6.7.3 Trust in the AI assistants and automation bias. Individuals tend to overly trust AI systems and favor AI-based decision-making while ignoring contradictory information made without automation, a phenomenon known as automation bias [20], which is a problem also related to the Elisa effect first described by Weizenbaum [64]. If the user's decisions are biased towards following a privacy recommendation proposed by an AI-driven PPA, the users' autonomy may be negatively impacted in practice. Hence, future research should examine if users may too easily trust and rely on proposed or nudged decisions by AI-driven PPAs without critically judging or adapting proposed decisions and how such a problem could be addressed by Human-Computer Interaction research.

7 THREATS TO VALIDITY SUMMARY

Here, we briefly address the SLR's threats to validity and explain the strategies used to mitigate them (see details in Appendix B). The three main threats are the following. **(1) Potential Planning Limitations:** biases may be introduced in the initial planning phase if research questions are not clearly defined and key topics are omitted, which could lead to flawed review results. To mitigate this threat, we clearly outlined two research questions and objectives and defined an SLR Protocol [14], piloted and refined before the review. **(2) Validity of the Search Process:** There is also a risk that relevant studies were missed during the search process, potentially impacting the accuracy of the SLR's findings. To mitigate such threats, we followed a stepwise search process involving literature screening, complete papers reading, forward and backward snowballing, and selection by two independent reviewers. We also base the SLR on four scientific databases (i.e., Scopus, Web of Science, IEEE Xplore, and ACM Digital Library) with high relevance to computer science, privacy, and data protection. **(3) Potential Bias in Synthesis Process:** The synthesis of data from reviewed studies could introduce biases, such as deriving a flawed taxonomy or missed potential research gaps. To avoid bias, we created a DEF with well-known classification schemes and involved all researchers in the synthesis process for consistency checking.

8 CONCLUSION

In this SoK, we provide a classification and common vocabulary to compare and discuss AI-driven PPAs. Although many papers (39 in our selection) have already been published on AI-driven PPAs in the last decade, they do not yet form a coherent body of literature. AI-driven PPAs can be improved by performing standard

[12] Note that their assistants do not support users in making privacy decisions.
[13] Art 5 (i) (e) GDPR indeed stipulates that data needs to be "accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('accuracy')".

[14] https://anonymous.4open.science/r/SoK_AI_PPA-E29F

evaluations (including their usability), integrating privacy by design in their design process, providing additional explanations for their decisions, and considering more system contexts and data sources. We hope this survey and its classification allows users and developers of AI-driven PPAs to compare different solutions and understand their pros and cons. Moreover, the recommendations should help improve AI-driven PPAs in different ways, addressing the challenges raised by AI's latest developments (including LLMs), data collection, and modern regulations.

ACKNOWLEDGMENTS

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

REFERENCES

[1] State of IoT 2024: Number of connected IoT devices growing 13% to 18.8 billion globally. URL https://iot-analytics.com/number-connected-iot-devices/.

[2] Davide Alberto Albertini, Barbara Carminati, and Elena Ferrari. Privacy settings recommender for online social network. In *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*. IEEE. ISBN 978-1-5090-4607-2. doi: 10.1109/CIC.2016.079. URL http://ieeexplore.ieee.org/document/7809750/.

[3] Md. Zulfikar Alom, Barbara Carminati, and Elena Ferrari. Adapting users' privacy preferences in smart environments. In *2019 IEEE International Congress on Internet of Things (ICIOT)*. IEEE, . ISBN 978-1-72812-714-9. doi: 10.1109/ICIOT.2019.00036. URL https://ieeexplore.ieee.org/document/8815665/.

[4] Md. Zulfikar Alom, Barbara Carminati, and Elena Ferrari. Helping users managing context-based privacy preferences. In *2019 IEEE International Conference on Services Computing (SCC)*. IEEE, . ISBN 978-1-72812-720-0. doi: 10.1109/SCC.2019.00027. URL https://ieeexplore.ieee.org/document/8813970/.

[5] Wolfgang Apolinarski, Marcus Handte, and Pedro Jose Marron. Automating the generation of privacy policies for context-sharing applications. In *2015 International Conference on Intelligent Environments*. IEEE. ISBN 978-1-4673-6654-0. doi: 10.1109/IE.2015.18. URL http://ieeexplore.ieee.org/document/7194273/.

