ACTA INFOLOGICA 2023;7(1):153-163
dergipark.org.tr/acin

DOI: 10.26650/acin.1231944

RESEARCH ARTICLE

Melting of Privacy with Machine Learning, Big Data, and Social Media

Pelin Canbay[1], Zübeyde Demircioğlu[2]

ABSTRACT

Every individual has the right to keep their information private. However, there is a big question: is this possible in the digital era? While social media attracts people to share personal data, most advanced technologies are continually developing in the area of how to exploit information from this personal data. Is it possible to talk about keeping personal data private? This study aims to investigate whether it is possible both to connect to the cyber-world and remain private in the digital era, where intensive studies have been conducted to protect privacy. This study discusses: (1) the social perception of privacy, (2) the contradiction between privacy expectations and behaviors, and (3) the current state of both disclosure and protection efforts of privacy with machine learning and big data techniques. As a result of our research, it was concluded that it is almost impossible to exist in the cyber/digital world and remain private, that most users are not uncomfortable with the current situation, and that institutions and technology developers should take more responsibility in this regard.

Keywords: Privacy, social media, big data, machine Learning

[1](Assist. Prof. Dr.), Kahramanmaraş Sutcu Imam University, Department of Computer Engineering, Kahramanmaras, Turkiye
[2](Assist. Prof. Dr.), Istanbul Medeniyet University, Faculty of Arts and Humanities, Department of Sociology, Istanbul, Turkiye

ORCID: P.C. 0000-0002-8067-3365; Z.D. 0000-0002-8749-006X

Corresponding author:
Pelin CANBAY
Kahramanmaraş Sutcuimam University, Department of Computer Engineering, Kahramanmaras, Turkiye
E-mail address: pelincanbay@ksu.edu.tr

Submitted: 10.01.2023
Revision Requested: 12.04.2023
Last Revision Received: 25.04.2023
Accepted: 17.05.2023
Published Online: 14.06.2023

Citation: Canbay, P., Demircioglu, Z. (2023). Melting of privacy with machine learning, big data, and social media. *Acta Infologica, 7*(1), 153-163.
https://doi.org/10.26650/acin.1231944

1. Introduction

Although privacy norms vary from culture to culture, society to society, and even from person to person, it is seen as a universal need as old as the history of humanity. Privacy is a right that should be protected and this means that an individual can determine what information or data are collected and analyzed about him/her in any given context. Privacy is one of the most critical concerns of the digital age where almost all personal information and data are stored electronically. Nowadays, people have a digital existence beyond their physical existence, and this digital existence tends to expand constantly. Therefore, people vacillate between the concern of protecting their personal information and the desire to be discovered and recognized (Bauman & Lyon, 2013). The rapid rise of social media allows users to share their personal data unquestioningly.

Social media is a computer-based application that uses digital entities such as texts, voice messages, images, or videos for interaction between individuals. Social media, also called Web 2.0, offers three specific services to users: ease of use, sociability, and uploading/sharing content the way they want. The distribution of data has gained a new form with social media (Fuchs, 2014). In the early days of social media, people who were enthusiastic about interacting, sharing, and collaborating through this application preferred these environments (Correa et al., 2010), and today it has become inevitable that personal data is stored in almost every service such as health and education. The volume and variety of data stored in digital media are increasing day by day. The necessity of managing this massive amount of data increases the use of big data technologies and the need for these technologies. Big data technologies with modern data science techniques, especially with machine learning methods, have the power to integrate data from multiple sources and reveal hidden patterns in them. The data collected from social media have led to many practical applications such as crime detection, recommendation systems, anomaly detection, behavioral analysis, bioinformatics, event detection, business intelligence, relationships, epidemics and opinion, sentiment, and emotions analysis, etc. (T.k. et al., 2021). It is possible to develop tools or applications that provide high benefits to humanity with the analysis of the personal data collected from digital media. Spoken dialog systems, for example, allow machines to help patients easily (Rosenthal et al., 2010; Balcı, 2019) or support therapy for less-abled persons (Matarić et al., 2007). On the other hand, the power of machine learning and big data to integrate data and reveal hidden patterns can have unpredictable negative consequences on data subjects (Kelleher & Tierney, 2018).

Many social and technological studies are carried out to protect the data owner. General Data Protection Regulation (GDPR) (Voigt & von dem Bussche, 2017) is one of the most important measures. The GDPR, established by the EU, legally enforces that the data subject's privacy must be respected. However, only a few governments have legislated the most basic measures. Although many privacy protection methods have been developed, such as anonymity, encryption, and distributed system privacy practices, new approaches are necessary. Since many privacy disclosure techniques have also been developed and used, it is obvious that more than current measures are needed to protect personal information while there are so many obligations to share data.

