



Bolton, T., Dargahi, T., Belguith, S., & Maple, C. (2023). PrivExtractor: Towards Redressing the Imbalance of Understanding Between Virtual Assistant Users and Vendors. *ACM Transactions on Privacy and Security*. <https://doi.org/10.1145/3588770>

Peer reviewed version

Link to published version (if available):
[10.1145/3588770](https://doi.org/10.1145/3588770)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the accepted author manuscript (AAM). The final published version (version of record) is available online via ACM at <https://doi.org/10.1145/3588770>. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: <http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

1 Privextractor: Towards Redressing the Imbalance of Understanding Between Virtual 2 Assistant Users and Vendors

3
4 Thomas Bolton

5 School of Science, Engineering and Environment, University of Salford, UK, t.j.e.bolton@edu.salford.ac.uk

6
7 Tooska Dargahi

8 Department of Computing and Mathematics, Manchester Metropolitan University, UK, t.dargahi@mmu.ac.uk

9
10 Sana Belguith

11 Department of Computer Science, University of Bristol, UK, sana.belguith@bristol.ac.uk

12
13 Carsten Maple

14 Warwick Manufacturing Group (WMG), University of Warwick, UK, Maple, CM@warwick.ac.uk

15
16
17 The use of voice-controlled virtual assistants (VAs) is significant, and user numbers increase every year. Extensive use of
18 VAs has provided the large, cash-rich technology companies who sell them with another way of consuming users' data,
19 providing a lucrative revenue stream. Whilst these companies are legally obliged to treat users' information 'fairly and
20 responsibly', artificial intelligence techniques used to process data have become incredibly sophisticated, leading to users'
21 concerns that a lack of clarity is making it hard to understand the nature and scope of data collection and use.

22
23 There has been little work undertaken on a self-contained user awareness tool targeting VAs. Privextractor, a novel web-
24 based awareness dashboard for VA users, intends to redress this imbalance of understanding between the data 'processors'
25 and the user. It aims to achieve this using the four largest VA vendors as a case study and providing a comparison function
26 that examines the four companies' privacy practices and their compliance with data protection law.

27
28 As a result of this research, we conclude that the companies studied are largely compliant with the law, as expected.
29 However, the user remains disadvantaged due to the ineffectiveness of current data regulation that does not oblige the
30 companies to fully and transparently disclose how and when they use, share, or profit from the data. Furthermore, the
31 software tool developed during the research is, we believe, the first that is capable of a comparative analysis of VA privacy
32 with a visual demonstration to increase ease of understanding for the user.

33
34 CCS CONCEPTS • Security and privacy • Human and societal aspects of security and privacy • Usability in security and
35 privacy

36 1 INTRODUCTION

37 Cash-rich, monolithic private-sector technology companies are significant consumers of personal information with their
38 primary goal being revenues achieved from targeted advertising. Four of the world's five richest corporations are
39 Microsoft, Apple, Amazon, and Google; as of 2021, the four have a combined market capitalisation of US\$7.47 trillion
40 [1]. All four hold vast quantities of data relating to individuals that they use to sell targeted advertising [2] [3] [4] [5].
41 Google's total company-wide revenue in 2021 was US\$257 billion, of which the majority, US\$209 billion, came via its
42 Google Ads platform [6]. These four companies are not alone in marketing a voice assistant (VA), or in using personal
43 data for advertising purposes – far from it. They are, however, the biggest technology companies in the world in monetary
44 terms. For this reason, and to make the research more manageable, Microsoft, Apple, Amazon, and Google will be the
45 subjects of the case study in this paper.

46
47 George Orwell's future-dystopian fictional novel Nineteen Eighty-Four, published in 1949, refers to an electronic device
48 called the 'Telescreen' that has the ability to see, hear, and broadcast; in Orwell's fictional world, the Telescreen is forced
49 into every citizen's home by law. Today, the VA, a centrally-controlled hearing and broadcasting device, is marketed as a
50 must-have lifestyle accessory in a range of prices and designs [7] for which people voluntarily pay. The VA forms part of
51 the private sector's move towards large-scale collection and processing of personal data.

52
53 The complex and lucrative business of brokering online advertising relies on data that describes the user and their
54 preferences. One common source of this data is a user's web browsing history – their 'click behaviour' [8]. This data has
55 traditionally been collected via a user's keyboard input; however, companies have recently begun to use newly emerging
56 computing devices, equipped with microphones to enable data capture in the form of a user's speech, to harness additional
57 forms of information.

58
59 VAs such as Apple's Siri and Amazon's Alexa are software applications with which users can interact verbally, almost
60 conversationally. In return, the assistant can provide information, or can interact with devices around the home to which it
61 is connected – for example, to play music or switch off a light. It is important to understand, however, that regardless of
62 the device upon which the VA is installed, that device is simply an endpoint – the majority of the work in servicing the
63 user's voice command is carried out on the provider's servers [9]. [Section 3.1: Forensic Recovery outlines some findings
64 that show artefacts recovered from both Amazon's Alexa and Microsoft's Cortana – both in text form, and in the form of
65 recorded audio. These artefacts have been transmitted to, and stored on, the vendors' cloud backend. This transmission and
66 storage constitute what GDPR defines as 'processing' and, as such, is subject to that data law.](#)

67
68 It is clear that in an industry that is home to large cash-rich private companies, whose profit is largely (or even partly)
69 reliant on the gathering of data from individuals, the balance of power lies with those companies and not with the end user
70 of the products. Understanding the mechanics which underpin the process of bidding for advertising requires a good deal
71 of knowledge of computer science; even understanding the impact of the mechanics is not straightforward. Smit et al.
72 conducted a study in which participants were questioned on their understanding of online advertising; 41.1% of participants
73 in the survey believed that "*When I visit a website, I see the same ads as someone else visiting that website*", contrasting
74 with the 82.5% of participants who believed that "*Your browsing history determines which ads you are going to see during
75 your next visit.*" [10]. This disconnect suggests that, whilst users are aware that their browsing history is being mined, they
76 do not necessarily understand the impact this use of their data has on what they see when viewing online advertising.

77
78 VA users are becoming concerned about the lack of clarity that makes it hard to understand the nature and scope of data
79 collection [11]. Any user who wishes to better comprehend how their data is collected and used is reliant on the vendors
80 to explain this. However, whilst there is data law such as the General Data Protection Regulation (GDPR) governing the
81 behaviour and responsibilities of the vendors, it has been shown by Linden et al. that “*many [vendors’ privacy] policies*
82 *still do not meet several key GDPR requirements or their improved coverage comes with reduced specificity*” [12].

83 **1.1 Related Work**

84 A VA is a software application; more accurately, it is a whole series of connected software applications. A VA’s client can
85 take many forms. Software translations (or ‘ports’) of commercial VA clients such as Alexa are available for smartphones,
86 tablets, televisions, TV ‘sticks’ such as Google’s Chromecast, video games consoles, and dedicated smart speakers – to
87 name a selection. 38.2% of adults in the United Kingdom have adopted a smart speaker in the home – a growth of 24.5%
88 during the global COVID-19 pandemic [13]. In 2019, an estimated 3.25 billion VAs were being used globally. Forecasts
89 further estimate that by 2023 this number might hit eight billion globally at which point VAs will outnumber humans [14].
90

91 Having thoroughly researched the problem, we could find no equivalent privacy dashboards which examine and compare
92 VAs in terms of privacy and data law compliance in the way that Privextractor does. Tools which address privacy on
93 mobile devices exist, as do tools which help a user manage the settings of their VA devices, but none is comparable directly
94 to PrivExtractor.
95

96 There does appear to be a requirement for such a tool: Sharma et al. made a study of VA users and their perceptions towards
97 Google’s Assistant dashboard and found some concerns [15]. 38.7% of users were unaware that Google collects audio
98 recordings; when shown transcripts of these interactions with Google, the authors found that the participants would be
99 ‘uncomfortable’ sharing around 18% of their individual conversations with the company. In studying the vendors, Liao et
100 al. [11] applied a comprehensive, quantitative technical approach to analysing the privacy policies of two VAs – Google
101 Assistant and Amazon Alexa – along with the policies ascribed to the software add-ins for each platform. Using data
102 mining and machine learning techniques, the authors analysed the vendors’ data practices and found some alarming results:
103 not only were there many incorrect privacy policy URLs and broken links, but some of Amazon and Google’s voice apps
104 violated their own policy.
105

106 Zibuschka et al. have identified these privacy concerns and presented ENTOURAGE – a ‘privacy and security reference
107 architecture’; part of the authors’ system presents a dashboard to the user through which they may control the privacy
108 settings of their VA and manage data [16]. ENTOURAGE does not, however, offer an insight into data law and privacy
109 compliance of similar devices competing in the VA marketplace.
110

111 Privacy Flag, a European research project co-funded by the European Commission and the Swiss State Secretariat for
112 Education, Research and Innovation, has produced a smartphone application that aims to inform users of privacy-related
113 risks emanating from applications installed on the same device [17]. Privacy Flag’s backend system gathers inputs from
114 ‘technical enablers’ and via crowdsourcing in order to inform a user that an application installed on their device is, or is
115 not, privacy friendly. This is an interesting project with much scope for restoring the imbalance between users and the
116 providers of the applications. Currently, the application exists for Android devices only – porting the work to other mobile

117 operating systems would enable a much larger user base to benefit. The work undertaken here mirrors, to some degree,
118 what we would like to achieve with VAs and VA devices - whilst it would be more difficult to create an app native to the
119 devices, which lack touch and screen interfaces, there are many parallels with our goals.

120
121 Finding that privacy policies are “...*excessively long and difficult to follow*”, Harkous et al. created Polisis – a framework
122 that uses machine learning-based analysis of privacy policies to divide a policy into smaller, self-contained fragments for
123 easier digestion by the user [18]. The authors found that their application was able to apply icons to the privacy policy such
124 as ‘Precise location’ and ‘Data retention’ to an accuracy of 88.4%. As a fundamental part of PrivExtractor is the analysis
125 of VAs’ privacy policies, this research is of interest as a means to avoid the manual analysis of the relevant policies.

126 *1.1.1 Data and the Law*

127 To regulate the privacy of its citizens whose details were increasingly being recorded in government databases, Sweden
128 introduced the Data Act in 1973; this was not the first data law, but was the world’s first of its kind to apply to an entire
129 nation [19]. The United Kingdom took until 1987 to introduce its law – the Data Protection Act (DPA) – which would be
130 enforced by the newly-assembled Information Commissioner’s Office (ICO) [20].

131
132 The General Data Protection Regulation (GDPR) comprises data security and privacy law and came into effect on 25 May
133 2018; it is claimed to be the “*toughest privacy and security law in the world*” [21]. It was drafted and passed by the
134 European Union (EU) and imposes legal responsibilities on organisations anywhere in the world so long as they target or
135 collect the data of people resident in the EU. The regulation, in its current state, comprises 99 articles [22]. It was the
136 introduction of GDPR that led to a new DPA being made law in 2018 [20]. It should be noted that GDPR applies by itself
137 and does not require national implementation. However, the DPA is the benchmark for UK data protection law and required
138 changing to better reflect the comprehensive new regulation laid down in GDPR.

