ARTICLE TEMPLATE

Privacy and Ethics in GeoAI

Grant McKenzie[1], Hongyu Zhang[1], and Sébastien Gambs[2]

[1]Platial Analysis Lab, Department of Geography, McGill University, Montréal, Canada
[2]Département d'informatique, Université du Québec à Montréal, Montréal, Canada

ARTICLE HISTORY
Compiled April 24, 2023

ABSTRACT
Any advancement in technology is accompanied by new concerns over its ethical use and impacts on privacy. While a notoriously difficult term to define, privacy as it relates to technology usage, can be described as the ability of an individual or group to control their personal information. Like many ethical concepts, this definition evolves with changes in societal and technical norms. The emergence of machine learning and related artificial intelligence techniques has again shifted societal concerns about the privacy of our persons, socio-demographic group membership, and personal data. Location data are particularly sensitive as they link information across sources and can be used to infer a wide variety of personal information. This makes data privacy one of the most important ethical discussions within the field of geographic artificial intelligence (GeoAI). The main objective of this chapter is to explore the unique privacy concerns associated with AI techniques used for analyzing geospatial information. After providing an overview of the topic, we describe some of the most common techniques and leading application areas through which data privacy and GeoAI are converging. Finally, we suggest a number of ways that privacy within GeoAI can improve and highlight emerging topics within the field.

KEYWORDS
Privacy; Ethics; Machine Learning; Anonymity; Location-aware Technology

1. Introduction

The number of companies, agencies, and institutions using artificial intelligence (AI) techniques has grown substantially over the past few years. Their goals are diverse and span application areas ranging from cashier-less grocery stores to breast cancer screening. As with any technology, these advancements have lead to important discussions related to ethics. In particular, ethical concerns associated with such technologies range from the collection and storage of personal data to biases in model development and implementation. These concerns also encompass questions on how best to explain their predictions. While ethics is its own domain of research, the rapid development and adoption of AI techniques in many sectors of society has given rise to the field of ethical artificial intelligence (Mittelstadt 2019). Researchers of ethics in AI aim to identify and investigate issues facing society that can specifically be attributed to the introduction and application of AI and related methods. Approaches to the topic most often include exploration and analysis of one or more themes such as privacy,

Corresponding Author: grant.mckenzie@mcgill.ca

surveillance, bias and/or discrimination (Stahl and Wright 2018, Naik et al. 2022).

Like many other aspects of AI, ethical concerns are also shifting. The field is changing so rapidly that legal experts, policy makers, and researchers are forced to continually revise their assessments of bias, transparency, social manipulation, and privacy in AI. Through increased public pressure, many leading technology companies have hired experts to help them navigate these waters and develop policies related to the ethical use of AI. Many private companies and government agencies regularly publish technical reports outlining AI guidelines and principles. A recent scoping review of 84 existing guidelines on ethical AI by Jobin et al. (2019) identified a set of ethical principles commonly included in these reports. The top five include *transparency*, *justice & fairness*, *non-maleficence*, *responsibility* and *privacy*. Each of these principles is worthy of its own book chapter, with numerous books having already been published on these topics (see Dubber et al. (2020), for instance).

In this chapter, we choose to focus our discussion on the ethical principle of privacy. To understand why, we must examine ethics as it relates to the topic of this book, namely *geographic* artificial intelligence (GeoAI). We argue that the same common set of AI ethical principles identified by Jobin et al. also apply to GeoAI, but that the relative importance, or ranking, of these principles has been modified. AI techniques that leverage the relationships of objects, events, and people in geographic space make GeoAI a unique subset of artificial intelligence. We argue here that ethical issues related to privacy are fundamentally different when viewed through a geographic lens. Thus, while a discussion on ethics and all of its themes are essential to the future of GeoAI research, the unique aspects of location privacy will be the focus of this chapter.

1.1. *Data privacy & AI*

In today's technocratic society, the privacy of one's personal information is of the utmost importance. Given "big tech's" penchant for collecting data for AI training purposes, people have become increasingly concerned about how their data are being used and how much control they retain over their data. Historically, the broader concept of privacy has been difficult to grasp, with definitions differing substantially depending on the domain considered. The word *private* is derived from the Latin *privatus*, which means to set apart from what is public, personal and belonging to oneself, and not to the state. Various efforts have been made to categorize privacy into different dimensions (Pedersen 1979, Finn et al. 2013) but many of them come to the conclusion that privacy is the right of an individual or group to control how information about them is accessed, shared, and used, thus being related to the concept of self-information determination. This is a data-centric definition of privacy, which is arguably the most applicable to the GeoAI context.

When the terms privacy, data, and AI are combined, most readers' minds go to a futuristic surveillance state reminiscent of George Orwell's Big Brother. While such a scenario is worthy of further discussion, there are a number of less Orwellian representations of privacy, or privacy violations, that should also be acknowledged. Many of these are less dramatic, but should be no less concerning to those that use AI technologies. As many have noted, the heart of most AI techniques is the data on which the models are trained – sometimes referred to the *petrol* of AI. The provenance of these data, and details on the individuals from which these data are collected, continue to be a topic of much discussion among privacy researchers. In this era of Big Data we have also seen the emergence of data brokers purchasing and selling data for a variety of uses. Ethics related to data handling, and the confidentiality, anonymity, and privacy of the data all then become topics for further investigation. As the commercial appetite for data grows, we have seen a societal shift from people trading commodities to the information of those people now *being* the commodities. This has led to a significant change in our perception of privacy and the steps we take to ensure it (Zhang and McKenzie 2022).

