Socially-Critical Software Systems: Is Extended Regulation Required?

Niall Dagg[1], Conor Kostick[1], James Fallon[1], Alex O'Neill[1], Murat Yilmaz[2] [0000-0002-2446-3224], Richard Messnarz[3], and Paul M. Clarke[1,4] [0000-0002-4487-627X]

[1] School of Computing, Dublin City University, Dublin, Ireland
{niall.dagg3, conor.mckeon22, james.fallon22, alex.oneill89}
@mail.dcu.ie

[2] Department of Computer Engineering, Gazi University, Ankara, Turkey
my@gazi.edu.tr

[3] ISCN, the International Software Consulting Network, Graz, Austria
rmess@iscn.com

[4] Lero, the Science Foundation Ireland Research Center for Software
paul.m.clarke@dcu.ie

Abstract. Data has become a prevailing aspect of our daily lives, becoming ever more present since the beginning of the 21st century. It is a commodity in today's world and the amount of data being produced has increased enormously. One of the major ways data is produced and collected is from the use of websites and web-based applications. This data is later used for many different purposes. This paper presents findings from a multivocal literature review, exploring the methods of how this data is collected, what the data is used for once it has been collected, the ethics of data and its collection, and the future of data collection. Among the possible futures, we introduce the concept of socially-critical applications, where data harvesting in web-based applications might require pre-market disclosure and evaluation by notified bodies (instructed by regulation) as a means to break the existing cycle of technology companies outpacing under-resourced and ill-equipped regulators. Rather than regulators continually falling short of enacting laws to satisfy the common good, a new class of socially-critical application could be created in law to permit pre-market evaluation of applications (or versions of applications) that could undermine or interrupt the common good.

Keywords: Socially-Critical Software Applications, Data Collection, Web Applications

1 Introduction

Data at its core is a collection of facts, statistics, or items of information. It can be seen as values of qualitative or quantitative items about an object or a person[1]. It is a present and ever-increasing part of our everyday lives. Data can be seen as a new commodity in today's world, sometimes referred to as the *new oil* of the digital economy.[37]

A significant way data plays a role in many facets of society today is through web applications and websites. Web applications are software that run on a web server, unlike desktop-based applications. They are accessed through a web browser and are usually interactive or even just static information sites. Data can be accessed and collected through these web applications and is used for various means [2]. Of course, it is not just web applications that are engaged in data collection, and therefore some of the perspectives examined in this research have broader relevance. In this paper we will investigate the process of data collection, what that data is used for, the ethics of data collection and the future of data collection. The primary aim of this research is to contribute to the understanding of data collection in software-based systems and implications for privacy and ethical consideration.

The remaining part of this paper proceeds as follows. Section 2 outlines the research methodology and identifies four research questions, with Section 3 detailing the major research analysis elements. Sections 4 discusses limitations and future work, with Section 5 presenting a detailed conclusion and examining the possible introduction of a new regulated class of software system: *Socially-critical applications.*

2 Research Methodology & Questions

2.1 Methodology

This research employed a Multivocal Literature Review (MLR) [36], enabling the inclusion of white (peer-viewed) and grey (non-peer-reviewed) literature. Google Scholar and other search engines including IEEE and Springer were utilised.

2.2 Search Queries

For the purpose of this research paper, the terms website and web application are used interchangeably as they both fulfil the same function as a method of data collection. A record of all research terms was created and maintained throughout the life cycle of this research. This record includes the search strings used, keywords and their relevance to the paper. Search strings used included: “Data Collection” “How Websites”, “Data Monetisation” “Web Applications”, “Future”, and “Web Big Data”. To keep search results relevant, we limited the search to papers published from 2015 onward. In total, there were over 70 papers that were initially found, but this number was cut down to 34 through elimination of works not directly related to the core research focus. Initially, the top ten results were considered but this had to be expanded as results were scattered in their relevance. Collective efforts researching data collection in web applications resulted in several research questions.

2.3 Research Questions

The following research questions (RQs) were identified.

RQ1. How is data collected by web applications?
RQ2. How do websites/companies use the data that their applications collect?
RQ3. What are the ethical considerations regarding data collection?
RQ4. What are the possible future directions for web-based data collection?

3 Analysis

Each of the following subsections addresses one of the research questions identified in Section 2.3.