[6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. URL http://arxiv.org/abs/1910.10045.

[7] Art. 29 Working Party. Opinion 2/2013: Apps on smart devices, 2013. URL https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp202_en.pdf.

[8] Gonul Ayci, Murat Sensoy, Arzucan Özgür, and Pinar Yolum. Uncertainty-aware personal assistant for making personalized privacy decisions. 23(1). ISSN 1533-5399, 1557-6051. doi: 10.1145/3561820. URL https://dl.acm.org/doi/10.1145/3561820.

[9] Tim Baarslag, Alper T Alan, Richard Gomer, and Muddasser Alam. An automated negotiation agent for permission management.

[10] Paritosh Bahirat, Yangyang He, Abhilash Menon, and Bart Knijnenburg. A data-driven approach to developing IoT privacy-setting interfaces. In *23rd International Conference on Intelligent User Interfaces*. ACM. ISBN 978-1-4503-4945-1. doi: 10.1145/3172944.3172982. URL https://dl.acm.org/doi/10.1145/3172944.3172982.

[11] Natã M. Barbosa, Joon S. Park, Yaxing Yao, and Yang Wang. *"What if?"* predicting individual users' smart home privacy preferences and their changes. 2019(4). ISSN 2299-0984. doi: 10.2478/popets-2019-0066. URL https://petsymposium.org/popets/2019/popets-2019-0066.php.

[12] Adam Barth, Anupam Datta, John C Mitchell, and Helen Nissenbaum. Privacy and contextual integrity: Framework and applications. In *2006 IEEE symposium on security and privacy (S&P'06)*. IEEE.

[13] Melek ÖNEN (EURC) Muhammad BARHAM Boris ROZENBERG Ron SHMELKIN (IBM) Monir Azraoui Sebastien CANARD Bastien VIALLA (ORA) Beyza BOZDEMIR, Orhan ERMIS. D3.1–preliminary design of privacy preserving data analytics, 2019. URL https://www.papaya-project.eu/sites/default/files/papaya/public/content-files/deliverables/PAPAYA_D3_1_Preliminary_Design_of_PPDA.pdf.

[14] Víctor Botti-Cebriá, Elena Del Val, and Ana García-Fornes. Automatic detection of sensitive information in educative social networks. In Álvaro Herrero, Carlos Cambra, Daniel Urda, Javier Sedano, Héctor Quintián, and Emilio Corchado, editors, *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)*, volume 1267. Springer International Publishing. ISBN 978-3-030-57804-6 978-3-030-57805-3. doi: 10.1007/978-3-030-57805-3_18. URL http://link.springer.com/10.1007/978-3-030-57805-3_18. Series Title: Advances in Intelligent Systems and Computing.

[15] André Brandão, Ricardo Mendes, and João P. Vilela. Prediction of mobile app privacy preferences with user profiles via federated learning. In *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy*. ACM. ISBN 978-1-4503-9220-4. doi: 10.1145/3508398.3511526. URL https://dl.acm.org/doi/10.1145/3508398.3511526.

[16] CEBMa. CAT manager app. URL https://cebma.org/resources/tools/cat-manager-app/.

[17] Hanbyul Choi, Jonghwa Park, and Yoonhyuk Jung. The role of privacy fatigue in online privacy behavior. 81. ISSN 07475632. doi: 10.1016/j.chb.2017.12.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0747563217306817.

[18] Lorrie Cranor, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, and Joseph Reagle. The platform for privacy preferences 1.0 (P3P1. 0) specification. *W3C recommendation*, 16, 2002.

[19] John W Creswell and J David Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, United States, 2017.

[20] Mary L Cummings. Automation bias in intelligent time critical decision support systems. In *Decision making in aviation*, pages 289–294. Routledge, 2017.

[21] Anupam Das, Martin Degeling, Daniel Smullen, and Norman Sadeh. Personalized privacy assistants for the internet of things: Providing users with notice and choice. 17(3). ISSN 1536-1268, 1558-2590. doi: 10.1109/MPRV.2018.03367733. URL https://ieeexplore.ieee.org/document/8490188/.

[22] Cailing Dong, Hongxia Jin, and Bart P. Knijnenburg. PPM: A privacy prediction model for online social networks. URL http://arxiv.org/abs/1606.07463.