With respect to AI, a lot of what is being discussed is not about individual privacy from a philosophical position, but rather *data privacy*, or the rights of the individual to control what information is being collected, accessed, shared and analyzed. More precisely, privacy has the potential to be viewed as a value to uphold or a right to be protected. This latter definition is less about the "right to be left alone" and more about the right to control one's own information. There is a separate philosophical discussion to be had about privacy and AI but in this work we focus on the ethical concerns over data privacy in AI, and specifically GeoAI.

1.2. *Geoprivacy & GeoAI*

It has been two decades since Dobson and Fisher (2003) published their paper *Geoslavery*, an evocative call to action showcasing how geographic information systems, global navigation satellite systems, and location-based services can be used to control individuals. While technology trends have deviated from those mentioned in the paper, the idea that location is a unique attribute capable of exposing highly sensitive information remains. Location is inherently tied to identity. Indeed, a plethora of research has demonstrated that socio-economic and demographic characteristics such as race, income, education, and many others correlate with location (Zhong et al. 2015, Riederer et al. 2016). The places that we visit (*e.g.*, restaurants, bars, parks, etc.) and times we visit them are also closely tied to our demographics characteristics (Liu et al. 2017, McKenzie and Mwenda 2021). The mobility behaviour of an individual uniquely characterizes them and can be used for re-identification even from so called "anonymous data" (Gambs et al. 2014a). Thus, publicly sharing the places that one visits, without their knowledge, can be a major violation of their privacy. For instance, exposing the bar one patrons on a Saturday evening may be of little concern for a cisgender male in a North American city, but it may be of appreciable concern to a non-binary gender individual living in a nation in which it is illegal to identify as such. The link between location and identity make such data particularly sensitive – and valuable. For developers of AI methods and tools, these data are an extraordinary resource on which to train models for applications areas such as human behavior and crime prediction, local business recommendations, or determining health insurance rates.

Geographers and demographers understand that access to an individual's location data is only the tip of the proverbial "privacy exposure iceberg." Paraphrasing the first law of geography, we know that things that are closer together in geographic space tend to me more similar (Tobler 1970). From a data privacy ethics perspective, this means that gaining access to socio-demographic information about my neighbor (*e.g.*, income, race and age) means that one can infer my socio-demographic characteristics with a high degree of accuracy. This presents the uncomfortable reality that the privacy of an individual's personal information depends on the privacy of information of those in close proximity. The dilemma here is that, while I do not have control over the personal information that my neighbor chooses to share, I am impacted by the disclosure of such content. In the era of social media, user-generated content, and other sources of geo-tagged data, this means that it is possible to infer information about me purely based on my location and the contributions of people around me (Pensa et al. 2019). This is often referred to as *co-location privacy*. AI technologies have amplified this allowing for data from multiple sources to be combined, multiplying probabilities by probabilities to infer details about people with shocking levels of accuracy. This leads to an entire new set of ethical considerations as we now see that sharing individual location information impacts collective or group privacy.

Despite the fact that location information is so important to our identity, it is surprisingly easy to capture. As outlined by Keßler and McKenzie (2018) in their *Geoprivacy Manifesto*, "ubiquitous positioning devices and easy-to-use APIs make information about an individual's location much easier to capture than other kinds of personally identifiable information." There are so many accessible data out there that the privacy of individual's locations has become a domain of research in and of itself. For instance, research has identified that the location of individuals can be inferred purely based on the text that people share online (Adams and Janowicz 2012), the photos they post (Hasan et al. 2020) or the time of day that they share information (McKenzie et al. 2016). Armstrong et al. (2018) provide an excellent overview of the domain of *geoprivacy* including examples of some of the leading issues in location privacy research. Additional work has specifically reviewed the state of location privacy issues in mobile applications (Liu et al. 2018) and cartographic publications (Kounadi and Leitner 2014). Like many research domains, those working in geographic information science have renewed calls to investigate ethics as it relates to location privacy and many other themes (Nelson et al. 2022). While not always purposeful, we are increasingly seeing GeoAI techniques used to de-anonymize location data, identify individuals, and violate individual privacy (Wang et al. 2016). As we witness the emergence of GeoAI built on massive amounts of personal, location-tagged content and geospatial data, scientists are reminded of Dobson and Fisher's warning from the early 2000s. If GPS and GIS were perceived to be the harbingers of a geotechnology-enabled surveillance state, what then is GeoAI?

It is not all doom and gloom. The emergence of GeoAI has substantially impacted our society in a number of positive ways (many of which are showcased throughout this book). From a data privacy perspective, advances in GeoAI and affiliated machine learning models have made major contributions to privacy *preservation*. Numerous research teams have contributed to the emergence of new methods, techniques, and tools for obfuscating, anonymizing, encrypting, and protecting location information (Jiang et al. 2021). Public-sharing location applications such as *Koi* (Guha et al. 2012) or *PrivyTo* (McKenzie et al. 2022) are being created that use many of these location obfuscation and data encryption techniques to put users back in control of their personal location information.

2. Data privacy methods in GeoAI

A wide range of artificial intelligence and machine learning techniques exist that touch on privacy as it relates to geospatial data. These can be split between one group that focuses on protection mechanisms such as privacy-preservation, anonymization, and obfuscation, and a second group dedicated to privacy attacks such as re-identification, de-anonymization, and privacy exposure.