In this context, this study aims to comprehend how to protect personal privacy with regard to users, governments, and technology developers. According to our research;

- In many digital media, users are obliged to share their data to benefit from the service, while in many applications such as social media they share their personal data voluntarily.
- By using big data technologies and machine learning methods, unexpected inferences about individuals can be found with high success from the data that seem unrelated to each other. On the other hand, technologies developed to protect the privacy of shared and processed data have lagged behind emerging technologies in information extraction.
- It is imperative that the sharing of personal information should not be left to the user merely through the information and permission procedures, that Governments take measures to protect the privacy of users with legislation and follow these regulations, and that technology developers are aware of their responsibility about personal privacy before unexpected violations lead to more significant problems.

The rest of this paper includes the explanation of individuals' digital existence in Section 2. Section 3 discusses the current power of machine learning and big data technologies in regard to privacy. While Section 4 gives the current situation regarding

technology related to privacy, Section 5 provides the users' privacy attitudes and behaviors. In Section 6, the conclusion of the study is provided.

2. Privacy and Individual As a Social Being

One of the most fundamental problems of the cyber/digital age that we live in is privacy. In the digital age, people have a digital presence beyond their physical existence, and the area of this digital asset tends to grow constantly. It has become almost impossible to carry out daily work without sharing personal data to use digital services (Zhu, 2011). However, people's use of e-services is not always a necessity. People also use such services when they think it offers them convenience, cheapness, speed, or some other benefit (Bennett, 2009). In this section, the preferences and obligations between individuals' commitment to digital environments and their privacy are discussed, and suggestions from some studies are presented.

2.1. Socialization As a Human Need in the Digital World

In the age of modern technology, the individual cannot live without leaving a data trail behind. Not being in a social network in the digital age can lead to consequences such as being unable to maintain communication in social and business life, not being able to develop new relationships, missing opportunities, and sometimes even being excluded (Fuchs, 2014). However, humans are social creatures, and information sharing is a central feature of human connection. Making oneself public, sharing, and disclosing provide numerous benefits, including psychological and physical health. The desire for interaction, socialization, disclosure, recognition or fame, and fear of insignificance are basic human motives like privacy. Social media has provided a unique opportunity to meet these basic human motives (Acquisti et al., 2015).

Both the internet and social media deliver messages to large audiences and make messages much more visible. The thought of "I am seen (watched, noted, recorded) therefore I am" (Bauman & Lyon, 2013) is becoming widespread as a view that dominates social media. According to Niedzviecki, who deals with the concept of "peeping culture," an individual becomes aware of being an individual when she/he makes herself/himself watched and when others comment on himself/herself (Niedzviecki, 2009).

2.2. Personal Data as a Valuable Asset

Personal data is any data that can be used to distinguish one person from another. These personal data, some of which are presented in Figure 1, are used to group people in digital environments as target audiences, especially for commercial and political purposes (see the Cambridge Analytica scandal (Confessore, 2018)).

Figure 1. Some personal information given to or obtained from digital media

Personal data can be easily extracted through continuous individual traces left behind when browsing social networks and websites, or using devices such as smartphones. These data may be collected and used for marketing, provocation, or experimental research. With the commercial value of data, every user has become critical; applications and platforms aim to obtain more data from more users to gain maximum profit. Precisely for this reason, these commercial enterprises are designed to meet the basic needs of users, such as communicating, socializing, and having fun, in order to encourage data disclosure (Acquisti et al., 2015).

The use of personal data represents a power that can affect society and individuals in an unprecedented way, not only in economic terms. Thus, most experts agree that the complex balance between individual rights and collective knowledge should not be entrusted solely to market dynamics (Politou et al., 2021).

2.3. Current Social Measures

Users have hardly any control over their personal information stored and analyzed by the relevant data controllers (Mittelstadt & Floridi, 2016; S. Yu, 2016). There are two contexts of digital data that pose a problem in terms of privacy. In some cases, even if the individual has chosen to share the data by her/himself, she/he may need to gain the knowledge and control that this data can be used and shared by third parties. The concepts of data shadow and data traces are used to explain this distinction. While the data shadow consists of the data collected without the individual's knowledge, consent, and awareness, the data traces comprise data that a person knowingly makes public (Koops, 2013). With up-to-date machine learning and big data methods and tools, sensitive information can be disclosed from seemingly unrelated data, and many parties can be involved in the distribution of this data. Therefore, the consent and knowledge of the individual are insufficient to solve privacy problems. Based on the seemingly unconnected data that the individual voluntarily shares on social media, extremely personal information about the individual that is undesirable to be disclosed can be accessed (Kelleher & Tierney, 2018).

As expressed in the Science and Technology Board (PCAST) Report 2014[1], notification and consent is the most widely used method to protect privacy. However, this means putting the entire responsibility of protecting privacy on the individual. It is optimistic to think that the user reads all the notifications and understands the legal implications, but the situation is different in reality. For this reason, there is a need for administrative rules to be in force in data collection and processing activities.

The vast amount of people's information on digital media offers severe vulnerabilities to the right to privacy. Governments have had to make more specific regulations to ensure privacy is protected in any transaction involving sensitive information. To mitigate the privacy risk, the European Parliament and Council enforced the General Data Protection Regulations (GDPR) (Voigt & von dem Bussche, 2017). Although privacy has to be protected by such kinds of regulations, there is more need for frameworks or practices to guide the implementation of these regulations (Jones & Kaminski, 2021). Many countries still do not have such regulations; so there is a greater need to establish country-specific and legal regulations with their framework and practices (Carey & Acquisti, 2018).

