AI versus AI for democracy: exploring the potential of adversarial machine learning to enhance privacy and deliberative decision-making in elections

Auliya, Syafira Fitri; Kudina, Olya; Ding, Aaron Yi; Poel, Ibo Van de

DOI
10.1007/s43681-024-00588-2

Publication date
2024

Document Version
Final published version

Published in
AI and Ethics

AI and Ethics
https://doi.org/10.1007/s43681-024-00588-2

ORIGINAL RESEARCH

AI versus AI for democracy: exploring the potential of adversarial machine learning to enhance privacy and deliberative decision-making in elections

Syafira Fitri Auliya[1] · Olya Kudina[1] · Aaron Yi Ding[2] · Ibo Van de Poel[1]

Received: 31 May 2024 / Accepted: 17 September 2024

Abstract
Our democratic systems have been challenged by the proliferation of artificial intelligence (AI) and its pervasive usage in our society. For instance, by analyzing individuals' social media data, AI algorithms may develop detailed user profiles that capture individuals' specific interests and susceptibilities. These profiles are leveraged to derive personalized propaganda, with the aim of influencing individuals toward specific political opinions. To address this challenge, the value of privacy can serve as a bridge, as having a sense of privacy can create space for people to reflect on their own political stance prior to making critical decisions, such as voting for an election. In this paper, we explore a novel approach by harnessing the potential of AI to enhance the privacy of social-media data. By leveraging adversarial machine learning, i.e., "AI versus AI," we aim to fool AI-generated user profiles to help users hold a stake in resisting political profiling and preserve the deliberative nature of their political choices. More specifically, our approach probes the conceptual possibility of infusing people's social media data with minor alterations that can disturb user profiling, thereby reducing the efficacy of the personalized influences generated by political actors. Our study delineates the boundary of ethical and practical implications associated with this 'AI versus AI' approach, highlighting the factors for the AI and ethics community to consider in facilitating deliberative decision-making toward democratic elections.

Keywords Deliberative decision-making · Elections · Privacy · Adversarial machine learning · Artificial intelligence

1 Introduction

The advancement of artificial intelligence (AI) is a double-edged sword. While AI applications benefit various areas of human life, their development presents broad challenges to many sectors, including constitutional rights. It has even been argued that the greatest social cost of AI may be an "erosion of trust in (..) our democratic institution" [1], owing to hypernudging, microtargeting, deepfakes, algorithmic bias, unexplainable AI, cyberattacks on elections, and digital gerrymandering [1–5]. In these discussions, one of the sub-values of democracy that frequently appears to be at risk with AI is privacy.

Philosophically, privacy and democracy are related in the sense that privacy encourages participation in democracy. Some researchers argue that having a sense of privacy allows people to practice their value of autonomy, specifically the ability to self-governance, self-determination, and self-development [1, 4, 6, 7]. All of these, we contend, are necessary for achieving the trust and participation of individuals in democratic practices. Moreover, we argue that a sense of privacy creates spaces for individuals to take a step back, reflect, and make their own deliberative decisions, which is crucial in a democracy[1]. In this paper, we

✉ Syafira Fitri Auliya
s.f.auliya@tudelft.nl

[1] Section Ethics and Philosophy of Technology, Department of Values, Technology, and Innovation, Faculty of Technology, Policy, and Management, TU Delft, Jaffalaan 5, Delft 2628 BX, the Netherlands

[2] Department of Engineering Systems and Services, Faculty of Technology, Policy, and Management, TU Delft, Jaffalaan 5, Delft 2628 BX, the Netherlands

[1] We do not argue that privacy is the sole aspect of democracy. Privacy is just one element among other interconnected variables that collectively construct democracy. Our argument emphasizes that having a greater sense of privacy can be a significant catalyst for people to

Published online: 16 October 2024

define democracy roughly as "a method of collective decision-making characterized by a kind of equality among the participants at an essential stage of the decision-making process" [8], with each participant "having equal and effective opportunities for learning about the relevant alternative policies and their likely consequences" [9]. Forms of democracy such as deliberative democracy emphasize equality and require "[collective] decision making by discussion among free and equal citizens" [10]. Thus, equal participation in the decision-making process is the defining characteristic of democracy. To achieve this equality, citizens must have autonomy in their political participation and be able to make independent, reflective decisions. In this paper, we shall refer to this set of practices as "deliberative decision-making."