2.1. *Obfuscation & anonymization*

A standard approach for preserving the privacy of a dataset involves obfuscating the dataset, or its properties, in some way. Typical approaches include adding noise either randomly or following some structured probability distribution. These approaches are not unique for location data, but location-specific noise-based obfuscation techniques have been developed. For instance, geomasking or spatial-temporal cloaking, refer to a broad set of methods used for obfuscating location data (Armstrong et al. 1999). Methods for obfuscating point coordinates include reporting a broader geometric region (*e.g.*, circle or annulus) in which the point exists, displacing the point by some distance and direction or reporting the political or social boundary in which the point is contained (Seidl et al. 2016). A variety of tools, such as *MaskMy.XYZ* (Swanlund et al. 2020) have been developed to help the average privacy-conscious user geomask their location content.

Anonymization is another way of preserving individual privacy, which aims to keep one's identity private but not necessarily one's actions. In contrast to obfuscation techniques, the objective is not necessarily to hide sensitive information through the addition of noise but rather to reduce the accuracy of the information disclosed in order to limit the possibility of re-identifying a particular mobility profile. Various approaches have been developed to guarantee some degree of geospatial data anonymity. For instance in k-anonymity, the objective is to hide the particular mobility behaviour of a user among other users sharing similar patterns. More precisely, a dataset is said to be k-anonymized if a record within the set cannot be differentiated from k-1 other records. While the seminal work on this topic (Sweeney 2002) did not specifically focus on location data, subsequent efforts have highlighted the ways in which one can k-anonymize spatial datasets (Ghinita et al. 2010). Geographic obfuscation methods such as Adaptive Areal Elimination (Kounadi and Leitner 2016, Charleux and Schofield 2020) leverage this k-anonymity property of the data to identify regions that offer a measurable level of privacy.

Differential privacy is often heralded as one of the field's most significant advances, offering strong and formal privacy guarantees (Dwork 2006). The objective of differential privacy is to extract and publish global usable patterns from a set of data while maintaining the privacy of the individual records in the set. This approach involves adding noise to a dataset such that exposure of one, or a set of attributes, will not expose the identity of a record or individual. Since 2015, differential privacy has been used by leading technology companies to monitor how products are used along with purchasing and mobility trends. Within the geographic domain, variations on differential privacy have been introduced, such as geo-indistinguishability (Andrés et al. 2013), that acknowledge the unique properties of geographic data and obfuscate location details through tailored geomasking techniques (Kim et al. 2021).

With the growth in GeoAI, a variety of new obfuscation and anonymity methods have emerged that leverage network graphs (Jiang et al. 2019), discrete global grids (Hojati et al. 2021), and decentralized collaborative machine learning (Rao et al. 2021), to name a few. In addition, the continued growth of contextually-aware devices has led to advances in obfuscation techniques for mobile device users (Jiang et al. 2021).

2.2. *Synthetic data generation*

An alternative to obfuscating or anonymizing real location data is to instead generate *synthetic* data. Sometimes referred to as fake or dummy data, the privacy of a dataset can be maintained by not reporting any piece of the original data at all. Instead, a new set of data are generated that exhibit similar properties of the original dataset. Such an approach can be tailored to specific use cases by only selecting the properties of interest from the original dataset. Methods of synthesizing data are often devised to protect the privacy and confidentiality of particular parts of a dataset, or the data as a whole. The generation of synthetic data through generative models is a hot topic in machine learning and numerous data synthesis methods have been developed and are actively in use in a variety of domains (Nikolenko 2021). With respect to geospatial data, synthetic population data has a long history in demography (Beckman et al. 1996) with governmental programs, such as the census, often generate synthetic data for regions with small or susceptible populations. In such cases, a population may be so small that even reporting aggregate values may expose unique individuals in a region. Synthetic data can be generated based on properties of the original data, but be adjusted such that the privacy of individuals can be maintained. With respect to location privacy, synthetic data have been used to understand crowd dynamics (Wang et al. 2019), analyze mobility trajectories (Rao et al. 2020) and more generally address a wide array of pressing geographic problems (Cunningham et al. 2021).

2.3. *Cryptography*

The previously mentioned techniques aim to preserve privacy either through distortion of the original data or generating dummy data. An alternative to these approaches is to simply hide the data using cryptographic techniques. Encryption is a widely used technique for storing and sharing information when the content needs to remain private. The limitation of such an approach is that once encrypted, the utility of the data is basically non-existent for someone that does not have the associated decryption key. Whereas geographic coordinates obfuscated to a neighborhood may still provide utility for location-based services, encrypted data are useless to anyone but those with the ability to decrypt them.

Researchers working with geographic data have proposed a variety of ways to encrypt geospatial data but still maintain some degree of utility. For instance, some approaches rely on partial encryption of the data meaning that some properties are exposed while others remain hidden (Sun et al. 2019, Jiang et al. 2021). Similar to some of the methods mentioned in the previous section, this means that identifiable and confidential information will be encrypted while spatial properties of a dataset (*e.g.*, degree of clustering), may be published. Geospatial communication platforms such as *Drift* (Brunila et al. 2022), are being developed that encrypt geospatial data but maintain utility.

On the advanced cryptographic primitives side, we have seen the recent adoption of homomorphic encryption in a variety of applications (Acar et al. 2018). Homomorphic encryption is an encryption method that allows one to analyze encrypted data without first decrypting it. Such analysis can result in the extraction of patterns and insight without having access to the original unencrypted private information. This technique is actively being used in health research and demography (Munjal and Bhatia 2022). There are limits to homomorphic encryption, not least of which are the types of analyses that can be performed and the computational costs of such analyses. The unique types of analyses that are conducted on geospatial data offer challenges for homomorphic encryption techniques (Alanwar et al. 2017) but advances in this area are sure to be made in the coming years.