3. Technological Privacy Protections and Disclosures

3.1. Machine Learning and Privacy

Machine Learning is the most significant part of Artificial Intelligence that has various algorithms for making the system learn itself. In artificial intelligence, learning means establishing a relation between the system's inputs and output(s). The system refers to the computational and algorithmic-based software. Researchers can use various software techniques to extract, identify, determine, predict or explore valuable information in the data of any domain. Artificial intelligence means to create a machine that has the ability to make one or more human-like behavior; machine learning means a computer program that can learn to produce a human-like behavior. This behavior is learned based on data, metrics, and feedback mechanisms, but the most crucial part is data (Joshi, 2020).

1 https://obamawhitehouse.archives.gov/blog/2014/05/01/pcast-releases-report-big-data-and-privacy

Machine learning techniques construct a mathematical model of data samples to make decisions, predictions, or information extraction, called a "train set." It is one of the essential parts of a machine learning system. In social media, there are comprehensive and streaming data that have to be stored and processed with high-level devices and mechanisms. Machine learning techniques are the possible ways to gain knowledge from these data by processing them. Machine learning has various interdisciplinary and intra-disciplinary domain applications such as opinion mining, medical sciences, textual forensics, etc. to obtain meaningful information from the data (Joshi, 2020; T.k. et al., 2021).

Machine learning and privacy are intertwined concepts. In the analysis or publication of personal data, machine learning techniques are primarily used to protect privacy, increase personal data privacy, and detect disclosure of personal data privacy. On the other hand, machine learning models, such as prediction models for insurance rates or stock prices, are also cases where the privacy of both models and their parameters should be protected (Liu et al., 2022).

A large part of communication in digital environments is done through digital texts. According to a survey published in 2021 (T.k. et al., 2021), more than 18.2 million text messages are transmitted in a minute. Natural Language Processing (NLP) is one of the fundamentals of Artificial Intelligence applications that use computational techniques in order to understand, learn and produce human language content (Hirschberg & Manning, 2015). NLP, also known as Computational Linguistics, focuses on generating technologies to discover the knowledge/wisdom of digital texts like a human does (Chowdhary, 2020). There have been significant improvements in NLP studies over the past 20 years. In the early times of NLP studies, scientists were interested in the automated analysis of linguistic structures like language translation. Today, NLP studies are in a position to make social media analyses that can reveal the user's depression level (Patidar & Umre, 2021), mental health status (Hao et al., 2013), demographic characteristics (Wang et al., 2017), personality (Tay et al., 2020), etc. With the increase in the computing power of computers, the amount of data in digital environments, and advanced models in artificial intelligence techniques, the diversity and success of NLP studies are also increasing. Figure 2 shows the process flow that some of the current text analysis processes. On the one hand, high-performance tools such as Stanford Core NLP (Manning et al., 2015) that can extract syntactic and semantic information in a text are being developed; on the other hand, some resources are still insufficient, especially for low-level languages.

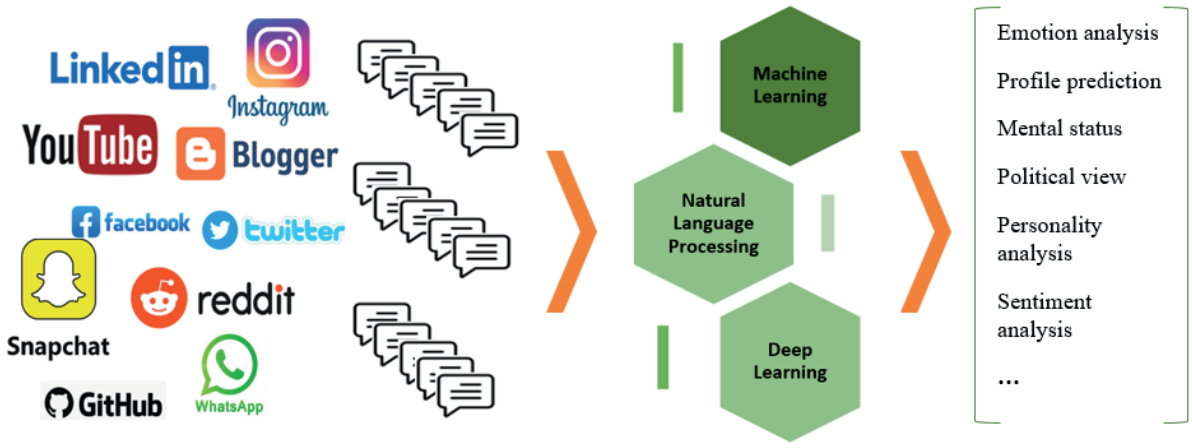

Figure 2. Process flow of some of the current applications through social media and text

NLP research has evolved with the increasing popularity of social media. Social media provide an incredible amount of information about users. Sources from these media, such as Facebook, Twitter, Instagram, LinkedIn, GitHub, Blogs, YouTube, forum sites, and more, are precious inputs for NLP studies. With these inputs, researchers are able to extract relations between social interactions, personal information, and language usage (Ali, 2015). NLP studies are frequently used in studies aimed at protecting or improving the privacy of the owner of the data used, as well as making valuable inferences from the data. Using NLP tools and methods, for example, domain-specific sensitive information can be identified from medical and legal documents and excluded from analyzes to enhance data privacy (Martinelli et al., 2020).