Numerous examples have emerged in recent years illustrating how privacy violation disturbs people's ability to make deliberative decisions, especially when AI is concerned. For instance, AI has a history of being used to influence people's choices in elections using unauthorized use of data. Cambridge Analytica is the best-known case in this context, with the private information of millions of users harvested without their consent to build psychographic profiles of voters. The profiles were then utilized to create microtargeting propaganda that affected the results of the US election and the UK Brexit referendum [11, 12]. Similarly, scholars have frequently demonstrated that publicly available data, such as social media data, can be used to accurately profile and predict individuals' political affiliation [13–18], which may become an input for some political actors to influence people into certain political agendas and prevent them from making deliberative decisions [3].

Because of these previous negative experiences, it is understandable that, predominantly, there is a focus on solving the negative influences of AI on the privacy of deliberation. Nevertheless, depending on how we conceptualize AI philosophically, AI can also be conceived of as a mediator or even an enabler of deliberative decision-making. Such a conception allows for a broader outlook on the potential of AI in facilitating better deliberative decisions, enabling scrutinizing with more nuanced interrelationships between AI, privacy, and democracy. Broadening the understanding of AI in such a way does not pretend to fix all the complex issues with AI. Rather, exploring the possibilities of technology can help us to become more reflective regarding its effects on democracy. Thus, instead of looking at the simplistic neutral-negative terms only when AI, privacy, and democracy are concerned, we need to consider a more complex position, acknowledging that AI always embodies certain values, affording a diverse range of societal implications [19].

In this paper, we propose an unconventional way of exploring AI's potential to facilitate better deliberative decision-making in non-ideal societies where political actors may influence our political agendas. More specifically, we focus on the privacy needed for deliberation in making a voting decision during elections, which is a nascent privacy response, i.e. the development of adversarial machine learning (AML). AML is a rising subject in the AI domain, whereby AI systems produce incorrect results as the result of intentionally providing them with deceptive input [20]. In this paper, we conceptualize the potential of AML for facilitating privacy in electoral decision-making. We especially focus on social media, where the algorithmic curation of user interaction is recognized and prevalent to influence one's decision-making ability in elections. In doing so, we position malicious AI systems deployed on social media (e.g., Twitter) to profile the users into clusters of certain political affiliations, making users vulnerable to targeted political advertising. In parallel, we conceive of AML as a potential approach to fool these malicious AI systems that attempt to profile and influence users into specific political agendas. Thus, the main question tackled in this paper is: *how can the design of AI-driven privacy-enhancing technology on social media platforms help facilitate deliberative decision-making in elections?*

To this end, we will build on the method of Value Sensitive Design (VSD), which starts from an assumption that all technologies explicitly or implicitly embed certain values and allows one to trace how those values materialize across different iterative states of design: conceptual, empirical, and technological [21]. In this paper, we primarily elaborate on the first conceptual stage of the study related to enhancing the value of privacy in how people make their deliberative decision-making in elections with the help of AML. Nonetheless, doing so will inevitably lead us to discuss also the technological stage, albeit in a preliminary rather than a definitive manner, e.g. related to understanding the potential and limitation of AML in this approach, relating it explicitly to the value of privacy and deliberative decision-making without providing yet final design decisions.

This paper is structured as follows. In Chap. 2, we discuss our underlying assumption that technology embodies values and emphasize the importance of responsibly designing technology. We then discuss the values pertinent to the design of AI technologies in connection to elections and specifically focus on the value of privacy in social media-based deliberative processes in Chap. 3. In Chap. 4, assuming that privacy plays a significant role in facilitating deliberative decision-making in elections, we explore the role of AI in this context and introduce the potential application of AML.

deliberately engage in democratic practices, leading to better democratic outcomes.

Chapter 5 discusses potential ethical and practical implications associated with the approach, which serve as a foundation for the future stages of the larger study. We conclude in Chap. 6 with reflections on the study and delineate the points for follow-up research.