2.4. *Re-identification methods & privacy attacks*

While the methods described in the previous sections aim at preserving privacy and anonymity, another set of methods relevant for privacy researchers are those used for de-anonymizing data and conducting other privacy attacks. While there is not a single leading approach to focus on, we instead highlight a few examples of how this is being done with location data.

De-anonymization approaches often involve the inclusion of an external dataset reflecting the knowledge of a potential adversary during analysis (Harmanci and Gerstein 2016). One possible approach to de-anonymization is through a linkage attack that leverage relationships between the external dataset and the anonymized one, reducing the anonymity of individual records in the process (Narayanan and Shmatikov 2008). Unique properties of location data such as the habitual movement patterns of people can also be leveraged to de-anonymize a dataset. For example, Gambs et al. (2014b) trained a Mobility Markov Chain model on a set of known mobility trajectories and used this model to identify individuals in an anonymized set of trajectories. When the data represents the location of individuals, *co-location analysis* can be used to reduce the privacy of seemingly obfuscated or anonymized data. For instance, geosocial media users frequently report their co-locations with other users through tags or photographs. Internet protocol (IP) addresses are also a means of co-location identification. Knowing the relationships in a social network can be leveraged to identify an individual (Olteanu et al. 2016). This is part of a broader discussion on *interdependent privacy* in which the privacy of one individual is impacted by the privacy decisions and data sharing of others (Liu et al. 2018). As mentioned in the introduction, if my neighbor chooses to share personal information and an adversary knows that we live in close proximity, they could infer a lot of information (*e.g.*, race, income, education) about me.

With the increase in computational power and access to massive amounts of data, GeoAI techniques are able to re-identify records (*e.g.*, people) in datasets through inference and probabilistic modelling. For instance, large language models use AI techniques to process large volumes of textual data, much of which include geographic elements. Trained on such data, these models can be used to infer mobility patterns, identify individuals, and re-identified seemingly anonymized datasets based on the massive amount of additional (contextual) data on which they are trained. Such models are concerning to privacy advocates as public facing tools built from these models (*e.g.*, chat bots) give immense power to average citizens, power that can be used to reduce the privacy of individuals (Pan et al. 2020).

3. Application areas

While privacy is a pervasive concern through arguably all application areas of GeoAI, we thought it useful to highlight a subset of sectors in which privacy is at the forefront of the discussion.

3.1. *Advertising*

Location-based advertising involves targeting advertisements to groups and individuals based on their geographic location. A study of user attitudes towards targeted advertising found that targeted ads were generally preferred to non-target ones but targeted ads were seen as a privacy concern (Zhang et al. 2010). While not new, the adoption of context-aware devices and advanced in predictive analytics have changed the landscape of location-based advertising. An analysis of mobile device ad libraries found that a large number of them track a user's location (Stevens et al. 2012), even if the location is not needed for the functionality provided by a particular application. Location data, along with a variety of other attributes are used by AI companies for tailored advertising and to target particular users and groups (Boerman et al. 2017). In addition, the knowledge of someone's location can be combined with other factors such as the time of day or mode of transportation to further refine targeted ads and track users across devices and platforms. Studies have shown that location-based tracking works (Dhar and Varshney 2011) and given the importance of training data for advertising models, significant efforts are underway to collect and sell such data. As these data are transferred between data providers, brokers, and agencies, maintaining the privacy of the individual records often falls by the wayside. For instance, in 2019 the New York Times was provided access to detailed information, including locations, for 12 million mobile devices (Thompson and Warzel 2019). The source of the data was apparently unauthorized to share such content, yet the full records were shared without any attempt to preserve the privacy of the individuals in the data. Though not an advertising example, this does highlight the market for private data. While location-based advertising is unlikely to disappear in the near future, advances in GeoAI will enable advertisers and advertisees to strike a balance between privacy preservation and advertising utility.

3.2. *Health care*

A large percentage of the research on location privacy preservation and spatial anonymization was originally done for the purposes of maintaining data confidentiality in health. Understandably, medical researchers and practitioners are highly incentivized to maintain the confidentiality and privacy of patient data yet it is necessary to share data to access the collaborative expertise of those in the medical field. While geomasking and other obfuscation techniques are used to preserve data privacy as well as maintaining utility, newer methods are being developed that guarantee privacy while still permitting a level of analysis. As discussed in Section 2.3, cryptographic techniques such as homomorphic encryption are on the verge of dramatically changing how medical health records are stored and analyzed.

AI techniques are also being actively used in disease prevention and epidemiological research with impressive results (Munir et al. 2019). GeoAI too is having a significant impact with methods having been designed to model unique conditions such as spatial non-stationarity, variation in scale, and data sparsity. These are relevant to fields such as environmental epidemiology (VoPham et al. 2018), precision medicine, and healthy cities (Kamel Boulos et al. 2019). All of these fields have a strong privacy and confidentiality component and many of the models being developed today are designed with privacy in mind. These are often referred to as *privacy-aware* or *privacy-enhancing* technologies. As mentioned previously, models that deal with location data are particularly vulnerable to privacy inference attacks as knowledge of one's location allows for the inference of different characteristics. Not surprisingly, this has impacted the other side of the medical industry, namely health insurance. While some of us are aware that AI techniques are being used to analyze our driving records (Arumugam and Bhargavi 2019), we should also be conscious that they are being used to estimate risk and set health insurance rates (Naylor 2018).