[23] European Commission. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, (April), 2016.

[24] European Commission. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (text with EEA relevance). *Official Journal of the European Union*, (June), 2024.

[25] Yuanyuan Feng, Yaxing Yao, and Norman Sadeh. A design space for privacy choices: Towards meaningful privacy control in the internet of things. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

[26] Dorota Filipczuk, Tim Baarslag, Enrico H. Gerding, and M. C. Schraefel. Automated privacy negotiations with preference uncertainty. 36(2). ISSN 1387-2532, 1573-7454. doi: 10.1007/s10458-022-09579-1. URL https://link.springer.com/10.1007/s10458-022-09579-1.

[27] Ricard L Fogues and Pradeep K Murukannaiah. SoSharP: Recommending sharing policies in multiuser privacy scenarios.

[28] Aamir Hamid, Hemanth Reddy Samidi, Tim Finin, Primal Pappachan, and Roberto Yus. GenAIPABench: A benchmark for generative AI-based privacy assistants. URL http://arxiv.org/abs/2309.05138.

[29] Ron Hirschprung, Eran Toch, and Oded Maimon. Simplifying data disclosure configurations in a cloud computing environment. 6(3). ISSN 2157-6904, 2157-6912. doi: 10.1145/2700472. URL https://dl.acm.org/doi/10.1145/2700472.

[30] Ron S. Hirschprung and Shani Alkoby. A game theory approach for assisting humans in online information-sharing. 13(4). ISSN 2078-2489. doi: 10.3390/info13040183. URL https://www.mdpi.com/2078-2489/13/4/183.

[31] Soheil Human, Harshvardhan J. Pandit, Victor Morel, Cristiana Santos, Martin Degeling, Arianna Rossi, Wilhelmina Botes, Vitor Jesus, and Irene Kamara. Data protection and consenting communication mechanisms: Current open proposals and challenges. In *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE. ISBN 978-1-66549-560-8. doi: 10.1109/EuroSPW55150.2022.00029. URL https://ieeexplore.ieee.org/document/9799369/.

[32] IFTTT. IFTTT - automate business & home. URL https://ifttt.com/.

[33] Mounika Kasaraneni and Johnson P. Thomas. A self–learning personal privacy assistant. In Leonard Barolli, Flora Amato, Francesco Moscato, Tomoya Enokido, and Makoto Takizawa, editors, *Advanced Information Networking and Applications*, volume 1151. Springer International Publishing. ISBN 978-3-030-44040-4 978-3-030-44041-1. doi: 10.1007/978-3-030-44041-1_26. URL http://link.springer.com/10.1007/978-3-030-44041-1_26. Series Title: Advances in Intelligent Systems and Computing.

[34] Harkeerat Kaur, Isao Echizen, and Rohit Kumar. Smart data agent for preserving location privacy. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, . ISBN 978-1-72812-547-3. doi: 10.1109/SSCI47803.2020.9308396. URL https://ieeexplore.ieee.org/document/9308396/.

[35] Harkeerat Kaur, Rohit Kumar, and Isao Echizen. Reinforcement learning based smart data agent for location privacy. In Leonard Barolli, Isaac Woungang, and Tomoya Enokido, editors, *Advanced Information Networking and Applications*, volume 226. Springer International Publishing, . ISBN 978-3-030-75074-9 978-3-030-75075-6. doi: 10.1007/978-3-030-75075-6_54. URL https://link.springer.com/10.1007/978-3-030-75075-6_54. Series Title: Lecture Notes in Networks and