2 Values in the design of technologies

The growing prevalence of technology has blurred the line between humans and technology. The extent to which technologies are integrated into human life is a subject of debate among academics. Peter-Paul Verbeek, for example, categorizes three modes of human-technology relationships [22]. In the first category, technology is viewed as neutral and proposes morally unjustifiable explanations such as "the machine made me do it" and "the gun caused the murder" [23]. In the second category, technologies are viewed as the externalization of humans [24]. The final category, which we agreed on, does not separate humans and technology; instead, it conceptualizes technologies as mediators of the interrelationship between humans and their worlds. The use of technology allows people to perform tasks that would not otherwise be possible. In the scenario involving the gun, the firearm does not have the intention to kill anyone. That intent belongs to the human. However, to some extent, the presence of a gun influences the person's decision to carry out their deadly intention, owing to the specific design of the artifact that makes it possible to pull up the trigger.

Thus, designers, in particular, must anticipate the direct and indirect implications of the technology that they create [25]. However, the implications of the new technologies are hardly fully predicted ahead of time. In some instances, consequences are not discovered until after the proposed technology has been widely adopted by the public and harder to alter– a classic dilemma known as the "Collingridge dilemma" [26]. The utilization of new technologies, which may result in unanticipated and unintended consequences, is frequently uncertain until the technologies are implemented and widely adopted [27, 28]. However, the fact that it is difficult to forecast all of the consequences of technology does not mean that we should not attempt to address this issue or create responsible sociotechnical experiments [29]. As technology is not neutral, but rather value-laden, it means that we can intentionally incorporate values into technologies. So, it is vital to ensure that it is acting as a mediator between humans and the world and that it is designed responsibly, and we must seek to maximize our efforts to predict any future consequences of the technology. One approach to responsibly design technologies by integrating values and considering the future consequences of the technology developed is by adopting a certain development framework, such as Value Sensitive Design (VSD).

VSD facilitates the design of technologies that incorporate values by embedding them throughout the design process. VSD acknowledges that design choices made during technology development inevitably reflect the values of the creator and manifest in various ways. In VSD, the designer of technologies can proactively consider the implications of their choices "that accounts for human values in a principled and comprehensive manner throughout the design process" [30]. VSD utilizes an iterative approach incorporating conceptual, empirical, and technical investigations. During the conceptual phase, VSD examines the values held by stakeholders. It considers how these stakeholders are impacted and which values are involved. Additionally, it explores how to handle the trade-offs that arise when competing values are at stake. The empirical phase involves investigating human activities associated with technology through the application of qualitative or quantitative methods. Technical investigations involve the proactive design of a system that will support the values identified in the conceptual phase. This phase also entails thoroughly investigating the issues raised by existing technological solutions.

In this paper, we primarily focus on the conceptual phase, exploring the potential values linked to deliberative decision-making in elections, in which we address that privacy in social media has a link to this narrative. Yet, this paper also serves as a preliminary technological investigation aimed at understanding the potential and limitations of AML in relation to the value of privacy and deliberate decision-making, lays the groundwork for future empirical investigations with relevant stakeholders and the subsequent technological investigation to refine the final design decisions in the next stages.

3 Deliberative decision-making and privacy in contemporary elections

3.1 Deliberative decision-making

We understand deliberative decision-making as a catalyst for deliberative democracy. Deliberative democracy is often defined as "[collective] decision making by discussion among free and equal citizens" [10]. Thus, equality in the discussion is important in deliberative democracy, as equality "[is the] key principle of deliberative democracy… [and] every person who is engaged in deliberation may have an equal opportunity to speak…" [31] and discussions serve as a tool for "producing reasonable, well-informed opinions in which participants are willing to revise preferences in light of discussion, new information, and claims made by fellow

participants" [32]. As deliberative democracy requires equal discussion, some scholars envision that the decisions made in ideal deliberative democracies should be made through small-scale discussions, or "mini-publics,", and require face-to-face communication. However, in this paper, we focus on the possibility that deliberative democracy can also occur in large-scale societies, such as aggregate societies that hold general elections. We defer to Robert Goodin and his principle of "first talk, then vote," which has become one of the central principles in modern democracies and gives significant depth to the theories of deliberative democracy [33]. In a modern democracy, the legitimacy obtained from deliberation comes not from the sole act of voting, but from the talk and general deliberation that precede voting [34, 35].