The Covid-19 pandemic gave rise to a new era of health-related privacy concerns with many agencies and industry partners using AI for contact tracing (Grekousis and Liu 2021) and predicting outbreaks (Vaishya et al. 2020). During the Covid-19 pandemic, many of the privacy mechanisms that went into securing public and private health care data were reduced or removed to support contact tracing and epidemiological modelling efforts. Ribeiro-Navarrete et al. (2021) provide an overview of Covid-19 related privacy discussions and surveillance technologies.

3.3. *Security & surveillance*

The quintessential domain that one thinks of when discussing privacy in GeoAI is surveillance. Concern over AI technologies used to monitor citizens has received quite a bit of attention in the news media in recent years. This is not unwarranted but the relationship between AI and surveillance is more complex than it is often made out to be. There are plenty of examples in the literature of machine learning methods and tools that are used to track the locations of objects (*e.g.,* people, vehicles). Tracking technologies range from collecting locations of people through GNSS, Wi-Fi, or cellular trilateration, to license plate identification on traffic cameras. Other surveillance efforts monitor animal movement through image recognition for habitat delineation, conservation, and poaching prevention (Kumar and Jakhar 2022). Tracking or surveilling an object, by definition, involves the collection of information about that object and while the act itself is not a privacy violation, in certain circumstances, it can be. Aside from the actual data collection, AI has contributed to advances in how such tracking data are analyzed. Improvements in image classification and high performance computing mean that people can be monitored across different regions through CCTV surveillance cameras (Fontes et al. 2022). Tracking and surveillance can be less explicit as well. Existing research has demonstrated that humans are creatures of habit and are highly predictable in their activity behavior. Through the analysis of user-contributed and crowd-sourced data, *social sensing* techniques can be used to identify when and where someone may visit a place (Janowicz et al. 2019).

Tools and methods for crime prediction and counter-terrorism are often seen as being at odds with privacy preservation. The role of AI in crime forecasting specifically has received considerable interest in recent years (Dakalbab et al. 2022, Kounadi et al. 2020). Many of the techniques used in these fields are design for de-anonymization and re-identification in the name of safety and security. Most of the discussion related to privacy stems from surveillance being viewed as an infringement on individual rights. Given that criminal activity clusters geographically, one must be concerned about the privacy of one's data and, when the data are exposed, how that data is being used. A large body of research has investigated mass surveillance for security purposes and few results have indicated that AI models built on such data are more accurate at predicting crimes (Verhelst et al. 2020) or identifying repeat offenders (Dressel and Farid 2018). Work by Mayson (2019) demonstrated that the personal data used as input to such prediction models have dire consequences on the resulting actions taken by law enforcement. Predictive AI modeling has been shown to incorrectly identify individuals as criminals (Crawford and Schultz 2014) and that some AI predictive recidivism tools demonstrate concerning bias in their recommendations either as a result of the input data or the model designs.

4. The future of privacy in GeoAI

In this section we look to how privacy within GeoAI is changing and identify some of the leading concerns that should be addressed by the community. Specifically, we outline three ways in which privacy within GeoAI can be improved and highlight three emerging topics related to location privacy.

4.1. *Suggested areas for improvement*

While there are multiple ways that privacy can be further addressed within GeoAI, we provide the following three suggestions as starting points.

- **Privacy by design**. Despite the significant body of work on privacy from legal experts, policy makers, and ethical AI researchers, privacy concerns are still typically a secondary factor in the advancement of artificial intelligence. This is not only true for GeoAI, but for the broader field of AI and related technologies. Rather than being considered as an after thought, future directions of GeoAI research should integrate data privacy principles from the outset. Furthermore, data privacy should be considered at all stages of development from conception through delivery. Those with expertise in privacy and ethics should be consulted in the development and assessment of new algorithms that will impact the privacy of individuals or certain demographic groups. Privacy impact assessments (Clarke 2009) or audits, similar to ethics-based audits (Mökander and Floridi 2021), may be one such solution.

- **Spatial *privacy* is special**. Building off *Spatial is Special*, the alliterative phrase commonly uttered by geographic information scientists, there continues to be a need for wider acknowledgement within the artificial intelligence community that geographic data are unique due to the relationship between entity similarity and spatiotemporal proximity. This is particularly true when the privacy of an individual is at stake. Ignoring spatial properties of a dataset can substantially impact one's privacy (Griffith 2018). Working with geographic data requires an understanding of basic geographic concepts such as spatial heterogeneity, auto-correlation, and inference, and how they can be leveraged to either preserve or divulge private details.

- **Enhancing regulations**. Since data are the foundation on which virtually all AI technologies are built, access to such data for AI development should be scrutinized. Currently there is very little oversight or transparency on what types of data are collected, how they are collected, and how they are being used. We need independent assessment and inter-governmental regulations pertaining to data collection, storage, and its use. The European Union's General Data Protection Regulation (GDPR) is a good, but flawed first step. For instance, each European country is responsible for investigating the companies that are registered within it. This means that a country like Ireland is responsible for regulating a massive percentage of big tech. The actual number of penalties placed on violators as a result of the GDPR are much lower than predicted five years ago (Burgess 2022). Additional efforts must be made to ensure that users of digital platforms have the right to control how their data are collected, stored, and analyzed. The need for such transparency is paramount.

4.2. *Emerging privacy topics in GeoAI*

Aside from these recommendations there are a number of new challenges and emerging opportunities within GeoAI privacy research (Richardson et al. 2015). Some of these are actively being investigated while other are merely proposal for future research directions within this domain. Below we identify three directions that we feel are of particular interest to the GeoAI community.