3.1 Data Collection Mechanisms in Web-based applications (RQ1)

Data collection through the use of and exposure of individuals to web applications is a vital part of the modern internet ecosystem and this is reflected in the staggering number of websites and web applications that participate in data tracking and collection [3]. Billions of interconnected devices and services can track and trace close to trillions of transactions and behaviours every day.[4]

There are various methods of collecting this data, traditionally this was achieved via forms, surveys, and registration/logins. Data can also be collected via more subdued/passive techniques such as web-browsing and through the collection of cookies. Cookies are small versions of state management that are stored within a user's browser. They are used to keep metadata based on various factors such as their web browsing habits.[5] The data stored within cookies can be used for a various array of functions, depending on the cookie being implemented. There are *Persistent Cookies*, that can store information for a long time. These are contrasted with *Session Cookies*, that can only store information for as long as the user is on the web page in question (it is deleted upon exit). These two types of cookies are usually used for functions such as the saving of login information, or the saving of the contents of a shopping cart on an e-commerce site.[6] Web-cookies exhibit further interesting characteristics. They can then fall into the categories of *First Party* and *Third-Party* cookies. The former is based on the website you are visiting. The latter come from a third-party website and usually have an ulterior motive, such as advertising.[7]

Cookies are typically accepted by the user using consent forms which are used to convey the information that will be collected from the user as they browse and use the website. There is some doubt regarding whether these consent forms faithfully implement the General Data Protection Regulation (GDPR) [8] which came into effect in May 2018. User literacy around cookies may be somewhat lacking, which can result in a user unconsciously handing over more of their data than they might realise. A recent phenomenon known as *dark patterns in website design* has emerged in recent years.[9] It is the idea that website designers can create a user interface that may influence a user and have them use the website in a certain way, usually to the website's benefit. [9] This can be achieved in various ways, sometimes through malicious interactions with the website itself. The issue of consent and accessibility may make it easier for websites to collect data on users, even if they are not given their explicit consent. A significant concern surrounds the finding that 92% of websites still use some form of tracking without a user's consent.[6] This may be enabled prior to a cookie consent form appearing on screen.

Websites primarily use cookies to access user data, which is a valued commodity. It has been noted that in practice, user data may be as valuable as any other goods or service.[8] This data is used to profile users, to characterise and predict their consumption habits across a wide variety of activities, ranging from audio and video content to e-commerce products, and also social media content. It has been observed that if a product or service is free, it is in fact the user that is the product, and a means to commercialise the user and their data.[9] There are many challenges where user data is concerned. Chief among them is the fine line between respecting a user's privacy and the exploitation of their data. The use of data for monetary gain [10] is embedded in many of the business models of some of the world's biggest web applications. As users are the product, the key challenge is regulating a balance between the protection of users and the monetary gain of their data. The internet and the various opportunities it provides has rapidly become expanded, but individuals interacting with various internet based services may be ill-equipped to manage their own personal data value and risks. Regulators have attempted to address this concern, but this research demonstrates that this remains an evolving landscape.

3.2 Uses For User Data Collected By Web-based Applications (RQ2)

Companies such as Meta (Facebook), Amazon, and Netflix are sometimes referred to as *data-driven companies*.[13] Organisations such as these are more inclined to use data as the cost of storing data has decreased sufficiently, while the ability to process large amounts of data is increasing.[10] Faroukhi and El Alaoui claim that data can be considered a raw material for organisations.[12] It has a sense of value and worth which is acquired through data analysis which leads to so-called *actionable insights*.[12] Through data analysis, organisations can provide data-based insights which can relate to customer interests and habits, as well as advertisement targeting. Organisations can add value for themselves and their customers through these insights, some examples of which include [13]:

- The pharmaceutical distributor Tamro supplies pharmaceutical manufacturers insights about people's spending in relation to their products. They also give pharmacies insights into their sales compared to their competitors.
- Apple's iTunes uses data to improve its service. In the past, music distributors didn't have knowledge of consumer preferences. Now, Apple can tell who bought the music, and determine what type of music that specific consumer is interested in based on their previous purchases. This gives them the ability to profile their consumers.
- Organisations are using user data for advertisement. Companies can target advertising so that their ads are displayed to those who have been browsing similar products online. Advertisers can focus their resources on the people most likely to be interested in the ads. For example, the software company Adara which uses data to provide their partners with customer information. By doing this, the hotels, travel-agents, and airlines that are partnered with Adara can provide highly personalised advertisements to users.