However, if legitimacy is to be obtained through deliberation, it is necessary to recognize the risk of manipulation occurring in the deliberation processes [34]. We live in a society in which there are continuously external interferences in our decision-making. Consequently, there are always opportunities to be manipulated. Nonetheless, deliberative decisions can still be made in such circumstances if we allow people enough space for their own deliberation and decision-making. So, if we assume that there are always interferences that might manipulate people, how can we still maintain deliberative decision-making?

The term "nudge" is useful here as it refers to the ways in which people are being influenced to alter their behavior in specific ways. The term "nudge" was popularized by Richard Thaler and Cass Sunstein in their book of the same title. They define a nudge as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" [36]. The concept of nudge has received much attention: empirical studies examining its effectiveness have been conducted in a variety of fields, such as organ donation, tobacco warning labels, and healthier consumption decisions [37–39]. Nonetheless, the nudge theory also has met with various criticisms. Research has linked nudges with the demotion of central moral values, and intentional nudges may be considered a form of manipulation or even power abuse [40]. Scholars who study manipulation, such as Michael Klenk, associate manipulations with an "intentional lack of care" for people to understand the bigger picture of their actions [41]. According to Klenk, even if the targeted individual is well aware of the underlying influence process or if the manipulators have made their aim to influence clear, it is still considered manipulative if the influencers intentionally lack care to help their target comprehend to consequences of their actions.

This discussion of nudges and manipulation is relevant because, when discussing deliberative decision-making in today's election process, it is especially important for participants to form critical and reflective opinions before making decisions. Ideally, given people's engagement with social media, people should be assisted in understanding the consequences of their social-media engagement in order to engage in deliberative decision-making. Ideally, under no circumstances should people be prevented from making deliberative decisions by manipulators taking advantage of social media data. However, we do not live in an ideal society. Social media data can be exploited to create individual profiles of users and to aid manipulative entities in influencing people to support certain agendas. Without people being helped to understand the consequences of their social media engagement, various AI algorithms are continually manipulatively nudging people to alter their decisions and follow certain political agendas, weaponized by the knowledge obtained from processing social media data, as will be explained in Sect. 3.3.

Therefore, we explore how individuals can make deliberative decisions in the context of elections when we are living in a world where AI algorithms actively influence our active engagement with social media. To promote deliberative decision-making in a world in which manipulative nudges are continuously being made, certain mechanisms are required to preserve people's ability to manage these influences while still being able to use social media and participate in the election process. We propose the privacy of social media data as one mechanism that could assist people in dealing with these manipulative nudges. Rather than only using the traditional widespread perception of privacy related to control over data or information, which will elaborate more in Sect. 3.2, we are interested in how privacy can afford individuals the space to reflect, as so act as a catalyst for making deliberative decisions in elections. This idea is new for two reasons:

First, scholars assert that unconscious feelings and emotions have significant roles in decision-making, often even preceding cognitive justifications [42–49]. This impact of unconscious judgment might be magnified by political actors using AI and social media data to influence people's decision-making, leading to phenomena such as echo chambers, filter bubbles, nudges, and microtargeting [44, 47, 50–53]. In order to counter these influences, particularly those that manipulate emotions during the decision-making process, scholars have proposed various response strategies, including public education, transparency, and early intervention [44, 45, 48, 54]. However, to the best of our knowledge, none of these works make explicit reference to privacy as a means to give people space to reflect, enabling them to assess emotional influences alongside their own cognitive justification before making decisions, as one of the potential remedies.

Second, a strategy for giving people space to think and reflect before voting has indeed been proposed by some scholars [45, 50], but their strategy emphasizes the need for public education. We contend that education alone is insufficient in providing people the space to reflect before making election decisions. This is due to the complexity of current manipulative political influences, which may go unnoticed even by educated individuals. In addition, owing to the previously stated definition of manipulation that does not require the target to be oblivious to the manipulation, people can still be manipulated even if they are capable of detecting the manipulations. Thus, merely educating them about manipulations is inadequate to enhance their deliberation in elections. Another form of endeavor to strengthen the space for deliberation is necessary. In this paper, we conceptualize privacy as one of the possible solutions. Owing to DeCew and Solove's description of a privacy invasion as "by being forced to hear propaganda, by being manipulated by subliminal advertisements, or by being disrupted by a nuisance that thwarts one's ability to think or read" [55, 56], improving privacy will make people to exert control over the propaganda they hear, to not being manipulated by subliminal influences, and to maintain their ability to think– including before making their decision in elections.