In addition to NLP studies, image processing and voice processing techniques in privacy protection are also among the current study topics. For privacy protection, the detection of sensitive objects in images with deep learning techniques (J. Yu et al., 2017), and the evaluation of voice conversion-based privacy protection against attacks (Lal Srivastava et al., 2020) are some of the studies in this area.

3.2. Big Data and Privacy

Machine learning systems work better as data sets are more extensive. The emergence of big data has contributed significantly to machine learning and data mining techniques, especially those for profiling. According to the UK ICO[2] (Information Commissioner's Office) report, big data analytics, mainly characterized by using machine learning algorithms with new types of data, are frequently reused and can benefit businesses, society, and consumers as citizens. Big data analytics can also help to deliver the public more effective and efficient services and produce positive outcomes that improve people's quality of life.

It is not possible to talk about social analysis without talking about big data. Today, data comes from everywhere: social media sites, sensors, cell phones, sharing sites, e-commerce sites, etc. A daily increase in the vast amount of data is the first indicator of the requirement for big data. Big data is not just a means to evaluate the vast amount of data; it also means to evaluate different properties of data that increase as data increases. In 2001, Laney defined the concept of big data with 3V: high Volume, high Velocity, and high Variety (Laney, 2001). This concept has then widened to 4V (adding high Value) and 5V (adding high Veracity) (Hashem et al., 2015). Big data is an important affair that aims to generate an effective alternative to traditional solutions regarding databases and data analysis (Bello-Orgaz et al., 2016).

Big data has become a significant issue in the field of privacy and machine learning with the emergence of cloud computing. Advances in various big data framework such as Spark (Zaharia et al., 2010) and Hadoop (White, 2012) have increased the use of machine learning application in different fields. Also, the increasing of machine learning libraries of big data such as Mahout (Owen et al., 2011) and SparkMLib (Deshai et al., 2019) has taken machine learning studies one step further. In order to gain more benefits from big data, there are also developed privacy-preserving big data publishing techniques by researchers (Canbay et al., 2019; Zakerzadeh et al., 2015). The vast amount of data with the big data facilities has been used to discover significant knowledge to improve decision-making processes. However, there are still some open problems and challenges, such as the determination of how much data is enough for high-quality data (quantity versus quality) or ensuring enough privacy (security and ownership) (Bello-Orgaz et al., 2016).

Big data analytics generate new knowledge by locating unexpected and previously unknown structures, correlations, and patterns by combining algorithms and information from large and various datasets (Hildebrandt, 2009). Thus, a person's online and offline activities are converted into profiling scores, while predictive algorithms extract personal information to make predictions about individuals' likely actions and behaviors. In summary, activities such as extensive profiling and scoring people based on their profiles with big data analytics and machine learning algorithms are now more suitable than ever to be used by private companies or public authorities (Politou et al., 2021).

4. Technology with Privacy

We live in a digital world, and many applications on the Internet, especially social media applications, are waiting to receive our personal information. These applications, which offer us small rewards or conveniences in return for our personal data, can make millions by processing, selling, or using this data. The general opinion is that people should protect their privacy, but when the practices and observations are examined , it has been shown that users can ignore permanent privacy concerns over their personal information against short-term benefits or some small rewards (Acquisti & Grossklags, 2005). By human nature, our abilities are limited by our bounded rationality (Simon, 2019). We have insufficient memory and processing power to calculate possible implications for the protection and publication of complex, branched data such as our personal information (Acquisti & Grossklags, 2005).

2 https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

Technically, it is not possible to walk around in digital environments without leaving a trace. Even with the most powerful defensive techniques used by the most privacy-conscious users, individuals seem to face great difficulties in avoiding tracking techniques (Acar et al., 2014). Therefore, at this point, there is a need for intelligence applications, responsible technology developers, and governmental regulations that will protect personal privacy. Providing a comprehensive data protection schema considering an application, organization, or government alone is challenging. There needs to be combined countermeasures to enhance privacy protection. While governments make regulations, organizations should employ experts to obey the regulations on data processing or transferring/transactions. Besides, there is a need for applications to help experts to set up and enhance privacy protection.

Many statistical, theoretical, and cryptographic methods have been developed so that personal data can be shared and processed in a way that does not disclose the privacy of the data owners (Churi & Pawar, 2019; Kreso et al., 2021; Majeed & Lee, 2021). The Differential Privacy method is used as one of the most up-to-date and reliable privacy protection methods today. Differential privacy (Dwork, 2008) provides a method that tries to keep the accuracy of the query requested from a database high while minimizing the chance of identifying a query's records. Differential privacy offers solutions to protect the privacy of unstructured data content and to share it with untrusted parties (Zhao & Chen, 2022).