These are just a few of examples where data was used to give actionable insights. Data can be used in many different ways: To personalise user experiences in web applications, to encourage healthier lifestyles, to measure business performance or to enhance web applications to give a competitive advantage. Through these insights, organisations can understand their customers and competitors better which can lead to better data-driven decisions.[13] Organisations seek profits, and through monetisation can generate revenue which can lead to the company making a profit.[11] Data monetisation refers to using data from a company in order to generate revenue and profit. Organisations today are using data assets to generate value, either for internal or external purposes.[10] A company can generate revenue through data monetisation both directly and indirectly.[12] Direct data monetisation refers to selling the data directly to the customers. On the other hand, indirect data monetisation is the selling of products or services that use the data. When a company sells data directly, this can be in the form of raw unstructured data, or it can be the results of data analysis. Examples of direct data monetisation include [13]:

- Vodafone sells its network data to the navigator organisation TomTom. As Vodafone has location-based data on their customers, they can provide this information to TomTom which helps inform about traffic congestions on the roads.
- Toyota sells traffic data to local government authorities to help with infrastructure development.

Wrapping is another method of direct data monetisation which involves firms wrapping data around their products and services. This allows them to stand out from their competitors and provide a product with a greater informational need. This can lead to an increase in revenue generation.[16] One example is software that has data contained in it that can be used by the customer.[11] *Packaging*, a practice that involves the collection of lots of data from different sources and selling it off as a package, has also identified as a method of direct data monetization.[11] This data can be structured or unstructured i.e. data in different formats packaged with the original raw data.[11] Google's *smart thermostat* packages energy, appliance and utility usage and monetises it by selling it to electricity providers [13]. Indirect data monetisation can also be established by selling information-based products and services. Not many companies focus solely on selling data and therefore indirect data monetisation is much more common.[11]

3.3 **Data Collection Ethical Considerations (RQ3)**

In web applications, accepting the privacy policy agreement is usually the only way to access the application. Companies use this privacy policy to communicate how they handle their user data. Many people often overlook the content of the privacy policy and agree with the company's data monetisation practices.[14] It has furthermore been suggested that it is up to the companies that own the web applications to take a moral responsibility for user data and to have privacy policies that act in accordance with the law.[15] Privacy policies are an important aspect of data monetisation. They limit what a company can do with user data and control the amount of data gathered.[11] A company's business model and data monetisation have to be thoroughly thought through, but this can be challenging for companies trying to generate revenue and profits. Data stores can contain sensitive information that requires secure storage. Sensitive data can be a person's full name, date of birth or address. Information security has to prevent access to this data and ensure data integrity [11].

The digital age has brought new challenges for human rights and fundamental freedoms. It raises questions on ethical data and the ethics regarding the collection of data.[19] Ethics covers a system of moral principles, what is good for individuals, and what is right and wrong behaviour [19], and it has been referred to as the *new self-*

regulation.[18] It is tempting for technology firms to focus only on profits, but these firms need to consider the impacts on users, regardless of the profit potential.[17]

The business world is experiencing a rapid transformation regarding the processing of digital information.[19] This causes a big problem with users and the ethics of their data. They lack an understanding about their data, which can lead to a breach in their privacy and exploitation of their personal data. This is frustrated by the lack of technological understanding of law and policy makers [18] who may underestimate how much effort companies invest in profiling users. A big question regarding ethics concerns whose responsibility is it to be ethical? [18]

The GDPR came into force in May 2018 to protect data relating to individuals (‘personal data’).[21] This European regulation demands significant data protection safeguards.[20] These safeguards specify what is ethical in relation to the use of a person's personal identifiable information (PII). PII can be data such as age, gender, and relationship status, and it may refer to information that can be correlated with other information to identify an individual, such as a credit card number or postal code [4]. This is the type of data that GDPR aims to protect. Before GDPR, there was the 1995 Data Protection Directive (DPD) [21]. It was effective for a period of time, but technological and digital consumption advances demanded a revised model of regulation. GDPR places the responsibility of robust privacy rights on the organisations producing the web-based applications. Privacy rights include the *Right to be Forgotten*, the *Right of Access to Data*, the *Right to Data Portability*, and the *Right to Explanation of Automated Decision-Making.*[20] Since GDPR has been introduced, a user can find out what personal data a company has collected about them and for what purpose, which is referred to as the *Principle of Transparency*.[20] A goal of GDPR is that a user is not only entitled to what information is being collected and how it is being used, but that clear and plain language and, additionally, where appropriate, visualisation be used.[22] The right to be forgotten is a crucial element of GDPR and ensures that the user has power over their data. If the customer is not satisfied with the way the company handles their personal data, the customer may request that the company to delete the personal data.[20]