139
140 Following the UK’s formal exit from the EU on 31 December 2020, most of the EU GDPR was retained in UK law; this
141 retained GDPR is known as the “UK GDPR” [23]. The DPA is still the UK’s primary data protection law; the UK GDPR
142 sits alongside the DPA and applies to controllers and processors based outside the UK if they should offer goods or services
143 to individuals in the UK, or monitor the behaviour of UK individuals.

144
145 In GDPR, an adult is any data subject aged 16 years or over; section 9 of the DPA lowers this to the minimum allowed
146 under GDPR – 13 years [24]. To bridge that gap, the ICO has written a code of conduct for organisations dealing with the
147 data of children and young adults. The ICO’s age-appropriate design is a code of practice for online services that came
148 into force on 2 September 2020. The code explains how organisations can ensure their online services appropriately
149 safeguard children’s personal data. It is intended that organisations can use it to demonstrate that they comply with GDPR
150 [25]. It is important to note that, whilst the GDPR and DPA are generally applicable laws, the ICO’s Age-appropriate
151 design code is limited in its applicability and does not have the same legislative significance.

152 *1.1.2 User Perceptions*

153 Lau et al. [26] found that people who choose to adopt a VA have worries that differ from those who do not. Those who
154 refuse to see the purpose of such a device are more likely to hold privacy concerns. It is these users who are, for example,
155 “*deeply uncomfortable with the idea of a ‘microphone-based’ device that a speaker company or an ‘other’ with malicious*

156 *intent could ostensibly use to listen in on their homes*". Users who are keen to use VAs hold fewer concerns; this lack of
157 worry is rationalised with the belief that the vendor can be trusted, and that it would be impossible for an unauthorised
158 individual to access their data.

159
160 In a control group of users, acceptance factors of various VAs were considered by Burbach et al [27]. Using a choice-based
161 conjoint analysis with three attributes - natural language processing (NLP) performance, price, and privacy - the authors
162 found that privacy is the chief concern among users. These findings appear to loosely tally with those of Lau et al [26].
163 However, the surveys used were quite different in design, and the primary goal of the two studies was also distinct. Burbach
164 et al.'s study was comprehensive in its design; the authors divided the survey participants into groups according to their
165 chief concern. Only one group, named the 'Thrifty' by the authors, were concerned by the price of the VA as opposed to
166 its privacy, or potential lack thereof. This segment, however, formed only 18% of the total user group who were,
167 overwhelmingly, concerned about privacy more than cost or NLP performance. Combining the attributes of cost and
168 privacy, Ebbers et al. analysed user preferences of key attributes of VA privacy features – the amount of personal data
169 shown to users, the explainability of the VA's decisions, and the gamification level of the UI [28]. A key finding showed
170 that 56.4% of participants would be willing to pay for privacy features; these users were young and concerned about
171 privacy.

172
173 Again taking a technical approach, this time combined with user education, Seymour et al. [29] developed the software
174 tool, Aretha. Designed to demonstrate to a user both the data coming in and out of the home and the ramifications of this,
175 Aretha was deployed in three users' homes and the users' behaviour was observed. One finding, in particular, is interesting
176 for our study: "*The lack of engagement with the firewall was instructive in its own way; while most participants found it*
177 *difficult to use effectively, due to having already observed, interpreted, and understood the underlying behaviour of their*
178 *devices they appeared better able to adapt, invent, or imagine other protective mechanisms, tools, and strategies.*" This
179 observation suggests that, when armed with clear information as to how their data is treated, users feel more empowered
180 to control what is shared with others.

181 *1.1.3 Security*

182 We can see then that privacy, and the security necessary to ensure that VA interactions and information remain private,
183 are important to users. The field of usable security examines, amongst other aspects, how security is traded with usability;
184 a review by Lennartsson et al. finds that "*Usability is hampered when users' primary tasks are disturbed.*" [30]. Further
185 findings, that "*Necessary security actions should be arranged in ways that minimize interruptions*" and "*compelling users*
186 *to remember passwords repeatedly interrupts other tasks as enforced context switches may cause confusion*" are of interest
187 when viewing security implications of a VA – a tool that is, by design, friction-free in its use of a voice interface.

188
189 Yan et al. conducted a comprehensive survey on VA cyber attacks and countermeasures that users might employ [31].
190 Perhaps counterintuitive to Lennartsson et al.'s findings on usable security, Yan et al. recommend to users that they might
191 avoid using their VA for bank account management or home unlocking, avoid leaving the VA unattended, disallow wake-
192 word detection, or disable the VA when the device is locked. These are all solid suggestions; however, they tend to fly in
193 the face of the VA's unique, straightforward mode of operation. Interestingly, one finding from Yan et al. is that "*Despite*
194 *that it is the users who actually suffer from the security consequences, there are relatively few [things] they can do to*
195 *prevent the attacks.*"

196 1.1.4 Children and VAs

197 McReynolds et al. researched the comparatively novel area of connected toys, such as Jibo and Hello Barbie [32]. These
198 were studied in parallel with ‘adult’ VAs to answer a number of questions – chiefly whether a child relates to a VA in the
199 same way it would the toys. The authors found that five of the nine parents who took part in the study – when asked if their
200 children interacted with ‘adult’ VAs - “...explicitly observed that Dino [toy] was similar to Siri and other artificial
201 intelligence voice recognition systems”. The relationship children have with VAs was the subject of a work by Girouard-
202 Hallam et al. who made a study revealing children’s perceptions of a VA and, in particular, whether the child thought of
203 the VA as having mental, social, and moral attributes [33]. This was a comprehensive study with some revealing findings
204 around how children perceive and interact with VAs; 65% of participants responded ‘yes’ to the question ‘Can [device] be
205 your friend?’. In summary, the authors note that “*Children’s beliefs about their social, or perhaps parasocial, relationships
206 with voice assistants may also influence their understanding of cybersafety. Believing that a voice assistant could keep a
207 secret, and that it is, at least in part, amoral entity, may contribute to young children oversharing information with internet-
208 based devices*”; the differences in the way in which VA vendors treat children and their data, and an adult’s, will be part
209 of this study.

210 1.1.5 Technology and Forensics

211 Javed et al. [34] conducted an in-depth study of what Alexa is listening to. Disputing Amazon’s claim that until the wake
212 word is used no recording will take place, the study found that the VA was indeed recording: 91% of a control group of
213 users had experienced such an unwanted episode. After investigation, it was discovered that passive sounds – radio,
214 television, background noise – were recorded in the majority of these cases. More seriously, and representing a privacy
215 concern, were the recordings made of sensitive information experienced by 29.2% of the study group.

216
217 There exists little documentation of the finer details of how VAs communicate with their cloud services; Ford et al.
218 undertook a study of Amazon’s Alexa and its voice streaming network traffic, ostensibly to discover if VA devices were
219 recording and streaming conversation without being prompted by the user [35]. Finding that Alexa’s internet traffic uses
220 Transport Layer Security (TLS) for its communications with the cloud service, and not having a key with which to decrypt
221 the traffic, the authors were forced to resort to observing patterns in the quantity of data that is exchanged between the
222 device and its cloud platform in order to make any useful analysis.

223
224 Akinbi et al. attempted to recover forensic data from an Android smartphone running both Google Home and Google
225 Assistant that was also used to control a Google Nest device [36]. The authors found useful forensic artefacts on the device,
226 along with a chronology of voice interactions. For the purposes of this study, the most interesting finding was the ability
227 to recover copies of past conversations from the user’s Google cloud service account. One of the more comprehensive
228 studies made of VA forensics was that by Chung et al [37]. The authors made a thorough examination of Amazon’s Alexa
229 ecosystem and were able to extract artefacts from both device and cloud, and develop a web-based dashboard to display
230 the information in a user-friendly manner. The cloud artefact extraction is of particular interest as the exercise revealed an
231 undocumented Web API that could be queried in order to retrieve data pertaining to both the user’s account and their
232 interactions with Alexa.

233
234 Microsoft’s VA, Cortana, was the subject of a study made in 2017 by Singh et al. [38]. The authors were able to extract
235 and examine forensic artefacts from local database storage and wrote Python scripts to simplify this process for future

236 investigators. Possibly due to the ‘walled garden’ nature of Apple’s mobile operating systems, there are fewer studies
237 available that focus on Siri. One such was made by Horsman in 2019 [39] in which the author noted the information that
238 could be extracted from Siri on a locked iOS device using carefully crafted voice interactions.

239 1.2 Research Questions

240 There has been little work undertaken in the development of a self-contained user awareness tool targeting virtual
241 assistants. This problematic imbalance of understanding between the vendors and the end user, and the lack of access to
242 clear information regarding the vendors’ adherence to data law, is what this research seeks to address. To this end, the
243 following research questions (RQ) are asked:

244
245 **RQ1:** If a user of a voice assistant wishes to know the extent to which their personal information is harvested by the vendor
246 of their chosen VA, does that vendor clearly and unequivocally state the exact nature of the information that they collect,
247 how securely they keep that data, what they are doing with it, and for how long they keep it?

248
249 **RQ2:** Using the UK and European data law as a basis – GDPR, the Data Protection Act 2018 (DPA) and the Information
250 Commission Office’s (ICO) age-appropriate code of conduct – can it be demonstrated to a user where the data collection
251 practices of the VA vendor conflict with the law?

252 1.3 Contributions

253 This paper makes the following contributions:

- 254
255 • We systematically analyse the privacy policies of four major VA vendors, as a case study, to determine if those
256 policies comply fairly with the UK and European data laws. We find that the problems are twofold: the vendors,
257 whilst ostensibly complying with data laws such as GDPR, give little information to enable the user to see exactly
258 how their data is manipulated. However, the blame for this cannot be placed solely upon the vendors: our analysis
259 demonstrates that the data law itself offers insufficient requirements for specific transparency on behalf of the
260 vendor.
- 261
262 • We use this analysis to tabulate where problems with vendors’ compliance lie and, importantly, how each
263 vendor’s compliance and transparency compare with that of the others in the study; we find that for each question
264 asked about the policies, there are varied results. There are areas such as ‘unintended processing’ – when a VA
265 listens and processes data without being asked to – where all four vendors fail.
- 266
267 • Using this information, we develop Privextractor: a web-based software tool that enables users to not only
268 understand the privacy issues that surround the use of VAs but to see, simply and clearly, how the VA that they
269 use compares with others on the market. We see Privextractor primarily as a decision-making tool for use when
270 selecting a VA; with more development of the forensic capabilities of the dashboard Privextractor might become
271 a companion for use throughout the time that the user owns their VA. As far as we are aware, this is the first
272 study to develop such a dashboard.

273

274 With this information, as well as the tool Privextractor, we enable users to see that the protection of law such as GDPR
275 does not necessarily bestow the protections that might be expected. Users might not be able to use Privextractor to see, for
276 example, how a vendor is processing their information, and demonstrate that the vendor is not prepared to disclose this
277 information which is a privacy concern. There is much scope for improved law and greater enforcement of that law. It is
278 important that users are able to understand how their personal and private data is being manipulated and, as such, vendors
279 need to improve both compliance and clarity.

280 **1.4 Organisation**

281 The remainder of this paper is organised as follows: Section 2 outlines the methodology for the research and practical
282 elements of the project. Section 3 expands on the methodology by describing exactly how the research and experiments
283 were carried out. The results are shown in Section 4; Section 5 concludes the paper and answers the research questions.