Systems.
[36] Barbara Kitchenham. Procedures for performing systematic reviews. Technical Report 2004, Keele, UK, Keele University, 2004. URL https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=29890a936639862f45cb9a987dd599dce9759bf5.
[37] Neil Klingensmith, Younghyun Kim, and Suman Banerjee. A hypervisor-based privacy agent for mobile and IoT systems. In *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications*. ACM. ISBN 978-1-4503-6273-3. doi: 10.1145/3301293.3302356. URL https://dl.acm.org/doi/10.1145/3301293.3302356.
[38] Nadin Kokciyan and Pnar Yolum. TURP: Managing trust for regulating privacy in internet of things. 24(6). ISSN 1089-7801, 1941-0131. doi: 10.1109/MIC.2020.3020006. URL https://ieeexplore.ieee.org/document/9178960/.
[39] Marco Kuhrmann, Philipp Diebold, and Jürgen Münch. Software process improvement: a systematic mapping study on the state of the art. 2. ISSN 2376-5992. doi: 10.7717/peerj-cs.62. URL https://peerj.com/articles/cs-62.
[40] Nadin Kökciyan and Pinar Yolum. Taking situation-based privacy decisions: Privacy assistants working with humans. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-00-3. doi: 10.24963/ijcai.2022/99. URL https://www.ijcai.org/proceedings/2022/99.
[41] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun Zhang, Norman Sadeh, Alessandro Acquisti, and Yuvraj Agarwal. Follow my recommendations: A personalized privacy assistant for mobile app permissions. Usenix Association. OCLC: 255312726.
[42] Sascha Lobner, Welderufael B. Tesfay, Toru Nakamura, and Sebastian Pape. Explainable machine learning for default privacy setting prediction. 9. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3074676. URL https://ieeexplore.ieee.org/document/9410256/.
[43] Ricardo Mendes, Mariana Cunha, João P. Vilela, and Alastair R. Beresford. Enhancing user privacy in mobile devices through prediction of privacy preferences. In Vijayalakshmi Atluri, Roberto Di Pietro, Christian D. Jensen, and Weizhi Meng, editors, *Computer Security – ESORICS 2022*, Lecture Notes in Computer Science. Springer International Publishing. ISBN 978-3-031-17140-6. doi: 10.1007/978-3-031-17140-6_8.
[44] Gaurav Misra and Jose M. Such. PACMAN: Personal agent for access control in social media. 21(6). ISSN 1089-7801. doi: 10.1109/MIC.2017.4180831. URL http://ieeexplore.ieee.org/document/8114620/.
[45] Victor Morel and Simone Fischer-Hübner. Automating privacy decisions-where to draw the line? In *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE.
[46] Toru Nakamura, Shinsaku Kiyomoto, Welderufael B. Tesfay, and Jetzabel Serna. Easing the burden of setting privacy preferences: a machine learning approach. In Olivier Camp, Steven Furnell, and Paolo Mori, editors, *Information Systems Security and Privacy*, volume 691. Springer International Publishing. ISBN 978-3-319-54432-8 978-3-319-54433-5. doi: 10.1007/978-3-319-54433-5_4. URL http://link.springer.com/10.1007/978-3-319-54433-5_4. Series Title: Communications in Computer and Information Science.
[47] Arvind Narayanan and Sayash Kapoor. *AI Snake oil: what artificial intelligence can do, what it can't, and how to tell the difference*. Princeton University Press.
[48] Helen Nissenbaum. PRIVACY AS CONTEXTUAL INTEGRITY. 79.
[49] Katarzyna Olejnik, Italo Dacosta, Joana Soares Machado, Kevin Huguenin, Mohammad Emtiyaz Khan, and Jean-Pierre Hubaux. SmarPer: Context-aware and automatic runtime-permissions for mobile devices. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. ISBN 978-1-5090-5533-3. doi: 10.1109/SP.2017.25. URL http://ieeexplore.ieee.org/document/7958625/.
[50] Frank Pallas, Katharina Koerner, Isabel Barberá, Jaap-Henk Hoepman, Meiko Jensen, Nandita Rao Narla, Nikita Samarin, Max-R. Ulbricht, Isabel Wagner, Kim Wuyts, and Christian Zimmermann. Privacy engineering from principles to practice: A roadmap. 22(2). ISSN 1540-7993, 1558-4046. doi: 10.1109/MSEC.2024.3363829. URL http://arxiv.org/abs/2404.03442.
[51] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, and David Fernandez Llorca. The role of explainable AI in the context of the AI act.
[52] Erika Puiutta and Eric MSP Veith. Explainable reinforcement learning: A survey. URL http://arxiv.org/abs/2005.06247.
[53] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson.