3.2 Privacy and deliberative decision making

Multiple scholars have associated privacy with a sense of control. Charles Fried declares that "privacy is not just an absence of information about us in the minds of others; rather it is the control we have over information about ourselves" [57]. According to Alan Westin, "privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" [58]. Article 8 of the European Charter of Fundamental Rights states that "everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified" [59]. While we agree that protecting one's privacy requires control of oneself, the majority of privacy-related theories focus on control over one's data and information. This limitation makes these theories too narrow for our purpose and overlooks the aspects of privacy pertaining to the fundamental ability to exercise control over our decision-making.

In this paper, we agree with Solove and DeCew that the disclosure of new data or information is not a prerequisite for privacy invasion. Instead, privacy can also be invaded "by being forced to hear propaganda, by being manipulated by subliminal advertisements, or by being disrupted by a nuisance that thwarts one's ability to think or read" [55, 56]. Therefore, despite heavy criticism for failing to "inform us about the matters in which we should be let al.one" [56] or failing to distinguish between "a punch in the nose. [and] a peep in the bedroom" [60], Judge Cooley's definition of privacy as "the right to be left alone" [61] is, in this paper, considered to be the core element of privacy. The essence of privacy, we argue, is the ability to make decisions about our lives even when others are not only watching us but are nevertheless potentially interfering with us. When we have some degree of privacy, regardless of whether we are aware of any ongoing interference, we are able to step back and create spaces for ourselves to reflect and make decisions.

To clarify, we do not argue that privacy equals isolation. Indeed, an individual living in isolation would not be concerned about any invasion of privacy. Rather, privacy is significant because we live within a society. As summarized by Moore, "without society there would be no need for privacy" [62]. Further, some scholars have explained that the purposes of privacy include the protection of "intimacy" [57] and "individuality" [63], both of which stem from our position as humans in society.

3.3 Privacy and deliberative decision-making in contemporary elections

In today's election processes, people ironically believe that they are deliberative and uninterrupted when—in fact—they are being observed and influenced by numerous entities. They often hold the belief that they read news X of their own volition and attend event Y out of personal interest. As stated in the definition of manipulative nudge in Sect. 3.1., these individuals are left unaware that their data, which was gathered from various sources including due to their engagement in social media, enabled other entities to tailor political materials according to their personal preferences and can result in the adoption of certain political viewpoints. For instance, news X is intentionally made visible to them, and the promotion of event Y is made appealing to them, which leads to an internalization of some political beliefs.

The largest-scale example of this phenomenon is that of Cambridge Analytica. Cambridge Analytica, in partnership with a company led by a Cambridge University academic, paid hundreds of thousands of users to complete a personality test for academic purposes. The app "thisisyourdigitallife" was used to collect data, along with a clear consent form for academic use of the data. However, in March 2018, Christopher Wylie disclosed a major data breach committed by Cambridge Analytica, his former employer. He leaked insider documents and reported the practices that his former company had allegedly used to illegally harvest 50 million profiles from Facebook by collecting information about the test-takers' Facebook friends. The data was used to develop psychographic profiles and create custom-tailored propaganda materials, delivered to prospective voters, which

aided in the success of numerous campaigns worldwide, including Brexit and Trump's election [11, 12, 64]. The Christopher Wylie quotation that follows is illuminating [64]:

> *We would know what kinds of messaging you would be susceptible to, including the framing of it: the topic, the content, the tone.… What you would be susceptible to, and where you are going to consume that, and then how many times did we need to touch you with that in order to change how you think about something.…. [We] then create that content, that then gets sent to a targeting team, which then injects it into the Internet.…**until they start to think something differently.** … [We] are whispering into the ear of each and every voter and [we] may be whispering one thing to this voter and another thing to another voter."*

Not only does Christopher's testimony imply that a popular social-media company that stores millions of people's personal information and interactions has failed to protect its users' data privacy, but also that social-media data can influence the political agendas of major nations. Furthermore, without Christopher's witness statement, it is unlikely that people would have recognized that major political decisions were being manipulated by intentionally creating individualized content targeting potential voters.