284 **2 METHODOLOGY**

285 To meet the goals of answering the research questions posed in Section 1, the following methodology is proposed. The end
286 solution takes the form of a prototype, proof-of-concept web application called Privextractor. Privextractor – a novel user
287 awareness dashboard – is presented as a web application built using a standard Microsoft technology stack: .NET Core
288 framework using a model-view-controller (MVC) design pattern.

289 **2.1 Comparison Matrix**

290 To present to the user a clear, unbiased picture of the VA vendors’ data practices and their compliance with data law, a
291 comparison matrix was developed for which a series of subject areas was chosen. For each subject area, several questions
292 were devised, decomposing the subject area into smaller specific areas of interest; the answers can not only inform a user
293 of the compliance of their chosen device vendor, but also compare that device with others on the market such that the user
294 gains an awareness when selecting a VA.

295
296 These subject areas were motivated by a TechDispatch article published by the office of the European Union’s Data
297 Protection Supervisor [40]. Whilst not reflective of official data policy, the article outlines areas of privacy concern
298 specifically pertaining to VAs; it is these areas which are specific to the voice interface of the VA that are of particular
299 interest. The individual questions’ levels – three for each, denoting a level of compliance – were devised during preliminary
300 research into the vendors’ privacy policies. This research enabled us to find the level for each, where one vendor might be
301 transparent and give plenty of compliance information (the ‘good’ level) and another might give little information (the
302 ‘poor’ level).

303
304 Each of the four vendors’ privacy policies, legal notices, and any advertising or cookie-specific disclaimers were examined
305 in detail to gain an insight into how transparently they are written and how much relevant detail is supplied. The questions
306 are then tested against each vendor’s policies and the answers are collected and written as objectively as possible to aid
307 further comparison between the vendors.

308
309 The answers were compared and each of the four VAs was assessed to give an immediate visual indication of a) how the
310 user’s chosen VA complies with data protection law, and b) how the user’s choice of VA compares – in terms of vendor
311 transparency – with the other three. The subject areas are as follows:

312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349

- **Transparency** – data controllers and processors, types of data processed, purposes of the processing, specific processing of biometric data
- **Consent** – decisions made on processing the data of a specific individual
- **Children** – distinguishing adults from children, age verification, parental responsibility and consent, parental controls
- **Unrequested Processing** – wake word confusion, deliberate wake word tampering
- **Data Repurposing** – data profiling and purposes of profiling, transfer of data to third parties
- **Data Retention** – length of time for which data is kept, user’s ability to control and delete (all or some) data, the ‘Right to be Forgotten’
- **Security** – access control (account and device/app), indications of security technologies employed by the provider to ensure protection of the user’s privacy
- **Government Surveillance** – handling of access requests from law enforcement and government

Each area considers both how the device and/or application operates. Additionally, each subject area considers the privacy policy information for each vendor and presents a clear picture to the user showing compliance (or otherwise) with GDPR. As GDPR and the DPA are almost identical for the purposes of this work, we will focus only on GDPR for the sake of simplicity. Any indication that the vendors had considered the ICO’s age-appropriate code of conduct (where appropriate) was also taken into account.

As well as an immediate visual indication of the user’s chosen VA both in the context of the questions asked of it and the corresponding performance of the other VAs, accompanying information is provided to the user placing the results in context with simple explanations and links to the appropriate data protection law. It is important to note here that a three-stage traffic light approach might seem odd at first glance to someone who practises law; strictly speaking, a legal requirement is either met or it is not. However, in the case of GDPR and the DPA, it will be seen that a law may be interpreted in different ways; this is particularly true if the law is insufficiently explicit in its requirements. With this comparison matrix, we are attempting to demonstrate to the user each vendor’s understanding of the law; the three-stage compliance levels indicate where each vendor is explicit themselves in how they adhere to data and privacy law, and where they might fall down in not imparting sufficient information to the user in their privacy and legal documents.

Privextractor’s user interface allows the user to select ‘their’ VA and always presents the matrix of information relative to that selection.

2.2 Forensic Recovery

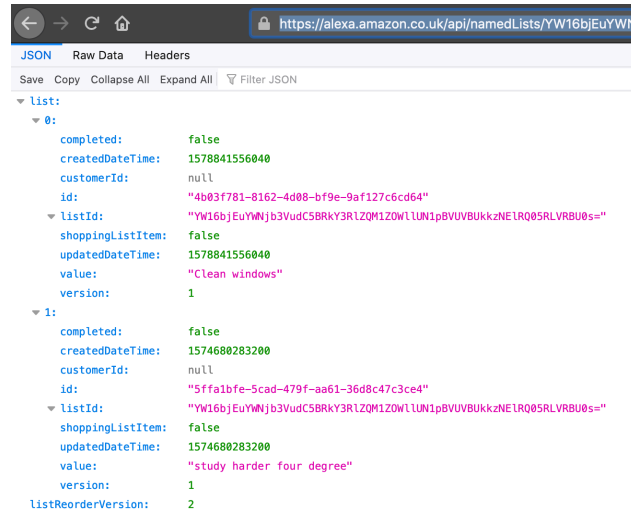
In addition to the comparison matrix, we undertook a forensic investigation of Amazon’s Alexa with interesting results. It was found that a user’s data could be recovered – using a session token found in the user’s web browser – from Amazon’s cloud service API. A range of artefacts was found including account information, and recordings of the user’s voice request and their associated text translation made by Amazon’s natural language processing. Alexa’s software-generated ‘replies’ were also found in text form. Similar artefacts were recoverable from Microsoft’s Cortana.

350 **3 PRIVEXTRACTOR DESIGN**

351 This section introduces the methodology used to recover forensic artefacts from two of the VA vendor’s ecosystems; we
352 also introduce the design of the matrix of questions for vendors that forms the basis of Privextractor.

353 **3.1 Forensic Recovery**

354 Preliminary testing revealed that Amazon exposes an API via which user data associated with Alexa interactions can be
355 read. After authenticating to the Alexa web client, the API can be queried and results returned in formatted JSON from the
356 Alexa account. To illustrate, Figure 1 shows an Alexa ‘reminder’ being served to the browser from Amazon’s API.
357
358



359
360
361 Figure 1: an Alexa ‘reminder’ in JSON format

362
363 Microsoft’s Cortana has shown to offer the same facility. Figure 2 shows that, after logging into Microsoft’s Privacy
364 Dashboard, an API can be queried to return data associated with Cortana interactions.
365
366

```

{
  "items": [
    {
      "displayText": "What is the weather?",
      "application": "Cortana",
      "deviceType": "Phone",
      "audioId": "29413984EB6D4D5AB28A71DA71878C0F",
      "timestamp": "2020-07-24T16:13:04.4452384+00:00",
      "id": "H4sIAAAAAAECXKM06CQBAF0Lv8VjB_Zwd2mG5h9QTGnoAFsAhdos7o7I",
      "cardType": "cardType_voice",
      "sources": []
    },
    {
      "displayText": "Who is prime minister of the United Kingdom?",
      "application": "Cortana",
      "deviceType": "Phone",
      "audioId": "B92FB924D8C643338AD8A3A4BF0B0851",
      "timestamp": "2020-07-24T15:21:08.9535859+00:00",
      "id": "H4sIAAAAAAECXKM06CQBAF0Lv8VjF_Z320drpdCCog9EYtiIWJoSPcXRI",
      "cardType": "cardType_voice",
      "sources": []
    },
    {
      "displayText": "What is the rweather?",
      "application": "Cortana",
      "deviceType": "Phone",
      "audioId": "73D4297FD0745AD98715E6B4BDC9A61",
      "timestamp": "2020-07-24T15:20:43.2040061+00:00",
      "id": "H4sIAAAAAAECXKM06CQBAF0Lv8VjF_h9kddzpg9ATG3ggFsTahdIS7Y-t",
      "cardType": "cardType_voice",
      "sources": []
    }
  ],
  "state": null
}

```

367
368
369

Figure 2: Cortana interactions stored by Microsoft and returned in JSON format

370 Google offers a web-based privacy tool which, upon inspection, retrieves data from an undocumented API to display
371 information pertaining to user interactions with the VA. However, this API was not able to run independently of Google's
372 host code. Apple's Siri has no privacy tool or other client available on the web; all user actions pertaining to Siri's data
373 must take place within native iOS or iPad OS tools.

374
375 Amazon, Microsoft, and Google all offer web-based privacy dashboards. Using browser-based developer tools, it could be
376 seen that asynchronous XMLHttpRequest (XHR) requests were being made to the vendors' servers – for example,
377 Amazon's Alexa dashboard was making requests to <https://alexa.amazon.co.uk/api/notifications>. By examining the header
378 for this request, it was possible to see the information that was passed to Amazon's server including, crucially, a session
379 token used to authenticate the API. Similarly, examining the response payload revealed that information was being returned
380 by Amazon's servers. A similar method revealed how Microsoft's privacy dashboard was communicating with its cloud
381 backend, and the information that was being transmitted and received.

382
383 Using some code written in Microsoft C#.NET, as part of the PrivExtractor system, it was possible to replicate these
384 requests for both Alexa and Cortana. A valid session token, acquired from the browser whilst logged into Alexa or
385 Microsoft Office365 (in the case of Cortana), was required in order to authenticate. Once an authenticated connection to
386 the APIs could be made, data could be retrieved in JavaScript Object Notation (JSON) format and rendered to the browser.

387
388 The data that was recovered from both Alexa and Cortana was real-world data –the information and voice recordings
389 accessible via the API were a result of voice interactions made by the author and, in the case of Alexa, the author's wife
390 using the same device. At this time, the forensics experiments should be considered as a lab study: despite integrating this

391 facility into PrivExtractor’s user dashboard, it is first necessary to obtain a valid session token which is not an everyday
392 user task.

393 3.2 Comparison Matrix

394 The eight sections for the comparison matrix to be displayed in Privextractor are described here in further detail. Each
395 section is designed to cover an aspect of GDPR and the UK’s DPA; Section 3 (Children) makes additional reference to the
396 ICO’s age-appropriate design code of conduct, designed to account for those users of information services aged between
397 GDPR’s default of 16 years and the DPA’s implementation that defines an adult as 13 years or over.

398
399 Each section of the comparison matrix is intended to cover a specific topic, as per the controller’s responsibilities laid
400 down in Article 24. Having studied and understood each of the four vendors’ privacy policies, legal notices, and any
401 advertising or cookie-specific disclaimers, that insight is used to devise a series of questions. These questions cover a broad
402 range of areas within each specific topic and are intended to indicate the vendors’ level of compliance with data law.

403
404 This is a qualitative assessment. To provide a measure of visual information, however, each question is assessed in terms
405 of the information provided by the vendor and how this appears to meet the requirements of data legislation. Each question
406 will be given a scale of three possible scores, of ‘good’, ‘average’, or ‘poor’. The definitions for each will be clearly signed.

407
408 The scores are not intended to be an immediate indication of quality when taken in isolation, as might be awarded to a
409 product in a consumer magazine review. However, taken together, the scores show broadly how each vendor is committing
410 to data protection law and, crucially, indicate to an end user how their choice of VA performs when compared with others
411 on the market. The results are tabulated in much the same way that commercial risks are evaluated using a matrix in
412 ISO27001 [41]; the final tables give a clear, colour-coded indication of performance both in isolation and in the context of
413 other VA devices.