Interface designs that try to guide end-users into desired behaviour through malicious interaction flows are referred to as *dark patterns.*[9] These patterns appear where user value is supplanted in favour of shareholder value.[23] and involve tactics that attempt to nudge users towards desired actions. This may be achieved by making desired actions appear to be part of the user’s task. A user may be made to think that they need to agree to some conditions regarding their personal data simply to access the website. If the online service is influencing users towards the *accept* option, then it is not in compliance with GDPR.[9] The interface design should not influence users using button placement, size, or colour, and it should be ‘as easy to withdraw as to give consent’.[9] If consent was given through a single keypress or keystroke, then revoking consent should be accomplished using the same amount of actions.

Dark patterns raise a large number of ethical concerns. User Interfaces (uIs) that deceive users for the benefit of other parties are known as *dark pattern UIs*.[23] How are dark patterns connected to the collection of user data? They deceive the user into granting a website permission to use personal data in ways that a user may not expect or want. New consent management platforms (CMPs) have been introduced to the web to conform with GDPR, particularly its requirements for consent when companies collect and process users' personal data.[9] GDPR uses two types of constraints to verify correctness, *automated* and others that require *human automation*.[24] Companies can start by building a generic model of the GDPR in relation to their website or application with the help of legal experts.[24] This will help the company understand what constraints will need to be imposed to ensure an ethical standard of personal data collection is upheld. There are directives given for a cookie consent form, but it is noted that many companies do not meet the requirements.[24] A GDPR-compliant cookie consent form must find a balance between two extremes. They must not contain so much information that the user can lose focus and not read the content, but they must also contain enough text for the purpose of the cookie to be understandable and help the user make a well-rounded choice.[24] Constant changes in the external environment have an impact on an organisation and its compliance with GDPR and need to be continuously assessed.[25] Another way to quantify a website's compliance is by seeing if all the data collected is strictly required. Data should be limited to what is necessary for the purposes of the processing [20] in relation to the users' activity.

Website privacy policies are often ignored by users, because these documents tend to be long and difficult to understand.[26] These policies are binding legal agreements between the website operator and the end user. User attention to privacy policies started to rise when 'big tech' companies were brought to court over their privacy policies. For example, Facebook received a $5 billion fine for violation of certain privacy rules.[20] Although this fine is large, it is dwarfed by Facebook's revenue in 2018, which while it was violating its own privacy policy, made $22 billion.[20] Technology companies bring significant benefit their users, this is clearly demonstrated in the widespread adoption of various web-based apps. As this research demonstrates, there is a complex interaction between technology companies, the very useful services their users enjoy, the laws enacted to protect user rights while using these services, and the capacity of users to make informed decisions regarding their own privacy. One further observation from this research is that tracking techniques seem to continually stay ahead in the race, the cookies that were once the focus of such attention, concern and ire, may already be supplanted by "opaque, stateless tracking techniques" such as *browser fingerprinting* [35], a method which can identify users by looking at specific configurations on their devices.

3.4 The Future of Web-based App Data Collection

It is safe to assume that businesses that are currently collecting data via web-based applications will continue to do so as web-based activity is continuing to climb all the time. There are also growing benefits to having large amounts of data collected, most notably in advertising.[28] Meta (previously known as Facebook) and Google, two of the largest firms involved in web-based data collection, make the majority of their revenue from advertising income.[29, 30] What is interesting to note from these two graphs however is that Google, known for its search engine and other online tools, seems to be slowly but gradually increasing its revenue in other areas such as cloud computing. Whereas Meta, known for its social media apps, seems more heavily reliant on its advertising revenue. This could suggest that going forward into the future, companies (such as Google) with the infrastructure to generate revenue from non-advertising means will gradually continue to attempt to grow into those areas thereby decreasing reliance on user data. User data itself might ultimately come to be considered as a type of toxic asset by consumers at some future point if awareness of data usage should rise or if data rights infringements are reported in the media.

Large amounts of data can be characterised by the *3Vs*: volume, velocity, variety. This can also be expanded to the *5Vs* which is the original three plus veracity and value.[31] Volume refers to the raw size of the data being considered. Velocity refers to the speed at which data is generated, processed, and moved on. Variety refers to the different types of data that can be used together to achieve the desired result. Veracity refers to the quality and trustworthiness of the data being used. Value refers to the benefits that can be gained from analysing the data, monetary value, social value, research value, etc. In terms of web-based app usage, it seems plausible to suggest that volume, velocity, variety and value will continue to rise. Veracity is important to consumers of the usage and user data, but it may also relate to privacy and accuracy regarding individual profiling. In this sense, veracity may also rise.