With the proliferation of distributed systems, which is a requirement today because of the amounts of stored and processed data, collaborative learning methods have been generated with respect to privacy protection. One of the popular collaborative learning frameworks is Federated learning (Zhang et al., 2021). Federated learning is a distributed machine learning approach that gives efficiency while providing parallel machine learning model training across many clients. Another popular framework for collaborative learning is Split Learning (Vepakomma et al., 2018). Split learning, another popular distributed machine learning approach, ensures a more private model than Federated learning because the machine learning model architecture is split between server and clients (Thapa et al., 2022).

As mentioned in the previous sections, the privacy risks of the data increase as the data grows. Thus, especially in big data analysis, there is a need for more applications to enhance privacy. In healthcare, the collected data from cyber-physical systems play an essential role in making decisions about the health of humans. The data of these heterogeneous systems are stored in private or public clouds and used for analysis. Many of the existing privacy-preserving data mining and privacy-preserving data publishing techniques do not seem proper to evaluate the unstructured, huge stream of data alone. Using differential privacy, homomorphic encryption, and key-based and clustering-based anonymization techniques alone is not found sufficient to ensure privacy on big data streams. Instead, a hybrid approach with both anonymization and encryption is a good choice (Maleh et al., 2019). Although studies in this field are increasing, data continues to grow, vary, and increase the privacy risk they carry.

5. Changed Privacy Perception (Post-Privacy) and Privacy Paradox

Discussions about the death of privacy have been going on since the mid-1960s. The widespread use of central hosts and monitoring technologies such as cameras and listening devices, which were gradually developing at that time, was the first source of discussion (Schulte, 2018). In the first book declaring the end of privacy, Rosenberg talked about a national computer system in which all citizens' information would be stored and argued that urgent and radical measures should be taken (Rosenberg, 1969). The second wave of privacy concerns escalated with the emergence of the Internet in 1983, the World Wide Web in 1993, and the simultaneous spread of personal computers. With the rapid rise of social media use at the beginning of the 21st century, concerns about "exhibitionist" privacy and sharing personal data with companies have been added to the privacy issue (Schulte, 2018).

Niedzviecki, on the other hand, accepts 2008 as the turning point of a fundamental cultural change and declares that the "age of peeping culture" has begun (Niedzviecki, 2009). As Mark Zuckerberg, the founder of Facebook which is one of the social media platforms, stated in his speech[3] in 2010, privacy is no longer a social norm, and people tend to share more and various information with more people openly. Thus, social media emerged and developed as an area of disclosure and peeping. Figure 3 gives the chronological order of socially melting privacy by the effect of technological and perceptual changes.

3 https://www.theguardian.com/technology/2010/jan/11/facebook-privacy

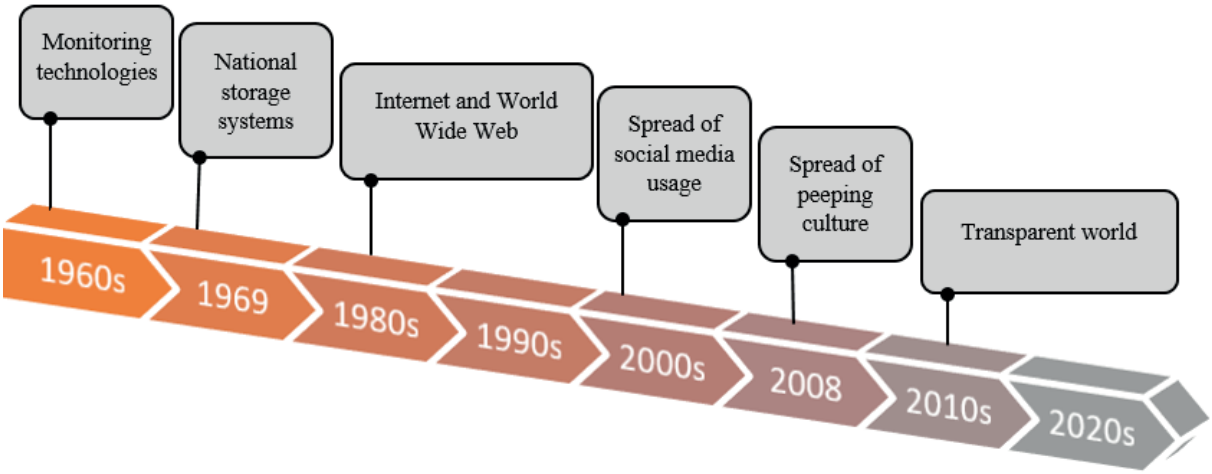

Figure 3. Chronological order of melting privacy by the effect of technological and perceptual changes

Han adapted the definition of Miller's "transparent world" (Miller, 1972) to the 21st century (Han, 2020). He stated that the transparency created by social networks penetrates society irreversibly, and people have to adapt to this new situation. Claiming that we now live in a post-privacy society, Han argues that paying attention to concerns about privacy does not match the realities of the age (Han, 2020). Acquisti et al., on the other hand, argue that sharing more personal data does not always mean development, productivity, and equality. They state that the erosion of privacy threatens the autonomy of individuals (Acquisti et al., 2015).