414 3.2.1 Transparency

415 **Question 1: Is it clearly stated who the data controllers/processors are?**

- 416 • (Good) Yes – name and address
- 417 • (Average) Yes – name only
- 418 • (Poor) Not stated

419
420 GDPR makes specific definitions of ‘controller’ and ‘processor’. This question asks if the vendors specifically outline
421 which parts of their businesses are responsible for each role and if any detail is given.

423 **Question 2: Are the types of data processed – such as a user’s name or location data - clearly listed?**

- 424 • (Good) Yes – examples are given covering GDPR
- 425 • (Average) Yes – generic classifications only, incomplete coverage of types stated in GDPR
- 426 • (Poor) No – data types not listed, even in generic form

427

428 GDPR gives a list of examples of ‘personal data’ that might be taken from an individual during the use of their products.
429 Question 2 asks how specific the vendors are when giving examples of those types of data that might be collected from a
430 user.
431

432 **Question 3: Are the purposes of processing clearly listed?**

- 433 • (Good) Yes – examples given covering GDPR
- 434 • (Average) Yes – generic classifications only, incomplete coverage of types stated in GDPR
- 435 • (Poor) No purposes given, even in generic form

436
437 GDPR gives clear examples of what it considers to be the ‘processing’ of data; these range from the simple act of collecting
438 the data in the first instance, to disposal at the end of the process. Question 3 asks how specific the vendors are when
439 outlining the processing purposes.
440

441 **Question 4: Is any processing of biometric data clearly explained?**

- 442 • (Good) Yes – examples given covering GDPR
- 443 • (Average) Yes – generic examples only, incomplete coverage of types stated in GDPR
- 444 • (Poor) No information about biometric data processing is given

445
446 GDPR has a definition of what constitutes ‘biometric data’. Voice recordings might not appear as categorically ‘biometric’
447 as, say, a fingerprint or retinal scan; however, each of the VA vendors does engage in some form of fingerprinting –
448 identifying a person using their data – to personalise the user experience upon recognising their voice. This voice
449 fingerprinting process has a significant precedent: when the UK’s governmental tax collection department – Her Majesty’s
450 Revenue and Customs (HMRC) – adopted voice authentication in 2017, complaints were made by industry watchdogs due
451 to the lack of transparency from HMRC [42].

452 *3.2.2 Consent*

453 **Question 1: Does the device feature a mechanism whereby it processes the data of only a specific individual?**

- 454 • (Good) Yes – data processing limited to a single user at the device level
- 455 • (Average) Only for specific features, or for personalisation
- 456 • (Poor) No mechanism offered – the device will process the data of any user who interacts with it

457
458 For the controller to process data, consent must be obtained from the user. This is important enough for GDPR to define
459 what it means by ‘consent’. The ‘data subject’ is the one it is assumed has given consent; another person using the same
460 VA might not have done. [Currently, none of the four VAs has the facility to perform voice identification without sending](#)
461 [the recording to the vendor’s cloud service, at which point the transmission of the data as well as the analysis at the vendor’s](#)
462 [end is considered ‘processing’ by GDPR. It is possible for this to happen without consent having been given.](#) However, as
463 VAs become more capable at the device level, Privextractor will be updated and the results of this question might change.

464 *3.2.3 Children*

465 This section of the matrix introduces the ICO’s age-appropriate design code. This code is not enshrined in UK law, rather
466 it sets standards and explains how UK GDPR ‘*applies in the context of children using digital services*’ [25]. Whilst GDPR

467 considers an ‘adult’ to be anyone over the age of 16 years, the DPA lowers this to 13. The ICO’s code helps bridge this
468 gap of three years with advice to providers of digital services whose services might be either aimed at children or whose
469 services might reasonably be accessed by a child.

470

471 **Question 1: Does the provider distinguish between adults and children as users?**

- 472 • (Good) Yes – with age explicitly stated
- 473 • (Average) Yes – no age stated
- 474 • (Poor) No distinction made based on the user’s age

475

476 As a baseline for Question 2, this question asks if the vendor states, in their privacy policies, what they consider to be the
477 age of an adult as distinct from a child?

478

479 **Question 2: If applicable, what form does the age verification mechanism take?**

- 480 • (Good) External verification using endpoints not easily obtainable by children (credit card)
- 481 • (Average) Basic input of age, with external verification using endpoints easily obtainable by children (email,
482 SMS)
- 483 • (Poor) External verification only as a means of two-factor authentication, age not considered

484

485 **Background:** GDPR requires that controllers ‘shall make reasonable efforts’ to verify the age of the primary user during
486 initial setup, or that consent is given by the responsible adult. The ICO’s Age-appropriate design code goes further and
487 suggests some mechanisms by which this might be done, from simple self-declarations to more complicated credit card
488 checks. However, even the strongest of these verification methods is not without issue; whilst credit card checks are
489 appropriate for children, they pose a problem for those aged between 13 years and 18 years who are considered adults by
490 the DPA but cannot – in the UK – legally hold a credit card with which to verify their age. For reference, the section in the
491 DPA which deviates from GDPR for the UK is shown in Table 1.

492

493

Table 1: excerpts from DPA Section 9

‘Child’s consent in relation to information society services’: In Article 8(1) of the GDPR (conditions applicable to
child’s consent in relation to information society services)
(a) references to “16 years” are to be read as references to “13 years”, and
(b) the reference to “information society services” does not include preventive or counselling services.

494

495 **Question 3: Is there a way of ensuring the person with parental responsibility has provided consent for a child’s**
496 **interaction with the device?**

- 497 • (Good) Yes – by full authorisation
- 498 • (Average) Yes – by optional ‘parental’ mechanisms
- 499 • (Poor) No mechanism present for giving parental consent

500

501 This question asks if it is possible that when a child is using the device after its initial setup, a parent can be assumed to
502 have given consent. There are ways in which this might happen, for example by the use of optional parental controls
503 offering the parent or guardian the ability to limit when the child uses the device or service.
504

505 **Question 4: Are there any parental controls?**

- 506 • (Good) Yes – fine-grained control on all devices
 - 507 • (Average) Yes – some control, or only on certain devices
 - 508 • (Poor) No parental controls present
- 509

510 The ICO’s age-appropriate design code offers useful insight into parental controls that can be “*used to support parents in*
511 *protecting and promoting the best interests of their child*” [25]. Does the vendor offer any controls and, if so, do they
512 operate across all devices on which their VA application might reasonably be expected to be used?
513

514 **Question 5: Are the parental controls made available with good accessibility for users?**

- 515 • (Good) Yes – clear instructions signposted in online support
 - 516 • (Average) Yes – but the information is difficult to find
 - 517 • (Poor) No – there is no information given regarding the controls
- 518

519 Parental controls are of little utility if they are hard to operate, or information explaining how they work is difficult to find.
520 For the purposes of this question, ‘difficult to find’ means the information is not clearly available from the vendor’s online
521 support. There is a further nuance – the child whose access is being controlled has, under GDPR’s edict that personal data
522 shall be “*processed lawfully, fairly and in a transparent manner in relation to the data subject*”, a right to know how the
523 controls are affecting their use of the service. The ICO’s code suggests that, for children aged between 13-15 years, the
524 vendor’s information should also clearly explain this.

525 *3.2.4 Unrequested Processing*

526 **Question 1: Is there evidence that a mistake could be made, confusing a spoken expression for the VA wake word?**

527

528 **Question 2: Is there evidence that VAs mishear their wake words, leading to accidental recordings?**

529

530 **Background:** VAs are activated by a wake word, after which an indication is given that it is ready for the user to interact
531 with it. To know if the wake word has been spoken by the user, the device needs to be constantly aware of the sounds being
532 made near it. A VA should not be recording or sending any information to its cloud server, however, until the wake word
533 has been spoken. Each device has a wake word; Alexa offers the facility to change the word from a predefined selection.
534 The words are ‘Alexa’, ‘Echo’, ‘Computer’, ‘Ziggy’, or ‘Amazon’ for Alexa devices; ‘Hey Siri’ for Apple devices; ‘Hey
535 Google’ for Google devices; ‘Hey Cortana’ for Microsoft devices.

536

537 Should the device mishear the wake word, it might activate and start sending audio to the cloud for recording without the
538 user’s knowledge – a clear privacy breach. VAs use audible alerts and visual indicator lights to mitigate the chance of this
539 happening without the user’s knowledge, but these only alert the user to the fact that a recording is being made, they do

540 not prevent it. The results of these questions are not presented in a chart but simply answered ‘yes’ or ‘no’ with evidence,
541 if applicable, to assert the answer.

542 3.2.5 Data Repurposing

543 According to a report made by Reuters Practical Law, ‘Big Data’ relies on three things: aggregation (size – vast volumes,
544 shape – text, sound); analysis (datasets are analysed in real time); and increasing value (enhancing competitiveness and
545 efficiency) [43]. Data is not in short supply for these companies and their VAs are adding to it; these questions ask if the
546 VA companies explicitly state how they repurpose or share this information.

547

548 **Question 1: Are the purposes of any data profiling explicitly stated?**

- 549 • (Good) Yes – examples given covering those explicitly stated in GDPR
- 550 • (Average) Yes – generic classifications only, and/or incomplete coverage of purposes
- 551 • (Poor) No examples or purposes of data profiling given

552

553 As distinct from what GDPR calls ‘processing’, which can be as simple as the collection of the data in the first instance,
554 profiling refers to the manipulation and mining of the data to infer the characteristics of the user. A common use for this is
555 targeting advertising, where the user’s interests have been built into a profile that matches that of a seller’s target – someone
556 who may be susceptible to buying the seller’s product.

557

558 **Question 2: Is the user’s data – according to the vendor’s policies – shared with other entities outside of the 559 organisation?**

- 560 • (Good) Yes – with explanations of what is shared, why, and with whom
- 561 • (Average) Yes – no explanation given
- 562 • (Poor) No – the user’s data is not shared outside the organisation

563

564 Recital 6 of GDPR highlights sharing of data as a growing area of concern, and one of the drivers in the introduction of
565 the regulation.

566 3.2.6 Data Retention

567 **Question 1: Can users find out how long data will be stored?**

- 568 • (Good) Yes – with specified timescales
- 569 • (Average) Yes – without specified timescales but within parameters of certain events
- 570 • (Poor) No – users are unable to find out how long their data will be stored for

571

572 GDPR is specific about how long user data should be kept for. Exceptions are made for cases where users’ data is processed
573 for archiving purposes in the public interest, for scientific or historical research purposes, or for statistical purposes. As it
574 is unlikely that Amazon, Apple, Google, or Microsoft are engaged in these activities, they are obliged to keep it for no
575 longer than they need to process it.

576

577 **Question 2: Is it possible for a user to delete their voice recordings?**

- 578 • (Good) Yes – clearly signposted in online support

- 579
- (Average) Yes – not clearly indicated in help guides
- 580
- (Poor) No – users are unable to delete their own voice recordings
- 581

582 When a user interacts with a VA, a recording is made of their voice and sent to the vendor's server for processing. These
583 recordings are kept, in the form of an audio file, alongside the user's account. It is important to note that it is not possible
584 to prove that a recording has been deleted – with this question, we are taking the vendor at their word. Even if a recording
585 appears to have been deleted, the vendor may have simply removed it from visibility of the user.