A side-effect of all this data collection will likely be the rise of cloud database systems, specifically NoSQL database systems [31, 32]. NoSQL systems are typically used for big data collection as they are a much more flexible alternative to SQL systems [31], making data entry easier but data retrieval more complicated. NoSQL databases have also proven themselves to be cost effective vs SQL equivalents [32] which of course is a fact many businesses would need to consider; however, another downside of NoSQL databases is that they do not always perform consistently and can struggle with analytics [32]. Another prediction for the future then should be the rise of NoSQL cloud database systems [31, 32] and something to facilitate standardising NoSQL databases to make data retrieval easier and deal with performance consistency issues [31, 32]. New techniques and technologies will be needed to keep up with the amount of big data processing that is being done.[32] There are many technologies that are useful in processing big data that we can expect to continue to grow such as cloud computing, especially with regards to parallel computing, and artificial intelligence (e.g. clustering and predictive analysis) which can be used to make intelligent devices that can help us understand and interpret large amounts of data quickly.

There are further aspects to consider regarding the possible social effects that web-based data collection can have. The danger of bias is ever-present in the collection and analysis of web-based applications, including *population bias* and *activity bias*.[33] Population bias refers to the idea that some online accounts are not a genuine reflection of a person, a website could have bot accounts, throwaway alt accounts or fan page accounts for example. These types of accounts can be accidentally confused for people showing their true selves and skew the statistics. Activity bias can happen if data is collected within a certain time frame and only captures data from accounts that happen to be online at that particular time. Both population and activity bias may be subject to manipulation by less-than-entirely-honest agents seeking personal gain, for example by generating fictitious data to be consumed by the data analysis and associated predictions. Current methods of data collection can be accurate but are also naive in their approach, resulting in “precisely inaccurate” results.[33] One potential solution for this problem involves segmenting the data collected into sections such as recency, frequency, and value. It is important to also note that the data being created rarely tells the full story of an event or action, but is instead “the traces left behind by the use of a large, complex, and constantly evolving software system”.[34] The understanding is therefore necessarily going to be imperfect, but segmentation could help to reduce this problem. Data analysis could lead to understanding society better than before, however it relies on cooperation from the users generating the data and that relies on the users having empathy for the people and companies trying to collect this data, and that unfortunately seems unlikely.[34] This is more-so a human and social limitation than a technological constraint.

4 Limitations and Future Work

Three primary limitations have been identified in this work. First, the topic of web-based application data collection is vast, far too great for one single research effort to cover (especially in the mode of an MLR which looks at both white and grey literature). Further frustrating the first limitation, this research was constrained to just six weeks and was conducted primarily by four final year undergraduate students as part of a software engineering assignment. To address these three primary limitations, the sample of works included was constrained to those that appeared high on the search result listings and which were deemed centrally relevant to this research. An element of researcher judgement is required and this can introduce subjectivity and inaccuracy. By breaking the work up into four distinct research questions, each of the four primary researchers was able to dedicate their focus on one key area of interest. This step made the work more manageable for novice researchers, who were further supported by interaction opportunities and paper writing support from more senior academics. Nevertheless, this work must therefore be considered relatively limited, but for the non-expert interested party, it can provide an up-to-date overview of some of the major themes related web based data collection. Future work should seek to extend this review in order to obtain a more comprehensive evaluation of the problem space.

A major lesson from this research is the acknowledgement that web-based application data collection and subsequent user data analysis is a large and contentious area. Future research efforts will need to work to balance the benefits for end users with the price those users pay for accessing interesting and valuable web-based services. It also seems that regulators will continually be in catchup mode as they do not have the resources and know-how to keep apace of large technology companies. One possible resolution to that particular challenge might involve technology companies disclosing their future technical ambitions and designs to regulators prior to deploying them to the market, with regulators acquiring the powers to reject or suggest modification to proposed innovations pre-market access.