[54] Norman Sadeh, Bin Liu, Anupam Das, Martin Degeling, and Florian Schaub. Personalized privacy assistant, March 23 2021. US Patent 10,956,586.
[55] Odnan Ref Sanchez, Ilaria Torre, Yangyang He, and Bart P. Knijnenburg. A recommendation approach for user privacy preferences in the fitness domain. 30(3). ISSN 0924-1868, 1573-1391. doi: 10.1007/s11257-019-09246-3. URL http://link.springer.com/10.1007/s11257-019-09246-3.
[56] Florian Schaub, Rebecca Balebako, Adam L. Durity, and Lorrie Faith Cranor. A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*. URL https://www.usenix.org/conference/soups2015/proceedings/presentation/schaub.
[57] Marc Serramia, William Seymour, Natalia Criado, and Michael Luck. Predicting privacy preferences for smart devices as norms. URL http://arxiv.org/abs/2302.10650.
[58] Yashothara Shanmugarasa, Hye-young Paik, Salil S. Kanhere, and Liming Zhu. Automated privacy preferences for smart home data sharing using personal data stores. 20(1). ISSN 1540-7993, 1558-4046. doi: 10.1109/MSEC.2021.3106056. URL https://ieeexplore.ieee.org/document/9530451/.
[59] M. Shaw. Writing good software engineering research papers. In *25th International Conference on Software Engineering, 2003. Proceedings*. IEEE. ISBN 978-0-7695-1877-0. doi: 10.1109/ICSE.2003.1201262. URL http://ieeexplore.ieee.org/document/1201262/.
[60] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.
[61] Anna Cinzia Squicciarini, Dan Lin, Smitha Sundareswaran, and Joshua Wede. Privacy policy inference of user-uploaded images on content sharing sites. 27 (1). ISSN 1041-4347. doi: 10.1109/TKDE.2014.2320729. URL http://ieeexplore.ieee.org/document/6807800/.
[62] Hua-Zhe Tan, Wei Zhao, and Hai-Hua Shen. A context-perceptual privacy protection approach on android devices. In *2018 IEEE International Conference on Communications (ICC)*. IEEE. ISBN 978-1-5386-3180-5. doi: 10.1109/ICC.2018.8422188. URL https://ieeexplore.ieee.org/document/8422188/.
[63] Onuralp Ulusoy and Pinar Yolum. PANOLA: A personal assistant for supporting users in preserving privacy. 22(1). ISSN 1533-5399, 1557-6051. doi: 10.1145/3471187. URL https://dl.acm.org/doi/10.1145/3471187.
[64] Joseph Weizenbaum. Computer power and human reason: From judgment to calculation, 1976.
[65] Alan F Westin. Privacy and freedom. 25(1).
[66] Roel Wieringa, Neil Maiden, Nancy Mead, and Colette Rolland. Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. 11(1). ISSN 0947-3602, 1432-010X. doi: 10.1007/s00766-005-0021-6. URL http://link.springer.com/10.1007/s00766-005-0021-6.
[67] Primal Wijesekera, Joel Reardon, Irwin Reyes, Lynn Tsai, Jung-Wei Chen, Nathan Good, David Wagner, Konstantin Beznosov, and Serge Egelman. Contextualizing privacy decisions for better prediction (and protection). In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3173842. URL https://dl.acm.org/doi/10.1145/3173574.3173842.
[68] Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. ACM. ISBN 978-1-4503-2476-2. doi: 10.1145/2601248.2601268. URL https://dl.acm.org/doi/10.1145/2601248.2601268.
[69] Jierui Xie, Bart Piet Knijnenburg, and Hongxia Jin. Location sharing privacy preference: analysis and personalized recommendation. In *Proceedings of the 19th international conference on Intelligent User Interfaces*. ACM. ISBN 978-1-4503-2184-6. doi: 10.1145/2557500.2557504. URL https://dl.acm.org/doi/10.1145/2557500.2557504.
[70] Jun Yu, Zhenzhong Kuang, Baopeng Zhang, Wei Zhang, Dan Lin, and Jianping Fan. Leveraging content sensitiveness and user trustworthiness to recommend fine-grained privacy settings for social image sharing. 13(5). ISSN 1556-6013, 1556-6021. doi: 10.1109/TIFS.2017.2787986. URL http://ieeexplore.ieee.org/document/8249924/.
[71] Xiao Zhan, Stefan Sarkadi, Natalia Criado, and Jose Such. A model for governing information sharing in smart assistants. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. ISBN 978-1-4503-9247-1. doi: 10.1145/3514094.3534129. URL https://dl.acm.org/doi/10.1145/3514094.3534129.
[72] Haoti Zhong, Anna Squicciarini, David Miller, and Cornelia Caragea. A group-based personalized model for image privacy classification and labeling. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-0-3. doi: 10.24963/ijcai.2017/552. URL https://www.ijcai.org/proceedings/2017/552.