586

587 **Question 3: Does the delete function remove all data (transcriptions) or just voice?**

- 588
- (Good) Yes – all data
- 589
- (Average) Voice data only
- 590
- (Poor) Some voice data cannot be deleted
- 591

592 Alongside the audio recordings of the user's voice interaction, a text translation is made by the vendor's speech recognition
593 software. However, as is shown in the 'Repurposing of Data' section, the user's information is not just used for responding
594 to queries. The data is used for advert profiling, 'personalisation', and any manner of other purposes; as long as the user
595 still consents to these practices, the providers are entitled to store the data.

596

597 **Question 4: Does the provider offer 'The Right to be Forgotten'?**

- 598
- (Good) Yes – clearly signposted with a selection of contact routes (verbal, writing)
- 599
- (Average) Yes – limited means of request
- 600
- (Poor) No – the provider does not offer the right to be forgotten
- 601

602 GDPR and the DPA offer what is called 'The Right to be Forgotten' which obliges the controller, when requested by the
603 user, to erase the user's personal data '*without undue delay*'. This can be triggered in several ways, for example where the
604 processing of the data is found to be unlawful or if there is a national or EU legal obligation to do so. Where VAs are
605 concerned, the important reason is when '*The data subject withdraws their consent and the controller has no other
606 legitimate ground for the processing of the data.*' In other words, when the user has decided that they no longer want the
607 provider to keep their data and withdraws their permission for the provider to do so. Users are entitled to be able to make
608 the request (withdraw their consent) verbally or in writing.

609

610 The right to be forgotten made news when, in 2019, Google fought the EU Court of Justice and won a landmark ruling
611 against the French privacy regulator Commission Nationale de l'Informatique et des Libertés (CNIL). The outcome of the
612 case was that Google only had to oblige the user's right to be forgotten in EU countries and not globally. Google argued
613 that they didn't wish to see totalitarian governments forcing their political will on their populations by removing and
614 therefore skewing search results in their favour [44].

615 **3.2.7 Security**

616 Where personal data is concerned, security is paramount in order to ensure the user's data remains private. VA devices
617 themselves have been the target of malicious attacks, as reported by various news agencies [45].

618

619 **Question 1: Is there any access control (authentication, authorisation) to the provider account?**

- 620 • (Good) Yes – credentials and 2FA
- 621 • (Average) Yes – credentials only
- 622 • (Poor) No access control in place

623

624 As it is necessary to create an account with each of the vendors, the security of that account is important. Access to the
625 account could give an intruder personal and private data; moreover, anyone with control of that account could use it to
626 impersonate the original user causing financial and reputational loss – some VAs allow the user to make purchases by
627 voice.

628

629 **Question 2: Is there any access control to the VA device or app?**

- 630 • (Good) Yes – the vendor’s VA is protected on all devices
- 631 • (Average) Yes – the vendor’s VA is only controlled on some compatible devices
- 632 • (Poor) No access control in place

633

634 Question 2 deals with the security on the VA device or application itself as opposed to the security protecting the user’s
635 account. With the proliferation of ways in which one single VA – i.e. Google Assistant – can be used, on smartphones,
636 tablets, and smart speakers, the methods in which the VA may be secured vary. A smart speaker may not have any inbuilt
637 security, allowing it to be used by anyone in its vicinity; however, a VA used on a smartphone may be protected by the
638 phone’s security, in the form of a PIN code or a fingerprint scan. GDPR is unclear on this definition – there is little
639 suggestion of how the controller might be required to implement any security on its endpoint software which has access to
640 its cloud servers.

641

642 **Question 3: Does the provider indicate that encryption is used for the protection of data in transmission or when**
643 **stored?**

- 644 • (Good) Yes – examples of technologies given
- 645 • (Average) Yes – no specific detail provided
- 646 • (Poor) No information given regarding security in transit or at rest

647

648 GDPR’s main focus – when discussing security – is the technologies the vendor uses to protect data in storage and data in
649 transit to ensure the information remains private. Question 3 asks if the vendors are upfront and give examples of the
650 security methods they use, and how specifically those measures are communicated to the user. Problems are further
651 compounded – and GDPR does make specific mention of this – when employees at the vendor are given access to voice
652 recordings [46].

653 *3.2.8 Government Surveillance*

654 **Question 1: Is the user informed if the vendor discloses information when an access request is made by law**
655 **enforcement or government agencies?**

- 656 • (Good) Yes – clearly states the user will be informed, and how
- 657 • (Average) Yes – no detail given
- 658 • (Poor) No – the user is not informed

659

660 In 2016, the UK Government’s then Home Secretary Theresa May introduced the Investigatory Powers Act (IP Act). This
661 act gave UK intelligence agencies (including MI5), and law enforcement, new powers to carry out interception of
662 communications and to collect communications data in bulk [47]. The London School of Economics believed at the time
663 that the IP Act could conflict with GDPR [48]. It will be of interest to see the vendors’ practices in this regard.

664 4 RESULTS AND TESTING

665 In this section, we carry out the research required to answer the questions that form the matrix designed in Section 3:
666 Privextractor Design. The answers to the questions, obtained from the companies’ privacy policies and legal statements,
667 are tabulated and shown as part of the user interface in the final software application. Additionally, we show how the
668 findings from Section 3.1: Forensic Recovery are incorporated into Privextractor’s dashboard.

669 4.1 Comparison Matrix

670 A sample of the matrices discussed in this section can be seen here as displayed in Privextractor’s user interface.

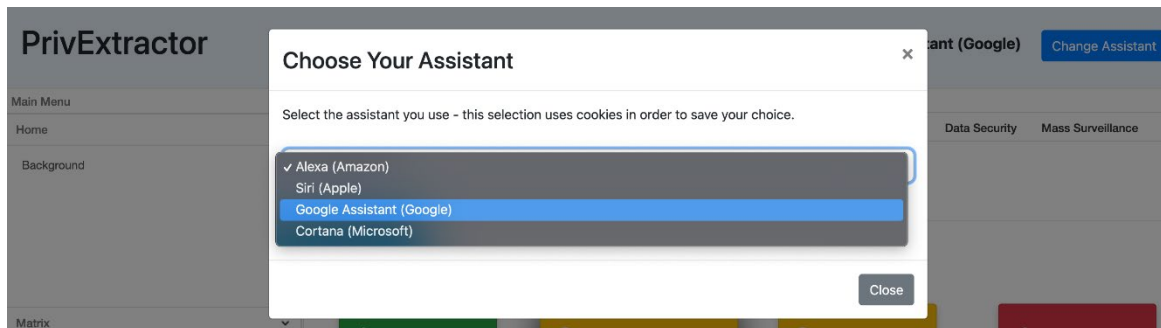
671 4.1.1 Tabulated Results

672 The following sections outline the findings from examining the vendors’ privacy policies, and answering the questions
673 posited in each area of the matrix designed in Section 3: Privextractor Design. In each table, a colour scheme is used where
674 green (happy face) = good, yellow (indifferent face) = average, and red (sad face) = poor. This colour scheme gives an
675 immediate visual indication of the standard of each VA vendor’s policies when asked a specific question. The scores are
676 intended as a comparison of the vendor’s privacy practices; should two VAs achieve the same score, the user can see that
677 choosing one of the VAs means that there is no advantage in either selection.

678

679 Firstly, Figure 3 demonstrates the user selecting their VA – in this case, Google Assistant.

680



681

682

683

Figure 3: user VA selection in Privextractor

684 In Figure 4 it can be seen that the user has selected Google Assistant and the score for that assistant is now highlighted.

685 Below can be seen the matrix for the first question in the ‘Transparency’ section.

686

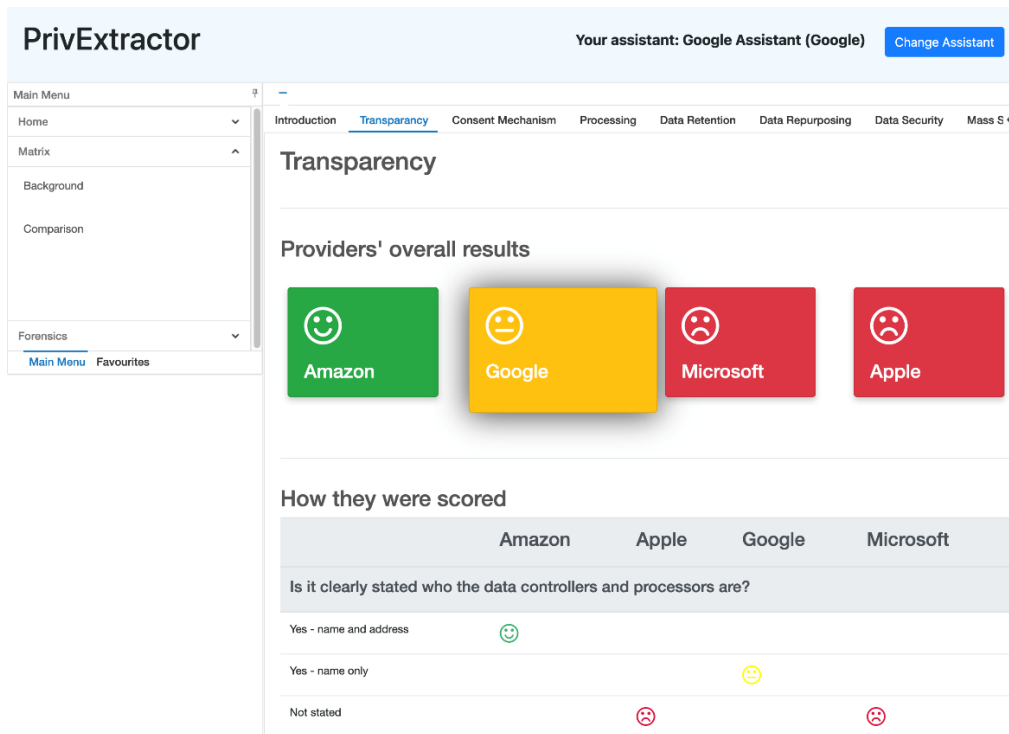


Figure 4: Transparency comparison with Google Assistant selected

687
688
689

690 Finally, in Figure 5, an expanding box has revealed the information used to populate the matrix for the first question.
691

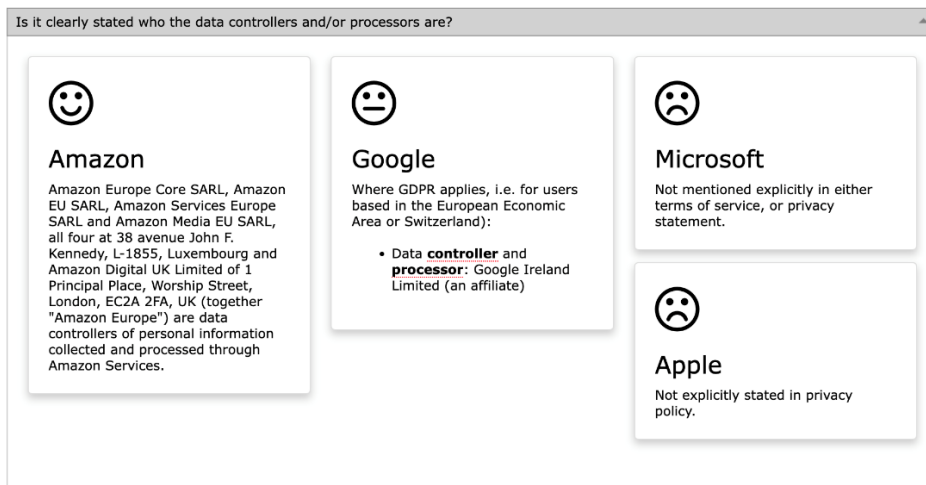


Figure 5: Transparency comparison with Google Assistant selected

692
693
694

695 4.1.2 Transparency

696 Table 2 shows the results of the first section of the matrix in which basic definitions were examined in each vendor’s
 697 privacy policies to compare how open each was about processors and processing, and the types of data that are collected
 698 by each.