The complete absence of data sources is neither practical nor desirable. For example, having access to the health records of an unconscious patient brought to the emergency room can save the patient's life. Therefore, keeping data recorded is necessary and essential in some cases. However, it is not cognitively possible for the individual to actively decide on his privacy in his daily life. This cognitive difficulty creates the privacy paradox: people worry about privacy, but in practice, they often do little to protect it (Stalder, 2002). This inconsistency between privacy attitudes and privacy behavior is called the privacy paradox (Norberg et al., 2007).

Privacy calculation theory assumes that individuals make decisions by calculating the potential gain of disclosure and the loss of privacy. For example, Debatin et al. argue that participation in social networks is associated with three needs: the need for entertainment, social connection, and identity construction (Debatin et al., 2009). Therefore, the individual gives up his privacy at the expense of meeting these needs. Also, in many cases, people may lack the cognitive ability or knowledge necessary to make an informed privacy decision. Limited rationality and incomplete information are the main factors determining the privacy decision (Acquisti & Grossklags, 2005). Incomplete and asymmetric information reveals the uncertainty of privacy. It is usual for individuals to be hesitant about sharing information, as they often do not clearly understand what information other people, firms, and governments have about them and how they use that information.

6. Conclusions

Via online activities, people leave easy-to-follow digital trails that reveal who we are, what we eat, where we go, whom we talk to, why we are happy, what we buy, and much more. Those trails are collected and stored somewhere. Once data is collected and stored, we have almost no control over who uses it or how it is used.

While technological developments that benefit humanity are realized in many areas with personal data, private information that even data owners cannot predict can be obtained through analyzing these data. The anonymity of digital environments, the vulnerability of users to attacks, the fact that users do not have enough information about the dangers of these environments, and perhaps most importantly, the inevitable sharing of information in these environments have increased privacy concerns.

Although informing the data owner and obtaining consent is presented as a solution in many studies, it is clear that this cannot prevent the risks of privacy violations in reality. Almost every step in the digital world is followed by applications

such as cookies. Even with the strongest protective measures taken by the most privacy-conscious people, it is almost impossible to stay away from these tracings. Especially without loss of content or functionality in digital media, it is challenging to prevent monitoring. Once tracing has occurred, it is almost impossible to start from a truly clean profile.

In the analysis or publication of personal data, big data, and machine learning techniques are mostly used to protect or enhance personal data privacy. However, the current technological countermeasures are not complete enough for unstructured data such as texts, images, videos, or voices. On the other hand, the privacy risk of data increases as the volume and variety of data continue to grow. Measures for privacy protections are not evolving as quickly as threats to violate privacy.

In the cyber/digital world, countries and institutions have to take measures to protect data privacy. While there is a need for additional resources to explain the necessary practices and frameworks in countries that have enacted measures such as GDPR, many countries still have not announced any legal regulations.

When an evaluation is made together with (1) today's technological developments, (2) individuals' enthusiasm and obligation to share information, (3) the economic and social value of data, and (4) the measures taken by governments against privacy violations, the theory of the death of privacy is increasingly valid from the 2000s. This theory, which was first discussed in the mid-1960s, is in a position that cannot be ignored today. In this digital age, transparency is at the forefront, and individuals are willing to accept temporary benefits in return for their personal data. For these reasons, there is an urgent need for lawmakers and technology developers to resort to advanced measures before privacy violations lead to more significant problems.

Peer-review: Externally peer-reviewed.
Author Contributions: Conception/Design of Study- P.C., Z.D.; Data Acquisition- P.C.; Data Analysis/Interpretation- P.C., Z.D.; Drafting Manuscript- P.C.; Critical Revision of Manuscript- Z.D.; Final Approval and Accountability- P.C., Z.D.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: This work was supported by a grant from the Kahramanmaras Sutcu Imam University Scientific Research Projects Unit, Project Number: 2021/2-38 M.

REFERENCES

Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., & Diaz, C. (2014). The web never forgets: Persistent tracking mechanisms in the wild. *Proceedings of the ACM Conference on Computer and Communications Security*. https://doi.org/10.1145/2660267.2660347

Acquisti, A., & Grossklags, J. (2005). Privacy and rationality in individual decision making. In *IEEE Security and Privacy* (Vol. 3, Issue 1). https://doi.org/10.1109/MSP.2005.22

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. In *Science* (Vol. 347, Issue 6221). https://doi.org/10.1126/science.aaa1465

Ali, D. (2015). Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, Github, and More, by Matthew A. Russell. *Journal of Information Privacy and Security*, *11*(2). https://doi.org/10.1080/15536548.2015.1046287

Balcı, E. (2019). Overview of Intelligent Personal Assistants. *Acta INFOLOGICA*, *3*(1). https://doi.org/10.26650/acin.571303

Bauman, Z., & Lyon, D. (2013). Liquid Surveillance. In *Statewide Agricultural Land Use Baseline 2015* (Vol. 1).