A SUPPLEMENTARY MATERIAL

A.1 Summary of the SLR Protocol

See also https://anonymous.4open.science/r/SoK_AI_PPA-E29F/Review_protocol.md.

A.1.1 Methods.

Design. This SoK study adopts the widely known methodology for systematic literature reviews (SLRs) proposed by Kitchenham [36]. The SLR methodology offers us a well-defined and rigorous sequence of methodological steps consisting of three main phases: (1) planning, (2) conducting, and (3) reporting the review.

Eligibility Criteria. Based on the research questions, and to reduce the likelihood of bias, the following inclusion and exclusion criteria are followed.

Inclusion Criteria
- Provides a technical solution (implemented or theoretical) to help end-users automate personal (and personalized) privacy decisions with an assistant (or artificial agent) in IT systems. - Papers from 2013 onward to concentrate on the state-of-the-art. - The concept of AI needs to be explicitly stated in the papers.
Exclusion Criteria
- Papers with solutions that are purely theoretical without substantial explanations on how they could be implemented in practice. - Papers with solutions that solely automate the analysis of privacy policies but without any type of personalization. - Papers with poor scientific quality (e.g., lack objectives or research questions, the methodology is not described, the solution is insufficiently/vaguely described, etc.).

Table 3: Criteria for the inclusion and exclusion of studies.

A.1.2 Information Sources and Search Strategy. Four scientific databases were selected, i.e., Scopus, Web of Science, IEEE Xplore, and ACM Digital Library, due to their high relevance to the areas of computer science and engineering, comprising the vast majority of published research in the field. We also specified inclusion and exclusion criteria (see Table 3) used during the screening of publications retrieved from the databases. Before starting the search process, two authors piloted the searches on all databases and ran a *calibration exercise* to verify the consistency of the inclusion criteria. For that, the authors independently screened 10% of the results and discussed their decisions. The conflicts were all discussed and solved, sometimes with the help of a third author. This process was repeated a second time, screening another 10% of the papers, at a point that the authors agreed with the consistency of the selection process.

Based on our RQs and previous preliminary searches when designing the SLR Protocol, we identified a list of nine relevant keywords, i.e., privacy, data protection, assistant, agent, artificial intelligence, machine learning, intelligent, automatic, and personalized. These keywords were used to construct the following search query:

```
(privacy OR "data protection") AND (assistant* OR agent*)
AND ("artificial intelligence" OR "machine*learning" OR
intelligent OR automat* OR personali*ed)
```

As such, the search query targets papers working on three joint topics: 1) privacy (or data protection), using either 2) an assistant or an agent, and leveraging 3) artificial intelligence or personalization.

A.1.3 Study Records.

Data management. To manage the screening process, we exported search results from each database and imported them to the RAYYAN software (https://rayyan.ai/), allowing two reviewers to select papers independently (i.e., double-blinded) and manage conflicts by a third reviewer. Duplicated publications were also removed using RAYYAN during the selection process. Bibliographies of final results were exported to Zotero (for citing and sharing research).

Data extraction. Preliminary components to be extracted:

- Bibliographic information, such as title, abstract, authors and affiliations, venues, year of publication, etc.
- Key information of the AI-driven PPA, such as its source(s) of data, its eventual architecture, its system context, the type of privacy decision considered, the accuracy of the decisions, the type of AI used, etc.
- The presence of a user study
- Extent of evaluation – scale of validation activity that is measured
- Quality assessment and critical appraisal of the studies that have validated or evaluated the AI-driven PPA
- Features of users control over decisions (initially guided by EU consent requirements)

Types of contributions. Inspired by Kuhrmann et al. [39] and Shaw [59], we classified publications by their types of contributions (i.e., multiple choice) according to the following:

Model Representation of observed reality by concepts after conceptualization.
Theory Construct of cause-effect relationships.
Framework Frameworks/methods (related to automated privacy decisions).
Guidelines List of advice.
Lessons learned Set of outcomes from obtained results.
Advice Recommendation (from opinion).
Tool A tool to automate privacy decisions.