Table 2: a comparison of vendors’ transparency

	Amazon	Apple	Google	Microsoft
Is it clearly stated who the data controllers/processors are?				
Yes – name and address	😊			
Yes – name only			😐	
Not stated		😞		😞
Are the types of data processed – such as a user’s name or location data - clearly listed?				
Yes – examples are given covering GDPR		😊		😊
Yes – generic classifications only, incomplete coverage of types stated in GDPR	😐		😐	
No – data types not listed, even in generic form				
Are the purposes of processing clearly listed?				
Yes – examples given covering those explicitly stated in GDPR				
Yes – generic classifications only, incomplete coverage of types stated in GDPR	😐	😐	😐	😐
No purposes given, even in generic form				
Is any processing of biometric data (voice) clearly explained?				
Yes – examples given covering GDPR	😊			
Yes – generic examples only, incomplete coverage of types stated in GDPR				😐
No information about biometric data processing is given		😞	😞	

701 4.1.3 Consent

702 Here we looked at the ability of the VA to respond only to its original user. As per Table 3, only Microsoft’s Cortana had
 703 this feature; however, it should be noted that Microsoft’s disclaimer stated Cortana will ‘try’ to respond only to the user
 704 whose voice it has been trained to recognise.

705
 706 Google offers ‘Family Link’ with which its VA can be trained to respond to a child in the family as part of a wider set of
 707 parental controls. Amazon’s Alexa and Apple’s Siri offer voice training that will personalise their responses to the user
 708 whose voice they recognise; they will not, though, prevent a stranger from conversing with the system.

709
 710

711

Table 3: a comparison of VA consent mechanisms

	Amazon	Apple	Google	Microsoft
Does the device feature a mechanism whereby it processes the data of only a specific individual?				
Yes – data processing limited to a single user at the device level				😊
Only for specific features, or for personalisation	😞	😞	😞	
No mechanism offered – the device will process data of any user who interacts with it				

713 4.1.4 Children

714 This section looks at the vendors' privacy policies compliance not just with GDPR, but with the ICO's Age-appropriate
715 code of conduct. This code was written to cater for a child's use of information services in general, but specifically, those
716 users who are still considered children by GDPR (aged 16 or under) but not by the DPA (13 or under). The ICO make
717 recommendations about age verification, consent and parental controls. The results can be seen in Table 4.

718
719

Table 4: a comparison of VA vendors' practices as pertain to child users

	Amazon	Apple	Google	Microsoft
Does the provider distinguish between adults and children as users?				
Yes – with age explicitly stated	😊	😊	😊	
Yes – no age stated				😞
No distinction was made on the user's age				
If applicable, what form does the age verification mechanism take?				
External verification using endpoints not easily obtainable by children (credit card)				
Basic input of age, with external verification using endpoints easily obtainable by children (email, SMS)		😞	😞	😞
External verification only as a means of two-factor authentication, age not considered	😞			
Is there a way of ensuring the person with parental responsibility has provided consent for a child's interaction with the device?				
Yes – by full authorisation			😊	
Yes – by optional 'parental' mechanisms	😞	😞		😞
No mechanism present for giving parental consent				
Are there any parental controls?				
Yes – fine-grained control on all devices		😊	😊	
Yes – some control, or only on certain devices	😞			😞

No parental controls present				
Are the parental controls made available with good accessibility for users?				
Yes – clear instructions signposted in online support		😊	😊	😊
Yes – but the information is difficult to find	😐			
No – there is no information given regarding the controls				

720 *4.1.5 Unrequested Processing*

721 All four VAs failed this test - research has shown evidence that all four can mishear the wake word, and record without
 722 the user’s consent and knowledge – for example when a user says something that the VA thinks is its wake word, or when
 723 the VA hears something on the television or radio that it mistakes for its wake word [49].

724
 725 Only Amazon allows the wake word to be changed – and then, the choice is limited to one of four options. It is worth
 726 noting that this cannot be done on every device running Amazon’s VA, Alexa [50].

727 *4.1.6 Data Repurposing*

728 As explained in Section 1 - Introduction, the repurposing of data – also known as ‘mining’ – is a profitable business for
 729 the vendors. The vendors’ comparative transparency can be seen in Table 5.

730
 731 Table 5: a comparison of VA vendors’ data profiling policies

	Amazon	Apple	Google	Microsoft
Are the purposes of any data profiling explicitly stated?				
Yes – examples given covering those explicitly stated in GDPR		😊	😊	😊
Yes – generic classifications only, and/or incomplete coverage of purposes	😐			
No examples or purposes of data profiling given				
Is the user’s data – according to the vendor’s policies – shared with other entities outside of the organisation?				
Yes – with explanations of what is shared, why, and with whom		😊		
Yes – no explanation was given	😐		😐	😐
No – the user’s data is not shared outside the organisation				

732 *4.1.7 Data Retention*

733 Data retention – specifically for how long VA vendors keep a user’s information – is a key tenet of GDPR. The vendors
 734 all showed broad compliance, apart from Microsoft (refer to Table 6). GDPR makes a provision for the ‘right to be
 735 forgotten’ which was mentioned in three of the vendors’ policies. GDPR makes no provision, however, for allowing the
 736 interim deletion of data by the user. All but Amazon failed here, and even they allow only the deletion of voice recordings.

737
 738

Table 6: a comparison of vendors' practices regarding data retention

	Amazon	Apple	Google	Microsoft
Can users find out how long data will be stored?				
Yes – with specified timescales			😊	
Yes – without specified timescales but within parameters of certain events	😐	😐		
No – users are unable to find out how long their data will be stored for				😞
Is it possible for a user to delete voice data?				
Yes – clearly signposted in online support	😊	😊	😊	😊
Yes – not clearly indicated in help guides				
No – users are unable to delete their own voice recordings				
Does the delete function remove all data (transcriptions) or just voice?				
Yes – all data				
Voice data only	😐			
Some voice data cannot be deleted		😞	😞	😞
Does the provider offer 'The Right to be Forgotten'?				
Yes – clearly signposted with a selection of contact routes (verbal, writing)	😊			
Yes – limited means of request		😐	😐	
No – the provider does not offer the right to be forgotten				😞

740 4.1.8 Data Security

741 The VA vendors' compliance with GDPR's security requirements is examined and the results shown in Table 7.

742
743

Table 7: a comparison of vendor and application security

	Amazon	Apple	Google	Microsoft
Is there any access control (authentication, authorisation) to the provider account?				
Yes – credentials and 2FA	😊	😊	😊	😊
Yes – credentials only				
No access control in place				
Is there any access control to the VA device or app?				
Yes – the vendor's VA is protected on all devices				
Yes – the vendor's VA is only controlled on some compatible devices	😐	😐	😐	😐
No access control in place				

Does the provider indicate that security is used for the protection of data in transmission or when stored?				
Yes – examples of technologies given				
Yes – no specific detail provided				
No information was given regarding security in transit or at rest				

744 4.1.9 Government Surveillance

745 As seen in Table 8, Microsoft did not mention in their privacy statement whether or not they inform users – in this case, it
746 is assumed they do not. Amazon expressly stated that they do not inform the user.

747 Table 8: a comparison of the vendors’ practices in dealing with access requests
748

	Amazon	Apple	Google	Microsoft
Is the user informed if the vendor discloses information when an access request is made by law enforcement or government agencies?				
Yes – clearly states the user will be informed, and how				
Yes – no detail was given				
No – the user is not informed				

749 **5 DISCUSSION**

750 The results obtained during this research raise several questions. The comparison matrix, ostensibly designed to test the
751 VA vendors’ compliance with data law, has done just that. Whilst there is room for improvement in specific areas of the
752 vendor’s adherence to data law, it has been shown that it is not so much the vendors’ compliance that is of concern but the
753 law itself. GDPR has proved to be quite vague in several areas, meaning that its purpose – to protect the user – is failing.

754
755 Whilst we feel that criticism of the vague nature of the requirements laid out in GDPR and the DPA is valid, it should be
756 pointed out that these regulations are, by necessity, designed to handle a large number of divergent cases of which user
757 data exchanged with a VA is just one. *This does not negate any criticism of under-regulation inherent in current data law,*
758 *and it is clear that GDPR and DPA must be regularly updated. Laws are not concrete and are open to interpretation, but*
759 *they must be considered in a way such that they provide a solid foundation for protecting the user.*

760
761 Whilst complying with GDPR, the vendors are acquiring large amounts of data and are not specifically informing the user
762 what they are doing with it. Despite declaring that they do not ‘sell’ data, the vendors are exchanging information for
763 money via advertising platforms. GDPR could improve in this area and require the vendors to explicitly state how and
764 when this happens, and when they profit, which would improve on the generic caveats given currently in privacy
765 statements. *The understanding of terms by the vendors must be as precise as possible – particularly here, where the terms*
766 *are applied in a specific case. It must not be possible for the companies to take a divergent line in their own legal terms*
767 *and conditions.*

768

769 Advertising is not the only issue. It has been seen that the UK government has previously collected the communications
770 and social media data of its citizens [51]; should they come into possession of the information collected by VAs, this could
771 be considered a worrying breach of privacy. The regulation allows ambiguity in the vendors' outlining how and when
772 information is shared with law enforcement and governments; two of the vendors openly admit that the user will not even
773 be informed when their personal data is shared with a government agency.

774
775 Moreover, GDPR is very specific about requiring user consent without offering any concrete guidance on how this might
776 be obtained by VAs. Again, regulation that deals with divergent cases – as here, where many devices are covered – must
777 be neutral to the technology. However, if consent cannot – for any reason – be effectively given, then users, in particular
778 children, are inevitably going to have their data processed without their consent and having no knowledge of the privacy
779 policies governing their data.

780 **5.1 Privextractor**

781 In order to convey this information to the user of a VA, we developed Privextractor – a web-based dashboard comparison
782 tool. Privextractor contains the information outlined in Section 4.1: Comparison Matrix and – via a mechanism whereby
783 the user can choose their choice of VA – offers an easy reference comparison for the user to decide how the vendor of their
784 VA is complying with data law. This information can help empower the individual to learn how their data is treated, and
785 when and with whom it is shared.

786
787 We envisage this tool to be used in two ways: firstly, as a reference point for a user to select a VA and, secondly, as a tool
788 for the user to reference throughout the time that the user interacts with their VA. We have seen that, whilst vendors are
789 largely compliant with data law, the law itself is not specific enough to enforce transparency on the part of vendors such
790 that users have a full and honest picture of what is happening to their data. A tool that can help redress this balance will
791 enable the user to interact with their VA in greater confidence. Privextractor is, we believe, the first of its kind to offer this
792 facility. We have seen from Section 1.1: Related Work that there are studies dedicated to user perceptions of VA privacy,
793 and that those users are concerned; Privextractor, we hope, will help to address those concerns when a user chooses a VA.