- **Fake geospatial data**. The methods introduced earlier in this chapter highlight techniques for preserving the privacy of real people sharing real data. Synthetic data generation is one such approach, but new disinformation campaigns are focused on generating *fake* location data. Similar to how *deep fake* algorithms have emerged as practical tools for communicating disinformation visually, we are beginning to see similar approaches used to generating fake, but geospatially probable data. We are already seeing the emergence of a new subdomain of deep fake geography (Zhao et al. 2021). The reasons for generating fake location data include identity theft, political or social disruption, or bypassing security protocols. Note that fake data generation, while similar to synthentic data generation, is substantially different in its design and motivation. As our security tools increasingly relying on location information for verification (e.g., known IP address for banking), a new focus on detecting fake location information is required and the GeoAI community is well situated to address this challenge.

- **Publicly accessible and integrated tools**. We have only just scratched the surface in developing techniques for privacy preservation. As AI development and data availability grow, so too will the need for privacy preservation tools. Similar to how efforts are under way to detect text generated by large language model chat bots, we need publicly accessible tools to help users detect privacy violations and help users take control of their data. While many of the techniques and tools mentioned in this chapter are realized through theoretical models published by academics, real-world applications of these approaches have been slow to emerge. This is doubly true for methods generated by GeoAI developers. Future research will involve 1) the further integration of privacy preservation methods into existing location data sharing platforms and 2) more investment in the development of publicly accessible location privacy tools. Finally, educational efforts from geographers and computational scientists will focus on investigating the ways in which these tools educate and inform the public as to what is possible with personal location information.

- **Policy development**. From a social, political, and ethical perspective, future research will undoutably focus on developing policies in partnership with commercial entities and government agencies. Historically, government regulation and laws follow technological advances – often years behind. As highlighted in our suggestions above, regulatory bodies need to rise to the occasion, but these regulations need to be driven by evidence produced by ethical AI researchers and domain experts. As GeoAI emerges as it's own subdomain from within AI and geography, we have an opportunity to include the study of ethical and privacy implications within our research principles. The inclusion and reporting of such research will help inform regulators and policy makers when considering the impact of GeoAI on local communities and the global population.

5. Summary

In this chapter we presented an overview of data privacy as it related to geographic artificial intelligence. Geographic data are a unique type of information in that knowledge of one person's location reveals highly sensitive information about nearby individuals or groups. The growth of AI and associated techniques has forced researchers, companies, governments, and the public to think seriously about the privacy implications of sharing, collecting, and analyzing such data. Within GeoAI, particular attention needs to be made to how personal location and movement data are being analyzed and what can be inferred through geospatial analysis. A growing body of AI methods and tools are focused on privacy preservation with respect to geographic data within a wide range of domains. We encourage continued discussion on ethics and privacy as advances in GeoAI continue to shape the world around us.

References

Acar, A., Aksu, H., Uluagac, A. S., and Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)*, 51(4):1–35.

Adams, B. and Janowicz, K. (2012). On the geo-indicativeness of non-georeferenced text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 375–378.

Alanwar, A., Shoukry, Y., Chakraborty, S., Martin, P., Tabuada, P., and Srivastava, M. (2017). Proloc: Resilient localization with private observers using partial homomorphic encryption. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pages 41–52.

Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., and Palamidessi, C. (2013). Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 901–914.

Armstrong, M. P., Rushton, G., and Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in medicine*, 18(5):497–525.

Armstrong, M. P., Tsou, M.-H., and Seidl, D. E. (2018). 1.28-geoprivacy. In *Comprehensive geographic information systems*, pages 415–430. Elsevier Inc.

Arumugam, S. and Bhargavi, R. (2019). A survey on driving behavior analysis in usage based insurance using big data. *Journal of Big Data*, 6:1–21.

Beckman, R. J., Baggerly, K. A., and McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6):415–429.

Boerman, S. C., Kruikemeier, S., and Zuiderveen Borgesius, F. J. (2017). Online behavioral advertising: A literature review and research agenda. *Journal of advertising*, 46(3):363–376.

Brunila, M., McConnell, M., Grigg, S., Appuhn, M., Sumner, B., and Bohman, M. (2022). Drift: E2ee spatial feature sharing & instant messaging. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising*, pages 1–11.

Burgess, M. (2022). How gdpr is failing. *Wired Magazine.*

Charleux, L. and Schofield, K. (2020). True spatial k-anonymity: Adaptive areal elimination vs. adaptive areal masking. *Cartography and Geographic Information Science*, 47(6):537–549.

Clarke, R. (2009). Privacy impact assessment: Its origins and development. *Computer law & security review*, 25(2):123–135.

Crawford, K. and Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.*, 55:93.

Cunningham, T., Cormode, G., and Ferhatosmanoglu, H. (2021). Privacy-preserving synthetic location data in the real world. In *17th International Symposium on Spatial and Temporal Databases*, pages 23–33.

Dakalbab, F., Talib, M. A., Waraga, O. A., Nassif, A. B., Abbas, S., and Nasir, Q. (2022). Artificial intelligence & crime prediction: A systematic literature review. *Social Sciences & Humanities Open*, 6(1):100342.

Dhar, S. and Varshney, U. (2011). Challenges and business models for mobile location-based services and advertising. *Communications of the ACM*, 54(5):121–128.

Dobson, J. E. and Fisher, P. F. (2003). Geoslavery. *IEEE Technology and Society Magazine*, 22(1):47–52.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

Dubber, M. D., Pasquale, F., and Das, S. (2020). *The Oxford handbook of ethics of AI*. Oxford Handbooks.