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, *28*. https://doi.org/10.1016/j.inffus.2015.08.005

Bennett, L. (2009). Reflections on privacy, identity and consent in on-line services. *Information Security Technical Report*, *14*(3). https://doi.org/10.1016/j.istr.2009.10.003

Canbay, Y., Vural, Y., & Sagiroglu, S. (2019). Privacy Preserving Big Data Publishing. *International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism, IBIGDELFT 2018 - Proceedings*. https://doi.org/10.1109/IBIGDELFT.2018.8625358

Carey, P., & Acquisti, A. (2018). Data protection: a practical guide to UK and EU law. In *Economics of Information Security*.

Chowdhary, K. R. (2020). Fundamentals of artificial intelligence. In *Fundamentals of Artificial Intelligence*. https://doi.org/10.1007/978-81-322-3972-7

Churi, P. P., & Pawar, A. v. (2019). A systematic review on privacy preserving data publishing techniques. In *Journal of Engineering Science and Technology Review* (Vol. 12, Issue 6). https://doi.org/10.25103/jestr.126.03

Confessore, N. (2018). Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. *The New York Times*.

Correa, T., Hinsley, A. W., & de Zúñiga, H. G. (2010). Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior*, *26*(2). https://doi.org/10.1016/j.chb.2009.09.003

Debatin, B., Lovejoy, J. P., Horn, A. K., & Hughes, B. N. (2009). Facebook and online privacy: Attitudes, behaviors, and unintended consequences.

Journal of Computer-Mediated Communication, *15*(1). https://doi.org/10.1111/j.1083-6101.2009.01494.x

Deshai, N., Sekhar, B. V. D. S., & Venkataramana, S. (2019). Mllib: machine learning in apache spark. *International Journal of Recent Technology and Engineering*, *8*(1).

Dwork, C. (2008). Differential privacy: A survey of results. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4978 LNCS*. https://doi.org/10.1007/978-3-540-79228-4_1

Fuchs, C. (2014). Social Media: A Critical Introduction. In *Social Media: A Critical Introduction*. https://doi.org/10.4135/9781446270066

Han, B.-C. (2020). The Transparency Society. In *The Transparency Society*. https://doi.org/10.1515/9780804797511

Hao, B., Li, L., Li, A., & Zhu, T. (2013). Predicting mental health status on social media a preliminary study on microblog. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8024 LNCS*(PART 2). https://doi.org/10.1007/978-3-642-39137-8_12

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98–115. https://doi.org/10.1016/j.is.2014.07.006

Hildebrandt, M. (2009). Who is Profiling Who? Invisible Visibility. In *Reinventing Data Protection?* https://doi.org/10.1007/978-1-4020-9498-9_14

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. In *Science* (Vol. 349, Issue 6245). https://doi.org/10.1126/science.aaa8685

Jones, M. L., & Kaminski, M. E. (2021). AN AMERICAN'S GUIDE TO THE GDPR. *Denver Law Review*, *98*(1).

Joshi, A. v. (2020). *Machine Learning and Artificial Intelligence*. Springer International Publishing. https://doi.org/10.1007/978-3-030-26622-6

Kelleher, J. D., & Tierney, B. (2018). Data Science. In *Data Science*. https://doi.org/10.7551/mitpress/11140.001.0001

Koops, B.-J. (2013). Forgetting Footprints, Shunning Shadows: A Critical Analysis of the "Right to Be Forgotten" in Big Data Practice. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1986719

Kreso, I., Kapo, A., & Turulja, L. (2021). Data mining privacy preserving: Research agenda. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 11, Issue 1). https://doi.org/10.1002/widm.1392

Lal Srivastava, B. M., Vauquier, N., Sahidullah, M., Bellet, A., Tommasi, M., & Vincent, E. (2020). Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2020-May*. https://doi.org/10.1109/ICASSP40776.2020.9053868

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, *6*(70).

Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2022). When Machine Learning Meets Privacy. *ACM Computing Surveys*, *54*(2), 1–36. https://doi.org/10.1145/3436755

Majeed, A., & Lee, S. (2021). Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access*, *9*. https://doi.org/10.1109/ACCESS.2020.3045700

Maleh, Y., Shojafar, M., Darwish, A., & Haqiq, A. (2019). Cybersecurity and Privacy in Cyber-Physical Systems. In *Cybersecurity and Privacy in Cyber-Physical Systems*. https://doi.org/10.1201/9780429263897

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2015). *The Stanford CoreNLP Natural Language Processing Toolkit*. https://doi.org/10.3115/v1/p14-5010

Martinelli, F., Marulli, F., Mercaldo, F., Marrone, S., & Santone, A. (2020). Enhanced Privacy and Data Protection using Natural Language Processing and Artificial Intelligence. *Proceedings of the International Joint Conference on Neural Networks*. https://doi.org/10.1109/IJCNN48605.2020.9206801

Marwick, A. E. (2012). The public domain: Social surveillance in everyday life. *Surveillance and Society*, *9*(4). https://doi.org/10.24908/ss.v9i4.4342

Matarić, M. J., Eriksson, J., Feil-Seifer, D. J., & Winstein, C. J. (2007). Socially assistive robotics for post-stroke rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, *4*. https://doi.org/10.1186/1743-0003-4-5

Miller, A. R. (1972). *Computers, Data Banks and Individual Privacy: An Overview*. Colum. Hum. Rts. L. Rev. 4.