794
795 During the user VA lifetime, a guide to how the user's data is being collected gives a useful overview to the information
796 that the user has shared with the device and, by extension, the vendor. Zibuschka et al. aimed to expose this information in
797 their own dashboard – ENTOURAGE [16]; we feel that the combination of this forensic work, already begun in
798 Privextractor, and the comprehensive overview of the vendor's privacy practices with regard to data law make a useful
799 'one-stop shop' that can act as a reference point for the VA user as long as they use their device.

800
801 Whilst these use cases, we feel, are advantageous to PrivExtractor's target audience – end users – there are also
802 shortcomings of such a system. The four vendors whose VAs are studied here – Apple, Amazon, Microsoft, Google – are
803 large and dynamic organisations whose legal terms and privacy policies are likely to change regularly. Privextractor, in its
804 current state, cannot dynamically accommodate such alterations to this source material in which it is based. Moreover, its
805 forensic capability is – at present – limited to what can be thought of as a laboratory experiment, due to the complications
806 with locating and identifying the necessary security token required to authenticate to the vendors' cloud services.

807

808 PrivExtractor’s main reason for existence is to assist users in making informed choices about their use of a VA and, as
809 such, its utility for practitioners such as regulatory bodies and the vendors themselves is, on first inspection, limited.
810 However, a study by Emami-Naeini et al. determined through interviewing users that there are many considerations a user
811 makes when purchasing a VA, amongst which privacy and security concerns were only mentioned by a few participants
812 [52]. Post-purchase, the number of participants reporting security and privacy concerns rose to around half of the total
813 number of interviewees. The authors proposed the use of a label, to be attached to the device at point of sale, containing
814 “...ratings from an independent privacy lab, an independent IT security institute, and Consumer Reports (CR).”
815

816 Should a regulatory body start to demand such transparency, a tool such as PrivExtractor could work in tandem with such
817 a system with the labelling offering a useful insight into the VAs security and privacy rating at point of sale, with
818 PrivExtractor as a companion throughout the life of the VA. Projects such as Polisis [18] which aim to automate privacy
819 policy analysis, and Privacy Flag [17] which assess a user’s smartphone apps for privacy risk, could be further systems
820 with which PrivExtractor might work. A painstaking and time-consuming part of this research was the manual analysis of
821 privacy policies, and a sophisticated automated system such as Polisis would be a great asset. Furthermore, VAs – as we
822 have seen – exist not only as standalone devices, but as smartphone applications. Privacy Flag’s important work in
823 determining the privacy risk of smartphone apps could be a very useful system with which PrivExtractor may interact. In
824 addition, we would be happy to share usage statistics and user feedback from the use of PrivExtractor with VA vendors
825 should they wish.
826

827 In academic research terms, we saw in Section 1.1: Related Work how Ford et al. attempted to analyse the traffic that is
828 exchanged between Amazon’s Alexa VA and its cloud platform [35]; unable to decrypt the TLS-encrypted traffic itself,
829 the authors had to resort to observing patterns in the quantity and timing of the data that passes across the network.
830 Privextractor continues the forensic work that has been carried out in previous studies [36] [37] [38]; our forensic work
831 does not offer anything new, but reinforces existing studies which helps in an understanding of how VAs handle data and
832 what the vendors are storing. We also saw in the literature review how many academics are interested in user perceptions
833 of VAs, and how they view these devices in privacy terms. We see Privextractor as a useful tool to help in this area, by
834 addressing one of our goals of redressing the imbalance of understanding between the user and the VA vendors.

835 5.2 Research Questions

836 In this section, each of the research questions is addressed individually.
837

838 **RQ1: If a user of a voice assistant wishes to inform themselves about the extent to which their personal information**
839 **is harvested by the vendor of their chosen VA, does that vendor clearly and unequivocally state the exact nature of**
840 **the information that they collect, how securely they keep that data, what they are doing with it, and for how long**
841 **they keep it?**
842

843 As can be seen in section 4.1: Comparison Matrix, where the results of this part of the research are described in full, the
844 vendors of virtual assistants (VAs) are largely compliant with data law and any deviations from the strict rule are minimal.
845 It is where GDPR itself becomes less clear that a corresponding lack of clarity is found within the vendors’ privacy policies.
846 For example, Google’s privacy statement appears quite specific in defining what the company does not share – “*We don’t*

847 *show you personalized ads based on sensitive categories, such as race, religion, sexual orientation, or health”* is one
848 example.

849

850 As described in Section 1: Introduction, Google makes the majority of its money by brokering online advertising [6]. The
851 user, however reassured that they will not be shown adverts based on their race, for example, might still like to know the
852 following:

853

- 854 • What data *do* you share?
- 855 • Exactly when do you share the data?
- 856 • With whom?
- 857 • Is the sharing for profit?

858

859 Amazon, similarly, gives plenty of information in its privacy policy that suggest it complies with GDPR’s requirements;
860 however, there are apparent contradictions. The company made over US\$10 billion in ad revenue in 2019 [53] despite the
861 claim made in their privacy policy that they “*are not in the business of selling our customers’ personal information to*
862 *others*”.

863

864 More clues are given in Amazon’s ‘Interest-Based Ads’ policy in which the company tells the user that they ‘*work with*
865 *third parties*’; from this, it is not clear if they mean ‘*share data with*’. These examples are in accordance with GDPR and,
866 as such, neither Amazon nor Google is in breach of the law. The lack of information that is conveyed to the reader, though,
867 suggests that in certain areas GDPR is not providing the overarching protection of the user that it is intended to.

868

869 **RQ2: Using UK and EU data law as a basis – GDPR, the DPA and the ICO’s age-appropriate code of conduct – can**
870 **it be demonstrated to a user where the data collection practices of the VA vendor conflict with the law?**

871

872 All four vendors perform reasonably well when questioned on their data collection practices in terms of what is required
873 to be outlined by GDPR. Some vendors perform better than others in certain areas – they provide more detail, but GDPR
874 as it stands is being adhered to. Despite this, there do remain some areas of concern, chiefly around the way the vendors
875 handle security, authorisation, and consent.

876

877 In general, there are two points where security might be a concern in the use of a VA; the first of which is the security of
878 the vendor-hosted service account that it is necessary to create in order to use the VA, and the second is the security of the
879 VA application itself. Both of these must be robust in order to ensure that any information exchanged with the device and
880 the vendor remains private. Transparency regarding account security is handled well by all the vendors, as is the security
881 itself: all vendors offer two-factor authentication, and these accounts are well-protected against a malicious threat actor
882 wishing to gain illicit access to a user’s personal information. There is some opacity in the information provided by the
883 vendors regarding the means of securing cloud data once it is in the possession of the vendor, but some information is
884 given.

885

886 Where all four vendors fall is access control to their VA application across devices. There is simply no way, on any of the
887 four vendors’ VAs, of ensuring that – at all times – consent has been given by the person interacting with the client

888 application. Whereas the individual who initialised the device and created the cloud service account has given consent
889 during the signup process, any further user who interacts with the VA has not. This is less of a problem when the VA is
890 used on a smartphone or tablet, as many of these offer device-level security such as PIN codes or fingerprints to access.
891 Smart speakers, however, do not.

892
893 Whilst a concern for any user, lack of consent and authentication mechanism becomes more problematic when the
894 subsequent user is a child. Amazon, for example, clearly states *“If you're under 18, you may use Amazon Services only
895 with the involvement of a parent or guardian.”* The ICO’s view is that children aged between 16 and 17 years *“are still
896 developing cognitively and emotionally and should not be expected to have the same resilience, experience or appreciation
897 of the long term consequences of their online actions as adults may have.”* [25]

898
899 All vendors distinguished children from adults in their privacy policies, with only Microsoft failing to state the age of what
900 it considers a child. All four vendors offered parental controls; Google’s controls in particular – ‘Family Link’ – are fine-
901 grained and offer voice fingerprinting as a mechanism of making sure the child is correctly authenticated.

902
903 There is no mechanism preventing a child from signing up for an account to use any of the assistants that they could not
904 easily circumvent. Other options available to the vendors, such as making a user confirm their adulthood via a credit-card
905 check, have been the subject of much debate by owners of pornography websites [54] so it is perhaps unfair to expect
906 similar implementation for a service such as a VA that a child may be reasonably expected to use. In the UK it is illegal to
907 obtain a credit card until the age of 18 which – given the age of an ‘adult’ in terms of some of the VA vendors is a maximum
908 of 16 years – presents a further issue.

909
910 Other checks recommended by the Information Commissioner’s Office for verifying the age of a user include artificial
911 intelligence or ‘hard identifiers’ such as a passport [25]. These could be considered equally obtrusive, further eroding
912 privacy; relying on mandatory confirmation of an adult account holder would appear to be a better compromise.

913
914 The fact remains, however, that VAs gather a lot of data – voice, geolocation – that when taken from a child could present
915 a safeguarding risk [55]. Whilst compliant with data law, simply relying on a self-declaration is insufficient to mitigate
916 this risk and would appear to represent a shortcoming in the law itself.

917 **5.3 Conclusion**

918 Privextractor is a novel proof-of-concept application that is capable of highlighting to a user the strengths and weaknesses
919 of a chosen VA. From a review of the literature this has not previously been reported; there has been little work undertaken
920 on a self-contained user awareness tool specifically targeting virtual assistants. Such a tool could significantly increase VA
921 users’ understanding of the privacy and security issues surrounding the use of an assistant.

922
923 The outcomes of the work are interesting – we started the research with an open mind, and did not know what to expect.
924 Two possibilities were that the vendors were a) complying with data law and there was no problem, or b) were not
925 complying with data law causing an obvious legal issue. Curiously, the outcome was, strictly speaking, neither – the
926 vendors are in compliance. However, the more we researched and studied this area, the more it became apparent that data
927 law such as GDPR is not specific enough to allow the user to make an informed choice in the VA market should they be

928 concerned about privacy. Ultimately, the law has to cover a lot of different cases and cannot be too specific – but we feel
929 a lack of specificity in what are quite tightly-defined areas (“We do not sell your data”) is allowing the VA vendors too
930 much latitude at the user’s expense.

931
932 Previous studies have examined user awareness and acceptance factors of VAs [26] [27]. Studies have also been made on
933 the forensic recovery of information from Amazon’s cloud service, work which resulted in a functional web application
934 [37]. However, there have been no studies that combine data law and compliance in the context of redressing the imbalance
935 or privacy between user and vendor.

936
937 In the introduction section, it was noted that Linden et al. (2020) observed that “*many [vendors’ privacy] policies still do*
938 *not meet several key GDPR requirements or their improved coverage comes with reduced specificity*” [12]. This, in a way,
939 can still be shown to be true – certainly in terms of reduced specificity. However much the vendors improve, though, there
940 is still a fundamental problem: GDPR, and its UK counterpart the DPA, do not specify any requirement for greater
941 transparency in how the vendors are using data for brokering advertising. Future study into the ways in which current data
942 law such as GDPR appears to be lagging behind the rapid uptake of VAs and, in particular, the use of the data therefrom
943 in the advertising industry could be of great benefit to the end user.