5 Conclusion

This research established four research questions (RQs) to underpin the research focus. Through RQ1, we examined how data is collected by web-based applications, finding that cookies are one of the dominant techniques for data collection and that legal vehicles, such as GDPR, while well-intended and having some effect, are not necessarily faithfully implemented through the use of so-called dark-practices that attempt to lead users to surrender their data (in a somewhat inadvertent user engagement). In RQ2, the research investigated how data collected from web-based applications is used, finding that the associated data analysis seeks to identify actionable insights. Such insights can assist in user profiling, raising knowledge of user preferences and habits. These insights can be triggers for revenue opportunities. RQ3 asks about the ethical considerations surrounding data collection and monetisation, the findings from which suggest that the picture is far from clear. Users do not always know their data rights; some might not seem to care. Firms can introduce features with good intentions, but those features might later be utilised in a manner not previously imagined. Regulators seek to protect the rights of users but struggle with key technical knowledge and resources. Plus, legal elaboration generally takes more time than technological innovation. In RQ4, we examined how data might be collected and utilised in the future. Might the present practice of user data collection ultimately come to be seen as distasteful? At some future point, might data assets and transactions be viewed by consumers as toxic assets? Exploring RQ4 has also confirmed that the ever-growing volume of user data will itself require technological innovation, perhaps parallelisation of processing, and even the segmentation of data to profile the data more efficiently and accurately. The influence of population and activity bias are also discussed, highlighting the imperfections of collected user data and its processing, and the dangers this might manifest.

Having examined the RQs, we looked to possible future directions. In the hotly debated arena of individual rights, company rights and associated legal infrastructure, there does not appear to be an easily identifiable consensus view. We find that regulators, whose role is to represent the common good, are often behind the technology curve. They lack the skills and resources to continuously remain abreast of technology innovations. Furthermore, legal constructs, especially large jurisdictional regulations such as GDPR, take many years to prepare and deploy. Important new questions emerged towards the end of the research: could practices adopted in other regulated domains be of utility to the data protection domain? And could other technology advances such as blockchain be harnessed to improve data privacy?

Those familiar with the safety critical domain, for example automobiles, nuclear and aviation, will be aware that there is a practice of pre-market evaluation/notification prior to market access for certain types of safety-critical products. This practice helps to avoid situations where products are released to market but are later found to be unsafe for consumers. If a new classification of software application was to be created, the *Socially-Critical Application*, then software systems matching this classification might be subject to a type of pre-market evaluation wherein companies would disclose their proposed products or features in advance of market deployment (perhaps in advance of implementation), thereby facilitating inspection of potentially socially damaging applications prior to market access. Applications for market access could identify the intended use of the application/feature and require the involvement of licensed notified bodies as independent evaluators.

There might be an instinctive suspicion and resistance to this idea from technology companies, who may fear heavy bureaucracy or external tinkering in their internal designs. But there could be benefits for the technology companies also: rather than facing regulators in court and risking heavy fines as is presently the case, firms could invest much less and obtain market access in cooperation with regulators. This could demonstrate that they had taken all reasonable steps to produce a socially-responsible system, thereby reducing future claims of reckless behaviour. The introduction of mechanisms such as an *intended use* could also protect the technology firms. Consider the case of an aircraft manufacturer seeking to take a new aircraft to market. They contact the relevant agency/regulator(s), engage in demonstrating that the aircraft is safely designed and built, then obtain market access. If some user of the aircraft employs the craft in a manner for which it was not intended, then the aircraft manufacturer has some protection in that they have clearly identified the intended use. The manufacturer cannot after all protect against all possible unintended use or nefarious agents. For example, an aircraft might be overtasked with a payload that it was not designed to carry. This could be the case for technology companies and their data and features. Features are designed for use in a certain way, data is made available for use in a certain way; agents straying from the intended use would become the guilty party (not the technology firm).

This research paper has been prepared for the 29th EuroSPI conference which has an established interest in the general area of user data and data generation. Earlier EuroSPI contributions (notably a 2019 keynote address from Prof. Hermann Maurer) highlighted a further major data concern: that some organisations (maliciously) create fake data in a structured manner, which upon consumption by current web applications creates a web based (wrong) truth. Such scenarios are designed to create a narrative that can support the ambitions of less-than-completely-transparent entities. This can, for example, enable the creation of new (false) truths about persons, states and intentions. Strengthening data collection and analysis can help to reduce the impact of destabilising forces working contrary to the common good – and adopting a new regulated class of *socially-critical applications* could have a role to play in achieving that objective.