Data synthesis. We collated and summarized results into classification tables. We composed a narrative synthesis for papers meeting our inclusion criteria.

Preliminary components of the data synthesis:

- Overall identification and classification of AI-driven PPAs in published research
- Classification tables presenting features of AI-driven PPAs used to coherently organize the solutions surveyed
- Comparison analysis – based on the features of AI-driven PPAs–, and narrative synthesis

[11] Only location
[12] Not necessarily, depends on what they call the Configuration Options
[13] Not by default, they have a sort of 'user settings' for expert users
[14] Not consistently

A.2 Further Results from Data Charting and Critical Appraisal

Countries	Total	Countries	Total
United States	13	China	2
United Kingdom	6	Israel	2
Japan	4	Portugal	2
Netherlands	4	Switzerland	1
Italy	4	Turkey	1
Germany	3	Canada	1
Spain	3	Australia	1
India	2		

Table 4: Countries of affiliation of authors of selected papers.

Year	N. of Publications	Year	N. of Publications
2013	0	2019	4
2014	1	2020	5
2015	3	2021	2
2016	3	2022	8
2017	6	2023	2
2018	5		

Table 5: Number of publications per year.

B THREATS TO VALIDITY

B.1 Threat I – Planning Limitations of the SLR

The first threat relates to the planning of the SLR in terms of identifying the need and justification for this study. Here, we were concerned with identifying existing reviews (systematic and non-systematic) on the topic of AI-driven PPAs. The initial searches did not reveal any review studies on the topic, as described in Section 3.1, pointing to a significant gap in secondary research AI PPAs. The planning phase of the SLR is also critical to outline the research questions and provide the basis for an objective investigation of the studies that are being reviewed. If the RQs are not explicitly stated or omit the key topics, the literature review results can be flawed, overlooking the key information. To mitigate this threat, we outlined two RQs and objectives (Section 3.1). In summary, we seek to minimize any bias or limitations during the planning phase when defining the scope and objectives of this SLR. As a last step in the planning phase, the team finalized and cross-checked the study protocol to minimize the limitations of the SLR plan before proceeding to the subsequent phases.

B.2 Threat II – Validity of the Search Process

Identifying and selecting the studies reviewed in the SLR are also significant processes to be observed. Selecting studies is a critical step; if any relevant papers are missed, the results of the SLR may be flawed. Therefore, we followed a stepwise process (Section 3.2), starting with a literature screening and followed by a complete reading of papers. This selection process was carried out independently by two reviewers. We also performed forward and backward snowballing, looking for references to other potentially relevant studies. Also, this SLR restricts the selection of publications to four scientific databases: Scopus, Web of Science, IEEE Xplore, and ACM Digital Library. These databases were used due to their high relevancy to computer science, privacy, and data protection, as well as to maintain a feasible search space. This search process gives us confidence that we minimized limitations related to (i) excluding or overlooking relevant studies or (ii) including irrelevant studies that could impact the results and their reporting in the SLR.

B.3 Threat III – Potential Bias in the Synthesis Process

Some threats should also be considered regarding the potential bias in synthesizing the data from the review and documenting the results. This means that if there are some limitations in the data synthesis, they directly impact the results of this SLR. Typical examples of such limitations could be a flawed research taxonomy and a mismatch of potential research gaps. To minimize the bias in synthesizing and reporting the results, we have created a data extraction form that uses well-known classification schemes, such as the ones proposed by Wieringa et al. [66] and Creswell and Creswell [19], or Arrieta et al. [6] for the classification of AI. Three researchers independently reviewed this data extraction form while revising the research protocol. While one of the researchers led the data extraction step, two other authors helped by cross-checking the work throughout the process for consistency. Three authors were involved in the creation of the classification scheme derived from the literature (i.e., shown in Section 2), actively working on reviewing the list of categories for consistency through a series of meetings. Furthermore, this SLR also offers a complete replication package, enabling researchers to reproduce or extend this review (https://anonymous.4open.science/r/SoK_AI_PPA-E29F).