Compared to traditional political campaigns, this scandal violated privacy and hindered individuals' deliberative decision-making in elections to higher degrees, mainly due to its enormous-scale unauthorized access and utilization of personal data without explicit concerns. Unlike other legitimate data-driven influences that rely on aggregated non-personal data or consented personal data, Cambridge Analytica's major scandal stems from their unlawful acquisition of millions of Facebook data without the data owners' concerns, directly violating their right to be left alone. Furthermore, the perpetrators for Cambridge Analytica demonstrated a clear absence of intention to disclose the potential repercussions of the subtle influences they exerted on people. As mentioned earlier in Sect. 3.1, the process of deliberative decision-making in elections may be hindered by manipulation. This manipulation transpires when individuals are not provided with an explanation of the consequences of their data and the nudges exerted on them.

Furthermore, unlike the authorized advertising industry's business model, which similarly creates profiles of prospective customers using various data points, the creation of political profiles of individuals goes beyond existing legal limits. As per the provisions outlined in Article 9 of the GDPR, the processing of personal data that reveals sensitive personal data, such as political opinions, is explicitly prohibited. Even major social media platforms, which often monetize their users' data, now explicitly prohibit the disclosure of users' political profiles through their data. Nonetheless, privacy protection, even in its broad context, in the modern era is a challenge. Non-expert citizens are generally unaware that their "simple" data can yield in-depth knowledge about them. Simple daily data, such as that produced by smartphones, can reveal detailed information about individuals [65]. Moreover, corporations frequently instill in users a false dilemma, an "either-or" in relation to privacy: specifically, users are presented with the "option" of *either* complying with the company's policy and renouncing their privacy, which legally grants the company permission to use their personal data, *or* not using the product at all. Another issue arises when individuals knowingly grant other parties access to their data and information in exchange for certain benefits. This "privacy paradox" reflects a dichotomy between attitudes toward privacy and actual behavior, in which people express a high level of concern regarding their privacy but are willing to expose themselves for small rewards or conveniences [66–68]. Therefore, conflicting values—regarding privacy, individual comfort, company profit, and so on—constantly emerge. This phenomenon is known to scholars in related subjects, and numerous works have attempted to address this issue. Helen Nissenbaum states, for instance, that "[the proponents of the technological systems] must be able to address, in systematic ways, conflicts between privacy and competing values served by the offending technologies" [69]. However, in the real world, active and transparent encouragement of privacy is a rarity in the business sector, as advocating privacy rarely brings immediate monetary advantages. To reconcile these conflicting values and avoid any dilemma or privacy paradox arising in the first place, especially those linked to people's ability to make deliberative decisions in elections, this paper recommends a responsible technological design as a potential solution to be considered.

4 Introducing the possibility: altering people's social media behaviors to fool profiling AI systems

As we have seen, deliberative decision-making in elections can be facilitated by enhancing privacy—in the sense that privacy means leaving people alone and providing space for them to slowly digest information, think, and reflect on their own political stance before voting. Hence, even if external forces are continually influencing individuals and may manipulate their emotions, this space will enable them to be critical and reflective prior to voting. Numerous examples exist of current AI systems that endanger privacy

and circumvent this space by profiling individuals and by seeking to influence them. Due the widespread use of social media, people's tendency to express their political opinions on social media rather than in person, and the fact that people are now obtaining more information from social media than from traditional news sources, these AI systems often use social media as their primary data source.

There have been several efforts to protect the privacy of social-media data. From a technical standpoint, many advanced privacy-preserving algorithms have been introduced to enhance privacy on social media. For example, in a recent study, Jiang et al. demonstrate that differential privacy—which is advocated by some as the most effective method and which works by injecting noise into a system such that the output cannot be used to infer much about individuals—can also be achieved in social-media networks [70]. However, this kind of approach relies primarily on social media service providers and provides no option for consumers to actively participate in and oversee the process. Moreover, we should not rely solely on providers but should instead have other options to avoid the worst-case scenarios in the events that the providers' privacy solutions are inadequate. In addition, providing the feeling of control over their privacy to users would also increase their trust and engagement [71].