Suggesting that a new classification of software application be legally established in an effort to address the greater social good is a bold move, we therefore only suggest that there might be some value in further discussion and reflection on the concept. It would not be simply achieved, but implemented correctly, it could make life easier and better for regulators and technology firms through fostering collaboration, and for end users, their interests would be upheld as an integral part of that discussion.

Acknowledgements. This research is supported in part by SFI, Science Foundation Ireland (https://www.sfi.ie/) grant No SFI 13/RC/2094 P2 to–Lero - the Science Foundation Ireland Research Centre for Software.

6 References

1. P. S. Pek, “Data Monetis–tion - How an Organization Can Generate Revenue with Data?,” p. 66.
2. “What is a Web Application? | How a Web Application Works.” https://blog.stackpath.com/web-application/.
3. B. Krupp, J. Hadden, and M. Matthews, “An Analysis of Web Tracking Domains in Mobile Applications,” in 13th ACM Web Science Conference 2021, New York, NY, USA, Jun. 2021, pp. 291–298. doi: 10.1145/3447535.3462507.

4. A. A. Alwabel, “Privacy Issues in Big Data from Collection to Use,” in Big Data and Security, Singapore, 2020, pp. 382–391. doi: 10.1007/978-981-15-7530-3_29.
5. A. Dabrowski, G. Merzdovnik, J. Ullrich, G. Sendera, and E. Weippl, “Measuring Cookies and Web Privacy in a Post-GDPR World,” in Passive and Active Measurement, Cham, 2019, pp. 258–270. doi: 10.1007/978-3-030-15986-3_17.
6. I. Sanchez-Rola et al., “Can I Opt Out Yet? GDPR and the Global Illusion of Cookie Control,” in Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security, New York, NY, USA, Jul. 2019, pp. 340–351. doi: 10.1145/3321705.3329806.
7. L. Gröndahl, Public knowledge of digital cookies : Exploring the design of cookie consent forms. 2020. Accessed: Feb. 04, 2022. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-281888
8. E. Papadogiannakis, P. Papadopoulos, N. Kourtellis, and E. P. Markatos, “User Tracking in the Post-cookie Era: How Websites Bypass GDPR Consent to Track Users,” in Proceedings of the Web Conference 2021, New York, NY, USA, Apr. 2021, pp. 2130–2141. doi: 10.1145/3442381.3450056.
9. M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. Kagal, “Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence,” in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–13. Accessed: Feb. 02, 2022. [Online]. Available: https://doi.org/10.1145/3313831.3376321
10. C. C. H. Liu and C.-L. Chen, A Review of Data Monetisation: Strategic Use of Big Data. 2015.
11. J. Fred, “Data Monetis–tion - How an Organization Can Generate Revenue with Data?,” Data Monetisation – Miten organisaatio voi tuottaa liikevaihtoa datan avulla?, 2017, Accessed: Jan. 24, 2022. [Online]. Available: https://trepo.tuni.fi/handle/123456789/24694
12. A. Z. Faroukhi, I. El Alaoui, Y. Gahi, and A. Amine, “Big12onetizationsation throughout Big Data Value Chain: a comprehensive review,” Journal of Big Data, vol. 7, no. 1, p. 3, Jan. 2020, doi: 10.1186/s40537-019-0281-5.
13. M. Laitila, 12onetizationsation: Utilizing data as an asset to generate new revenues for firms,” p. 107.
14. J. Gerlach, T. Widjaja, and P. Buxmann, “Handle with care: How online social network providers’ privacy policies impact users’ information sharing behavior,” The Journal of Strategic Information Systems, vol. 24, no. 1, pp. 33–43, Mar. 2015, doi: 10.1016/j.jsis.2014.09.001.
15. M. J. Culnan and C. C. Williams, “How Ethics Can Enhance Organizational Privacy: Lessons from the Choicepoint and TJX Data Breaches,” MIS Quarterly, vol. 33, no. 4, pp. 673–687, 2009, doi: 10.2307/20650322.
16. S. L. Woerner and B. H. Wixom, “Big Data: Extending the Business Strategy Toolbox,” Journal of Information Technology, vol. 30, no. 1, pp. 60–62, Mar. 2015, doi: 10.1057/jit.2014.31.
17. B. Arogyaswamy, “Big tech and societal sustainability: an ethical framework,” AI & Soc, vol. 35, no. 4, pp. 829–840, Dec. 2020, doi: 10.1007/s00146-020-00956-6.
18. S. Jannick Kirk, H. Van den Bulck, and S. Kosta, “Privacy Policies Caught Between the Legal and the Ethical: European Media and Third-Party Trackers Before and After GDPR.,” Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3427207, Jul. 2019. doi: 10.2139/ssrn.3427207.
19. B. Ibiricu and der M. M. L. van, “Ethics by design: a code of ethics for the digital age,” Records Management Journal, vol. 30, no. 3, pp. 395–414, Jan. 2020, doi: 10.1108/RMJ-08-2019-0044.