944
945 Privextractor, in its current form, is a proof of concept. Future work in the form of a comprehensive study of the way in
946 which the vendors’ APIs, if they exist, would give Privextractor the ability to perform more comprehensive forensic
947 extraction, for example; something that could demonstrate to the user exactly how their data is stored. A further research
948 direction could work towards a tool that could demonstrate to the user how voice interactions with their VA can influence
949 targeted advertising; this would be of great help in demonstrating to the user the value of their data.

950
951 VA manufacturers and vendors are likely to make changes to their privacy policies and statements as circumstances detect;
952 as this happens, the information within Privextractor will become out-of-date and unreliable. The ideal goal, in this
953 instance, would be for the dashboard to include a form of automatic updating – as a minimum, the tool should be aware
954 that changes to the statements have been made. An interesting future work direction could focus on how Privextractor
955 achieves this; real-time updates of the content within Privextractor based on the contents of the new privacy policy will be
956 more of a challenge but would increase the utility of the tool, and the trust placed in it by users wishing to base their
957 decisions on the information contained therein. In a similar vein, Privextractor might be made more useful with the addition
958 of other VA vendors; we selected the four used in this study by market share but they are by no means the only VAs in use
959 today.

960
961 Finally, as we have concluded that GDPR’s – and, by extension, the DPA’s – vagueness is failing users, we must address
962 the ways in which it may reform. The California Consumer Privacy Act (CCPA) specifically mentions the ability for the
963 user to opt out of having their personal data sold [56]. We have seen, however, that the specific way in which online
964 advertising is brokered does not necessarily constitute a sale; rather, the data is exchanged on the basis that money may
965 change hands later down the line if specific transactional parameters are met (a ‘clickthrough’). The CCPA is right, then,
966 to address this but we may find that it has little effect on the distribution of a consumer’s data and the ability for the user
967 to know exactly where their private information is ending up.

968

969 Data laws must cover a lot of bases in one legislation and making them too specific might be detrimental in other,
970 unexplored ways. We feel that it is important that future work looks into ways in which the law may walk the line of being
971 general enough to protect in all cases, but specific enough that gaps are not there to be exploited. One issue with data law
972 is the rate at which technology advances; the current, huge rise in VA adoption took place in a few short years and any
973 future data law must be prepared for the ‘next big thing’ that may open up whole new areas of concern for data privacy.

974 6 REFERENCES

- 975 [1] Statista The 100 largest companies in the world by market capitalization. *Statista* (2022).
976 [2] Amazon Amazon Advertising. *Amazon* (2022).
977 [3] Apple Apple Search Ads. *Apple* (2022).
978 [4] Google Google Ads. *Google* (2020).
979 [5] Microsoft Microsoft Advertising. *Microsoft* (2022).
980 [6] Clement, J. Advertising revenue of Google from 2001 to 2021. *Statista* (2022).
981 [7] Gibbs, S. Google Nest Audio review: smart speaker gets music upgrade. *The Guardian* (2020).
982 [8] V M Radhika, A. T., M Abdul Nizar An enhanced model for behavioral targeting in online advertising. *2016*
983 *International Conference on Data Science and Engineering (ICDSE)* (2016).
984 [9] Hoy, M. B. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services*
985 *Quarterly* (2018).
986 [10] Edith G.Smit, G. V. N., Hilde A.M. Voorveld Understanding online behavioural advertising: User knowledge,
987 privacy concerns and online coping behaviour in Europe. *Computers in Human Behaviour*, 32 (2014), 15-22.
988 [11] Song Liao, C. W., Long Cheng, Hongxin Hu, Huixing Deng *Measuring the Effectiveness of Privacy Policies for*
989 *Voice Assistant Applications*. City, 2020.
990 [12] Thomas Linden, R. K., Hamza Harkous, and Kassem Fawaz The Privacy Policy Landscape After the GDPR.
991 *Proceedings on Privacy Enhancing Technologies* (2020), 47-64.
992 [13] Kinsella, B. *UK Smart Speaker Adoption Surpasses U.S. in 2020 – New Report with 33 Charts*. City, 2021.
993 [14] Tankovska, H. Number of digital voice assistants in use worldwide from 2019 to 2023 (in billions). *Statista* (2020).
994 [15] Mondal, V. S. a. M. *Understanding and Improving Usability of Data Dashboards for Simplified Privacy Control of*
995 *Voice Assistant Data*. City, 2022.
996 [16] Zibuschka, J. A. H., Moritz AND Kubach, Michael *The ENTOURAGE Privacy and Security Reference Architecture*
997 *for Internet of Things Ecosystems*. City, 2019.
998 [17] Commission, E. *Privacy Flag*. City, 2022.
999 [18] Aberer, H. H. a. K. F. a. R. L. a. F. S. a. K. S. a. K. *Polis: Automated Analysis and Presentation of Privacy Policies*
1000 *Using Deep Learning*. City, 2018.
1001 [19] Erdos, D. Dead Ringers? Legal Persons and the Deceased in European Data Protection Law. *University of*
1002 *Cambridge Faculty of Law Research Paper No. 21/2020* (2020).
1003 [20] Office, I. C. s. Our history. *Information Commissioner’s Office* (2018).
1004 [21] Wolford, B. What is GDPR, the EU’s new data protection law? *GDPR.EU* (2020).
1005 [22] EUR-Lex The General Data Protection Regulation. *EUR-Lex* (2022).
1006 [23] ICO *The UK GDPR*. City, 2022.
1007 [24] Government, H. Data Protection Act 2018. *gov.uk* (2018).
1008 [25] Office, I. C. s. Introduction to the Age appropriate design code. *Information Commissioner’s Office* (2022).
1009 [26] Josephine Lau, B. Z., Florian Schaub sac Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-
1010 seeking Behaviors with Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction* (2018).
1011 [27] Laura Burbach, P. H., Nils Plettenberg, Johannes Nakayama, Martina Ziefle, André Calero Valdez x[S1] “Hey,
1012 Siri”, “Ok, Google”, “Alexa”. Acceptance-Relevant Factors of Virtual Voice-Assistants. *IEEE Xplore* (2019).
1013 [28] Frank Ebbers, J. Z., Christian Zimmermann, Oliver Hinz User preferences for privacy features in digital assistants.
1014 *Electronic Markets*, 31 (2021), 411-426.
1015 [29] William Seymour, M. K., Reuben Binns, Max Van Kleek *Informing the Design of Privacy-Empowering Tools for*
1016 *the Connected Home*. City, 2020.
1017 [30] Markus Lennartsson, J. K., Marcus Nohlberg Exploring the meaning of usable security – a literature review.
1018 *Information and Computer Security* (2021).

1019 [31] Chen Yan, X. J., Kai Wang, Qinhong Jiang, Zizhi Jin, Wenyuan Xu A Survey on Voice Assistant Security: Attacks
1020 and Countermeasures. *ACM Computing Surveys* (2022).

1021 [32] Emily McReynolds, S. H., Timothy Lau, Aditya Saraf, Maya Cakmak, Franziska Roesner Toys that Listen: A
1022 Study of Parents, Children, and Internet-Connected Toys. *CHI '17: Proceedings of the 2017 CHI Conference on Human*
1023 *Factors in Computing Systems* (2017).

1024 [33] Lauren N. Girouard-Hallam, H. M. S., Judith H. Danovitch Children's mental, social, and moral attributions toward
1025 a familiar digital voice assistant. *Human Behaviour and Emerging Technologies* (2021).

1026 [34] Yousra Javed, S. S., Akshay Jadoun x[S13] Alexa's Voice Recording Behavior: A Survey of User Understanding
1027 and Awareness. *ARES '19: Proceedings of the 14th International Conference on Availability, Reliability and Security*
1028 (2019).

1029 [35] Marcia Ford, W. P. Alexa, are you listening to me? An analysis of Alexa voice service network traffic. *Personal and*
1030 *Ubiquitous Computing*, 23 (2019).

1031 [36] Alex Akinbi, T. B. Forensic Investigation of Google Assistant. *SN Computer Science* (2020).

1032 [37] Hyunji Chung, J. P., Sangjin Lee Digital forensic approaches for Amazon Alexa ecosystem. *Elsevier*, 22 (2017),
1033 S15-S25.

1034 [38] Bhupendra Singh, U. S. A forensic insight into Windows 10 Cortana search. *Computers & Security* (2017), 142-154.

1035 [39] Horsman, G. Loose-Lipped Mobile Device Intelligent Personal Assistants: A Discussion of Information Gleaned
1036 from Siri on Locked iOS Devices. *Journal of Forensic Science*, 64 (2019).

1037 [40] Lareo, X. TechDispatch #1: Smart Speakers and Virtual Assistants. *European Data Protection Supervisor* (2019).

1038 [41] Irwin, L. What is an ISO 27001 risk assessment and how should you document the process? *itgovernance.eu* (2020).

1039 [42] Wood, S. Blog: Using biometric data in a fair, transparent and accountable manner. *Information Commissioner's*
1040 *Office* (2019).

1041 [43] Kemp, R. Big data and data protection (GDPR and DPA 2018). *Reuters Practical Law* (2020).

1042 [44] Samonte, M. Google v CNIL Case C-507/17: The Territorial Scope of the Right to be Forgotten Under EU Law.
1043 *European Law Blog* (2019).

1044 [45] BBC Amazon Alexa security bug allowed access to voice history. *BBC News* (2020).

1045 [46] Cook, J. Amazon employees listen in to thousands of customer Alexa recordings. *The Daily Telegraph* (2019).

1046 [47] GCHQ Investigatory Powers Act. *GCHQ* (2019).

1047 [48] Economics, L. S. o. Could the European GDPR undermine the UK Investigatory Powers Act? *London School of*
1048 *Economics* (2016).

1049 [49] Daniel J. Dubois, R. K., Anna Maria Mandalari, Muhammad Talha Paracha, David Choffnes, Hamed Haddadi
1050 When speakers are all ears - Understanding when smart speakers mistakenly record conversations. *20th Privacy*
1051 *Enhancing Technologies Symposium (PETS2020)* (2020).

1052 [50] Amazon Set Up Alexa Hands-Free on Your Phone. *Amazon* (2020).

1053 [51] Waranch, R. S. Digital Rights Ireland Deja Vu: Why the Bulk Acquisition Warrant Provisions of the Investigatory
1054 Powers Act 2016 Are Incompatible with the Charter of Fundamental Rights of the European Union. *George Washington*
1055 *International Law Review* (2018).

1056 [52] Cranor, P. E.-N. A. H. D. A. Y. A. A. L. F. *Exploring How Privacy and Security Factor into IoT Device Purchase*
1057 *Behavior*. City, 2019.

1058 [53] eMarketer Amazon's ad revenue in 2020 is set to grow 23.5% despite the pandemic. *Business Insider* (2020).

1059 [54] Burgess, M. This is how age verification will work under the UK's porn law. *Wired* (2019).

1060 [55] Jenny Radesky, Y. L. R. C., Nusheen Ameenuddin, Dipesh Navsaria Digital Advertising to Children. *American*
1061 *Academy of Pediatrics* (2020).

1062 [56] *California Consumer Privacy Act (CCPA)*. State of California Department of Justice, City, 2022.

1063

1064