Publication	Accuracy	User study	Critical appraisal	Type of contribution					
				Framework	Tool	Model	Theory	Lessons learned	Advice
Xie et al. [69]	68%	Online user experiment α	–	•				•	
Apolinarski et al. [5]	–	No	–	•	•				
Hirschprung et al. [29]	–	Online qualitative survey	D-, very low (55%)	•		•			
Squicciarini et al. [61]	92.53%	Cross sectional study [15]	D, very low (55%)	•		•			
Liu et al. [41]	78.7%	Randomized controlled studies [16]	A, high (90%)		•			•	
Albertini et al. [2]	–	Cross-sectional study	D, very low (55%)		•				
Dong et al. [22]	89,8% F1	Case studies α	–			•		•	
Baarslag et al. [9]	–	Randomized controlled study [17]	A, high (90%)		•	•		•	
Fogues and Murukannaiah [27]	Around 50%	Online survey α	–		•				
Zhong et al. [72]	79%	Survey α	–			•			
Misra and Such [44]	91.8%	Non-controlled before-after study [18]	C, limited (70%)		•				
Nakamura et al. [46]	85%	Cross-sectional study [19]	D, very low (55%)			•			
Olejnik et al. [49]	More than 80%	Yes, for data collection α	–	•	•				
Das et al. [21]	–	No	–		•				
Tan et al. [62]	95% [20]	No	–		•				
Wijesekera et al. [67]	95%	Experience Sampling Method	D, low (60%)		•			•	
Yu et al. [70]	–	Cross-sectional study [21]	D, very low (55%)			•	•		
Bahirat et al. [10]	81.54%	No	–			•			
Klingensmith et al. [37]	–	No	–		•				
Barbosa et al. [11]	86.8% [22]	Survey α	–			•			•
Alom et al. [4]	Up to 72.2% (satisfaction)	Cross-sectional study [23]	D, very low (55%)	•					
Alom et al. [3]	96.4% and 94.5% [24]	Yes, for labeling and evaluation α	–	•					
Kasaraneni and Thomas [33]	–	No	–		•	•			
Kaur et al. [34]	–	No	–			•			
Botti-Cebriá et al. [14]	– [25]	No	–		•				
Kokciyan and Yolum [38]	–	No	–			•			
Sanchez et al. [55]	84.74%	Online survey to build their dataset α	–			•			
Kaur et al. [35]	–	No	–			•			
Lobner et al. [42]	83.33% [26]	Survey α	–			•			
Filipczuk et al. [26]	65% [27]	Between-subject experimental design	C, limited (70%)	•	•				
Hirschprung and Alkoby [30]	–	Cross-sectional study	D, low (60%)	•		•			
Kökciyan and Yolum [40]	Between 41 and 92% [28]	No	–			•			
Ulusoy and Yolum [63]	Around 75% [29]	No	–			•			
Zhan et al. [71]	74%	No	–			•			
Brandão et al. [15]	Between 82 and 88%	Field study (cross-sectional study)	D, low (60%)		•				
Mendes et al. [43]	92%	Cross-sectional study [30]	D, low (60%)		•	•		•	
Shanmugarasa et al. [58]	92.62%	Cross-sectional study [31]	D, low (60%)		•	•			
Ayci et al. [8]	89%	No	–		•	•			
Serramia et al. [57]	3.78/5 [32]	Cross-sectional study [33]	D, very low (55%)		•	•			

α Alpha means that the user study is not meant to assess the solution, but only meant to collect data

16 A survey-based study and a direct user evaluation

17 Two surveys

18 Between-participants design

19 Online survey

20 Online questionnaire

21 For privacy leakage detection (not to be confused with preferences detection)

22 To measure the interpretability of the approaches

23 AUC of binary allow/deny for a given scenario

24 User satisfaction

25 Accuracy based on appreciation of evaluators

26 The accuracy presented is for the right category of data

27 With interpretability of the results

28 On average, but seems higher in specific case

29 Depends on several parameters

30 Difficult to assess because they measure utility of decisions in a simulated setting

31 Field study

32 Field study

33 Acceptability rate, not accuracy

34 To measure the level of comfort of the norms inferred

Table 6: Supplementary table of our results. We present the accuracy of the predictions (see Section 5.6); the presence or not of a user study, and the type of user study if applicable; the results of our critical appraisal (see Section 4); and finally the type of contribution, informing whether the surveyed solution presents a *framework*, a *tool* (i.e. with an implementation), a *model*, *lessons learned*, or *advice*. We denote with – when the criterion is not applicable (no implementation/tool is presented)