Mittelstadt, B. D., & Floridi, L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. In *Science and Engineering Ethics* (Vol. 22, Issue 2). https://doi.org/10.1007/s11948-015-9652-2

Niedzviecki, H. (2009). *The peep diaries: How we're learning to love watching ourselves and our neighbors*. City Lights Publishers.

Norberg, P. A., Horne, D. R., & Horne, D. A. (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, *41*(1). https://doi.org/10.1111/j.1745-6606.2006.00070.x

Owen, S., Anil, R., Dunning, T., & Friedman, E. (2011). Mahout in Action. In *Online*.

Patidar, H., & Umre, J. (2021). Predicting depression level using social media posts. *International Journal of Research -GRANTHAALAYAH*, *8*(12). https://doi.org/10.29121/granthaalayah.v8.i12.2020.1972

Politou, E., Alepis, E., Virvou, M., & Patsakis, C. (2021). Privacy and Data Protection Challenges in the Distributed Era. In *Learning and Analytics in Intelligent Systems*.

Rosenberg, J. M. (1969). *The Death of Privacy*. Random House (NY).

Rosenthal, S., Biswas, J., & Veloso, M. (2010). An effective personal mobile robot agent through symbiotic human-robot interaction. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *2*.

Schulte, P. (2018). David Vincent, Privacy. A Short History. Cambridge, Polity Press 2016. *Historische Zeitschrift*, *307*(2). https://doi.org/10.1515/hzhz-2018-1399

Simon, H. A. (2019). Models of Bounded Rationality. In *Models of Bounded Rationality*. https://doi.org/10.7551/mitpress/4711.001.0001

Stalder, F. (2002). Opinion. Privacy is not the antidote to surveillance. In *Surveillance and Society* (Vol. 1, Issue 1). https://doi.org/10.24908/ss.v1i1.3397

Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and Validity Issues in Machine Learning Approaches to Personality Assessment: A Focus on Social Media Text Mining. *European Journal of Personality*, *34*(5). https://doi.org/10.1002/per.2290

Thapa, C., Mahawaga Arachchige, P. C., Camtepe, S., & Sun, L. (2022). SplitFed: When Federated Learning Meets Split Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(8), 8485–8493. https://doi.org/10.1609/aaai.v36i8.20825

T.k., B., Annavarapu, C. S. R., & Bablani, A. (2021). Machine learning algorithms for social media analysis: A survey. In *Computer Science Review* (Vol. 40). https://doi.org/10.1016/j.cosrev.2021.100395

Vepakomma, P., Gupta, O., Swedish, T., & Raskar, R. (2018). Split learning for health: Distributed deep learning without sharing raw patient data. *ArXiv Preprint ArXiv:1812.00564*.

Voigt, P., & von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR) A Practical Guide. In *The EU General Data Protection Regulation (GDPR)*.

Wang, Q., Ma, S., & Zhang, C. (2017). Predicting users' demographic characteristics in a Chinese social media network. *Electronic Library*, *35*(4). https://doi.org/10.1108/EL-09-2016-0203

White, T. (2012). Hadoop: The definitive guide 4th Edition. *Online*, *54*. https://doi.org/citeulike-article-id:4882841

Yu, J., Zhang, B., Kuang, Z., Lin, D., & Fan, J. (2017). IPrivacy: Image Privacy Protection by Identifying Sensitive Objects via Deep Multi-Task Learning. *IEEE Transactions on Information Forensics and Security*, *12*(5). https://doi.org/10.1109/TIFS.2016.2636090

Yu, S. (2016). Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data. *IEEE Access*, *4*. https://doi.org/10.1109/ACCESS.2016.2577036

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. *2nd USENIX Workshop on Hot Topics in Cloud Computing, HotCloud 2010*.

Zakerzadeh, H., Aggarwal, C. C., & Barker, K. (2015). Privacy-preserving big data publishing. *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*, 1–11. https://doi.org/10.1145/2791347.2791380

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. *Knowledge-Based Systems*, *216*. https://doi.org/10.1016/j.knosys.2021.106775

Zhao, Y., & Chen, J. (2022). A Survey on Differential Privacy for Unstructured Data Content. *ACM Computing Surveys*. https://doi.org/10.1145/3490237

Zhu, L. (2011). Privacy in Context: Technology, Policy, and the Integrity of Social Life. *Journal of Information Privacy and Security*, *7*(3). https://doi.org/10.1080/15536548.2011.10855919