20. H. Li, L. Yu, and W. He, “The Impact of GDPR on Global Technology Development,” Journal of Global Information Technology Management, vol. 22, no. 1, pp. 1–6, Jan. 2019, doi: 10.1080/1097198X.2019.1569186.
21. K. Kollnig et al., “Before and after GDPR: tracking in mobile apps,” Internet Policy Review, vol. 10, no. 4, Dec. 2021, doi: 10.14763/2021.4.1611.
22. “The right to be informed (transparency) (Article 13 & 14 GDPR) | Data Protection Commission,” The right to be informed (transparency) (Article 13 & 14 GDPR) | Data Protection Commission. https://www.dataprotection.ie/individuals/know-your-rights/right-be-informed-transparency-article-13-14-gdpr.
23. C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs, “The Dark (Patterns) Side of UX Design,” in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal QC Canada, Apr. 2018, pp. 1–14. doi: 10.1145/3173574.3174108.
24. H. J. Pandit, D. O’Sullivan, and D. Lewis, “Test-Driven Approach Towards GDPR Compliance,” in Semantic Systems. The Power of AI and Knowledge Graphs, Cham, 2019, pp. 19–33. doi: 10.1007/978-3-030-33220-4_2.
25. D. Torre, M. Alferez, G. Soltana, M. Sabetzadeh, and L. Briand, “Modeling data protection and privacy: application and experience with GDPR,” Softw Syst Model, vol. 20, no. 6, pp. 2071–2087, Dec. 2021, doi: 10.1007/s10270-021-00935-5.
26. S. Wilson et al., “The Creation and Analysis of a Website Privacy Policy Corpus,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Aug. 2016, pp. 1330–1340. doi: 10.18653/v1/P16-1126.
27. N. Fabiano, “The value of personal data is the Data Protection and Privacy preliminary condition: synthetic human profiles on the web and ethics,” in Proceedings of ^{t}he 3rd International Conference on Applications of Intelligent Systems, New York, NY, USA, Jan. 2020, pp. 1–5. doi: 10.1145/3378184.3378231.
28. E. C. Malthouse and H. Li, “Opportunities for and Pitfalls of Using Big Data in Advertising Research,” Journal of Advertising, vol. 46, no. 2, pp. 227–235, Apr. 2017.
29. “Meta: quarterly segment revenue 2021,” Statista. https://www.statista.com/statistics/277963/facebooks-quarterly-global-revenue-by-segment/.
30. “Google: distribution of revenue by segment 2021,” Statista. https://www.statista.com/statistics/1093781/distribution-of-googles-revenues-by-segment/.
31. M. Younas, “Research challenges of big data,” SOCA, vol. 13, no. 2, pp. 105–107, Jun. 2019, doi: 10.1007/s11761-019-00265-x.
32. “Big data: From beginning to f–ture - ScienceDirect.” https://www.sciencedirect.com/science/article/pii/S0268401216304753#sec0070 (accessed Feb. 17, 2022).
33. D. A. McFarland and H. R. McFarland, “Big Data and the danger of being precisely inaccurate,” Big Data & Society, vol. 2, no. 2, p. 2053951715602495, Dec. 2015, doi: 10.1177/2053951715602495.
34. R. Shaw, “Big Data and reality,” Big Data & Society, vol. 2, no. 2, p. 2053951715608877, Dec. 2015, doi: 10.1177/2053951715608877.
35. U. Iqbal, S. Englehardt and Z. Sh“fiq, "Fingerprinting the Fingerprinters: Learning to Detect Browser Fingerprinting Beha”iors," 2021 IEEE Symposium on Security and Privacy (SP), 2021, pp. 1143-1161, doi: 10.1109/SP40001.2021.00017.
36. Garousi, V, Felderer, M. and Mäntylä, M.V., Guidelines for including grey literature and conducting multivocal literature reviews in software engineering, Information and Software Technology, Volume 106, 2019, Pages 101-121. ISSN 0950-5849.
37. Data is the New Oil of the Digital Economy, available from https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/ (accessed May. 05, 2022).