Therefore, in this paper, we explore the possibility of enhancing the privacy of social media through user participation. Specifically, we investigate the possibility of employing "adversarial machine learning" (AML) to manage the threat of AI systems that violate the privacy of users and impede the deliberative decision-making process in relation to elections. AML is a growing area of computer science that involves fooling AI systems by using the characteristics of the deep neural networks employed in AI systems, making use of the fact that, while a little disturbance cannot affect the item category of an image in object-recognition tasks, an unnoticeable non-random disruption to a test image might arbitrarily alter the network's prediction [72]. Although researchers believe that adversarial attacks can degrade utility and infringe privacy—for example, by obtaining access to training data and manipulating the model's outputs [73]—we think the technique is worth exploring. Specifically, we explore the possibility of employing this concept to enhance privacy by impeding the capabilities of other AI systems that pose a threat to privacy. As we believe that the power of AI systems will continue to grow exponentially in the future, the cognitive capacity of humans will be incapable of balancing the development of these systems. Therefore, we believe it is worthwhile to investigate the possibility of using *other* AI systems to counter these AI systems' threats.

There are multiple ways of using AML to protect privacy while ensuring that the data is still statistically relevant to the original data—for example, generating synthetic data, reducing the amount of stored information, and adding noise to data. In this paper, we explore the conceptual elucidation of a model that uses the first of these options—namely, the generation of synthetic social media data. Particularly, we assume that AI systems play a role in generating user profiles by analyzing social media data, and external entities, such as politicians or political consultants, may use the user profiles as a foundation to generate personalized influences in order to steer the targets toward specific agendas. By exploring the possibility of using AML for this purpose, it might be possible to fool AI systems and make them generate wrong profiles about their intended targets. User profiling is a privacy violation, as it disregards the advice of Warren and Louis to leave users alone and collects more information than the users have voluntarily revealed. Assessing a user's profile, these entities generate influences specifically suited to that user. As DeCew and Solove explain in relation to forcing people to hear propaganda and manipulating them with subliminal marketing, providing individuals with personalized information to persuade them is also a breach of privacy. These entities exert highly personalized influences on their targets, exposing them to propaganda until—as in the case of Cambridge Analytica—the target "start[s] to think differently" [64]. As people's thoughts begin to shift as a result of external influences, their capacity to make deliberative decisions is impaired.

Survey papers on AML provide numerous classifications of the concept, with some based on the goals of the attackers, some on the threat models, and some on the targeted phases [73–76]. But in a broad sense, adversarial attacks can be categorized as "poisoning attacks," in which the adversary manipulates the training data to degrade the performance of the targeted AI model, or "evasion attacks," in which the adversary manipulates the data to deceive trained classifiers [77]. The AI model that is under attack will perform poorly in the first category but not in the second. In this paper, our exploration of AML operates under the second classification by misleading classification results *without* diminishing the capacity of the AI being targeted. Particularly, we investigate the possibility of generating synthetic social media data of individuals in order to mislead the classification result of AI systems attempting to build user profiles based on these individuals' social media data. Because our approach is part of evasion attacks, the AI systems that will be tricked will not be harmed. For instance, if *John* uses the above-mentioned approach to modify his social-media data in the hopes of preventing influencers from gathering accurate data on him and, thus, creating a stronger space for himself to think before voting, only John's classification results (in

this case, his "user profile") will be affected, which results in the creation of an erroneous profile of John and weaker influences upon him. However, because it is not a poisoning attack, it will not hinder this profiler AI system's ability to create profiles of *other* social media users. To be more specific, we explore the possibility of modifying a small portion of users' social media data in response to external influences. As suggested by the concept of AML, this subtle change should be undetectable by humans or straightforward data analysis. Nonetheless, this minor modification might affect the classification output of AIs attempting to infer the political affiliations of targeted individuals. This preventive counterattack strategy will vary depending on the degree of knowledge and information access of the AI. In white-box (counter)attacks, where we have access to and knowledge of the targeted AI's model, the attempts will be simpler than with black-box AI, where knowledge about the targeted model is unavailable.

To the best of our knowledge, the utilization of synthetic social media data to mislead the classification result of profiler AI, with a focus on safeguarding privacy and facilitating deliberative decision-making in elections, is a novel way to use AML. Previous research on putting adversarial attacks on social media platforms connects them with deceiving models detecting rumors, fake news, spam, hate speech, and sentiment analysis [78]. However, none of them specifically targeted the manipulation of political profiler AI systems reliant on social media data. A notable parallel in AML implementation is Fawkes, which helps individuals add imperceptible cloaks in their own pictures to avoid unauthorized facial recognition models and, thus, enhance their privacy [79]. However, Fawkes operates within the realm of image-based and does not extend to social media or political contexts. Therefore, the exploration of AML in the direction outlined in this paper holds promise for advancing the field and critically addressing the issues of privacy and deliberative decision-making in elections.