Dwork, C. (2006). Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer.

Finn, R. L., Wright, D., and Friedewald, M. (2013). Seven types of privacy. *European data protection: coming of age*, pages 3–32.

Fontes, C., Hohma, E., Corrigan, C. C., and Lütge, C. (2022). Ai-powered public surveillance systems: why we (might) need them and how we want them. *Technology in Society*, 71:102137.

Gambs, S., Killijian, M., and del Prado Cortez, M. N. (2014a). De-anonymization attack on geolocated data. *J. Comput. Syst. Sci.*, 80(8):1597–1614.

Gambs, S., Killijian, M.-O., and del Prado Cortez, M. N. (2014b). De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*, 80(8):1597–1614.

Ghinita, G., Zhao, K., Papadias, D., and Kalnis, P. (2010). A reciprocal framework for spatial k-anonymity. *Information Systems*, 35(3):299–314.

Grekousis, G. and Liu, Y. (2021). Digital contact tracing, community uptake, and proximity awareness technology to fight covid-19: a systematic review. *Sustainable cities and society*, 71:102995.

Griffith, D. A. (2018). Uncertainty and context in geography and giscience: reflections on spatial autocorrelation, spatial sampling, and health data. *Annals of the American Association of Geographers*, 108(6):1499–1505.

Guha, S., Jain, M., Padmanabhan, V. N., et al. (2012). Koi: A location-privacy platform for smartphone apps. In *NSDI*, volume 12, page 14.

Harmanci, A. and Gerstein, M. (2016). Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature methods*, 13(3):251–256.

Hasan, R., Crandall, D., Fritz, M., and Kapadia, A. (2020). Automatically detecting bystanders in photos to reduce privacy risks. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 318–335. IEEE.

Hojati, M., Farmer, C., Feick, R., and Robertson, C. (2021). Decentralized geoprivacy: leveraging social trust on the distributed web. *International Journal of Geographical Information Science*, 35(12):2540–2566.

Janowicz, K., McKenzie, G., Hu, Y., Zhu, R., and Gao, S. (2019). Using semantic signatures for social sensing in urban environments. In *Mobility patterns, big data and transport analytics*, pages 31–54. Elsevier.

Jiang, H., Li, J., Zhao, P., Zeng, F., Xiao, Z., and Iyengar, A. (2021). Location privacy-preserving mechanisms in location-based services: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(1):1–36.

Jiang, J., Han, G., Wang, H., and Guizani, M. (2019). A survey on location privacy protection in wireless sensor networks. *Journal of Network and Computer Applications*, 125:93–114.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Kamel Boulos, M. N., Peng, G., and VoPham, T. (2019). An overview of geoai applications in health and healthcare. *International journal of health geographics*, 18:1–9.

Keßler, C. and McKenzie, G. (2018). A geoprivacy manifesto. *Transactions in GIS*, 22(1):3–19.

Kim, J. W., Edemacu, K., Kim, J. S., Chung, Y. D., and Jang, B. (2021). A survey of differential privacy-based techniques and their applicability to location-based services. *Computers & Security*, 111:102464.

Kounadi, O. and Leitner, M. (2014). Why does geoprivacy matter? the scientific publication of confidential data presented on maps. *Journal of Empirical Research on Human Research Ethics*, 9(4):34–45.

Kounadi, O. and Leitner, M. (2016). Adaptive areal elimination (aae): A transparent way of disclosing protected spatial datasets. *Computers, Environment and Urban Systems*, 57:59–67.

Kounadi, O., Ristea, A., Araujo, A., and Leitner, M. (2020). A systematic review on spatial crime forecasting. *Crime science*, 9(1):1–22.

Kumar, D. and Jakhar, S. D. (2022). Artificial intelligence in animal surveillance and conservation. *Impact of Artificial Intelligence on Organizational Transformation*, pages 73–85.

Liu, B., Zhou, W., Zhu, T., Gao, L., and Xiang, Y. (2018). Location privacy and its applications: A systematic study. *IEEE access*, 6:17606–17624.

Liu, Y., Liu, C., Lu, X., Teng, M., Zhu, H., and Xiong, H. (2017). Point-of-interest demand modeling with human mobility patterns. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 947–955.

Mayson, S. G. (2019). Bias in, bias out. *The Yale Law Journal*, 128(8):2218–2300.

McKenzie, G., Janowicz, K., and Seidl, D. (2016). Geo-privacy beyond coordinates. In *Geospatial Data in a Changing World: Selected Papers of the 19th AGILE Conference on Geographic Information Science*, pages 157–175. Springer.

McKenzie, G. and Mwenda, K. (2021). Identifying regional variation in place visit behavior during a global pandemic. *Journal of Spatial Information Science*, 1(23):95–124.

McKenzie, G., Romm, D., Zhang, H., and Brunila, M. (2022). Privyto: A privacy-preserving location-sharing platform. *Transactions in GIS*.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, 1(11):501–507.

Mökander, J. and Floridi, L. (2021). Ethics-based auditing to develop trustworthy ai. *Minds and Machines*, 31(2):323–327.

Munir, K., Elahi, H., Ayub, A., Frezza, F., and Rizzi, A. (2019). Cancer diagnosis using deep learning: a bibliographic review. *Cancers*, 11(9):1235.

Munjal, K. and Bhatia, R. (2022). A systematic review of homomorphic encryption and its contributions in healthcare industry. *Complex & Intelligent Systems*, pages 1–28.

Naik, N., Hameed, B., Shetty, D. K., Swain, D., Shah, M., Paul, R., Aggarwal, K., Ibrahim, S., Patil, V., Smriti, K., et al. (2022). Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Frontiers in surgery*, page 266.

Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE.

Naylor, C. D. (2018). On the prospects for a (deep) learning health care system. *Jama*, 320(11):1099–1100.

Nelson, T., Goodchild, M., and Wright, D. (2022). Accelerating ethics, empathy, and equity in geographic information science. *Proceedings of the National Academy of Sciences*, 119(19):e2119967119.

Nikolenko, S. I. (2021). *Synthetic data for deep learning*, volume 174. Springer.

Olteanu, A.-M., Huguenin, K., Shokri, R., Humbert, M., and Hubaux, J.-P. (2016). Quantifying interdependent privacy risks with location data. *IEEE Transactions on Mobile Computing*, 16(3):829–842.

Pan, X., Zhang, M., Ji, S., and Yang, M. (2020). Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE.

Pedersen, D. M. (1979). Dimensions of privacy. *Perceptual and motor skills*, 48(3_suppl):1291–1297.

Pensa, R. G., Di Blasi, G., and Bioglio, L. (2019). Network-aware privacy risk estimation in online social networks. *Social Network Analysis and Mining*, 9:1–15.

Rao, J., Gao, S., Kang, Y., and Huang, Q. (2020). LSTM-TrajGAN: A Deep Learning Approach to Trajectory Privacy Protection. In Janowicz, K. and Verstegen, J. A., editors, *11th International Conference on Geographic Information Science (GIScience 2021) - Part I*, volume 177 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 12:1–12:17, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Rao, J., Gao, S., Li, M., and Huang, Q. (2021). A privacy-preserving framework for location recommendation using decentralized collaborative machine learning. *Transactions in GIS*, 25(3):1153–1175.

Ribeiro-Navarrete, S., Saura, J. R., and Palacios-Marqués, D. (2021). Towards a new era of mass data collection: Assessing pandemic surveillance technologies to preserve user privacy. *Technological Forecasting and Social Change*, 167:120681.

Richardson, D. B., Kwan, M.-P., Alter, G., and McKendry, J. E. (2015). Replication of scientific research: addressing geoprivacy, confidentiality, and data sharing challenges in geospatial research. *Annals of GIS*, 21(2):101–110.

Riederer, C., Kim, Y., Chaintreau, A., Korula, N., and Lattanzi, S. (2016). Linking users across domains with location data: Theory and validation. In *Proceedings of the 25th international conference on world wide web*, pages 707–719.

Seidl, D. E., Jankowski, P., and Tsou, M.-H. (2016). Privacy and spatial pattern preservation in masked gps trajectory data. *International Journal of Geographical Information Science*, 30(4):785–800.

Stahl, B. C. and Wright, D. (2018). Ethics and privacy in ai and big data: Implementing responsible research and innovation. *IEEE Security & Privacy*, 16(3):26–33.

Stevens, R., Gibler, C., Crussell, J., Erickson, J., and Chen, H. (2012). Investigating user privacy in android ad libraries. In *Workshop on Mobile Security Technologies (MoST)*, volume 10, pages 195–197.

Sun, G., Cai, S., Yu, H., Maharjan, S., Chang, V., Du, X., and Guizani, M. (2019). Location privacy preservation for mobile users in location-based services. *IEEE Access*, 7:87425–87438.

Swanlund, D., Schuurman, N., and Brussoni, M. (2020). Maskmy. xyz: An easy-to-use tool for protecting geoprivacy using geographic masks. *Transactions in GIS*, 24(2):390–401.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570.

Thompson, S. A. and Warzel, C. (2019). Twelve million phones, one dataset, zero privacy. In *Ethics of Data and Analytics*, pages 161–169. Auerbach Publications.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240.

Vaishya, R., Javaid, M., Khan, I. H., and Haleem, A. (2020). Artificial intelligence (ai) applications for covid-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4):337–339.

Verhelst, H. M., Stannat, A., and Mecacci, G. (2020). Machine learning against terrorism: how big data collection and analysis influences the privacy-security dilemma. *Science and engineering ethics*, 26:2975–2984.

VoPham, T., Hart, J. E., Laden, F., and Chiang, Y.-Y. (2018). Emerging trends in geospatial artificial intelligence (geoai): potential applications for environmental epidemiology. *Environmental Health*, 17(1):1–6.

Wang, Q., Gao, J., Lin, W., and Yuan, Y. (2019). Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8198–8207.

Wang, R., Zhang, M., Feng, D., Fu, Y., and Chen, Z. (2016). A de-anonymization attack on geo-located data considering spatio-temporal influences. In *Information and Communications Security: 17th International Conference, ICICS 2015, Beijing, China, December 9–11, 2015, Revised Selected Papers 17*, pages 478–484. Springer.

Zhang, H., Guerrero, C., Wheatley, D., and Lee, Y. S. (2010). Privacy issues and user attitudes towards targeted advertising: A focus group study. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 54, pages 1416–1420. SAGE Publications Sage CA: Los Angeles, CA.

Zhang, H. and McKenzie, G. (2022). Rehumanize geoprivacy: from disclosure control to human perception. *GeoJournal*, pages 1–20.

Zhao, B., Zhang, S., Xu, C., Sun, Y., and Deng, C. (2021). Deep fake geography? when geospatial data encounter artificial intelligence. *Cartography and Geographic Information Science*, 48(4):338–352.

Zhong, Y., Yuan, N. J., Zhong, W., Zhang, F., and Xie, X. (2015). You are where you go: Inferring demographic attributes from location check-ins. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 295–304.