However, adapting AML to facilitate privacy and deliberative decision-making in elections would require different approaches compared with existing works about AML-based technological solutions. Unlike more straightforward applications of AML in existing literature, implementing AML in our purpose requires more nuanced and interdisciplinary approaches. For example, while facial recognition protection in [79] straightforwardly adds noise to images to confuse detection algorithms, political profiling on social media involves broader and more complex data sources. To illustrate, profiler AI systems may analyze non-political data, such as posts about religious activities, to infer citizens' political preferences. This data can be combined with other parameters that can be derived online, such as age, gender, and issues of interest, to infer about political choices. Consequently, AML-based technologies need to obscure not only direct political content but also understand the nuance and often seemingly unrelated data that can infer political profiles from people. Additionally, the correlation between political preferences and individuals' data may differ depending on the national context. Some data on social media might indicate stronger political preferences in one country but not in another, requiring AML-based solutions that are attentive to the national context to determine which data to protect to obscure individuals' political preferences.

In addition, adding another layer of complexity, people's election decisions are often complex and not only influenced by direct exposure to political messages but also shaped by interpersonal trust and the alignment of the messages with preexisting beliefs held by these individuals or their social groups [80, 81]. AML-based technologies may reduce the precision of manipulative micro-targeting political campaigns that target the existing individuals' susceptibility, providing people with less exposure to campaigns targeting this susceptibility that often stirs their emotions and obstructs reflective decision-making, thus offering people more opportunities to engage with a broader range of nuanced information. Yet, to encourage people to go outside their preexisting polarization and construct social trust, AML-based technologies must be integrated with other approaches and embedded in broader systems that are linked to sociocultural factors.

5 Exploring potential ethical and practical concerns about AML

The prevalent perception of AML is that it is "threatening," with negative connotations of intentionally deceiving other AI systems [82]. Still, some researchers have been exploring the potencies of AML to benefit society instead. Chen describes, in his presentation, how reversed designs and the concept of AML have been applied for positive purposes in recent years [82]. For instance, Sablayrolles et al. introduce the use of CNN algorithms to track data by embedding a tiny watermark—which does not affect the accuracy of the model—into images in a dataset to track whether the dataset has been used to train a model and to make the model trackable. They succeeded in tracking the dataset, even when only 1% of the watermarked dataset was used [83]. In another instance, Shan et al. propose a system for evading user-recognition devices by adding imperceptible alterations at the pixel level to a user's own images before they are published. Their experiments found that this cloak provided 95%+ protection against user-recognition services [79]. With this new trend, AML is no longer solely

an "adversary," as it can also be used to deceive other malicious AI systems for social benefits.

While the pursuit of using AML for social benefits is potential, we also realize that this possibility will be met with pervasive skepticism. Inquiries into the ethical and practical implementations of this approach are anticipated to be more in-depth and nuanced than those typically posed on "regular" developments in AI. Even with the present level of AI, some ask, "Who is accountable if something goes wrong?" and "How can I verify that AI systems are not harmful?" [84]. Thus, it is envisaged that these questions will be expanded for our exploration to employ AML to manipulate social media data. We categorize the potential questions raised by this possibility of using AML in our direction into four categories: ethical, utilization, social-media environment, and effectiveness concerns.

First, regarding ethical concerns. The use of AML to deceive other AI systems, even for beneficial purposes, may raise ethical concerns. To address ethical considerations surrounding the use of AML [85], provides valuable insights into this aspect by evaluating the potential collateral harms that may arise from the development of this technology to humans and non-humans. In the paper, they scrutinized the potential collateral harms, such as human safety in different scenarios, to assess the broader implications of AML-based interventions. In the context of this paper, one of the potential collateral harms that may arise is when unauthorized entities with malicious intent try to modify users' social media data using AML. Additionally, extensive access to users' data required for AML-based intervention may also raise harmful scenarios of unethical processing of this gathered data. Due to these potential issues, the developer of the mentioned technology should prioritize users' authorization and informed decision-making. By ensuring that users have control over and be informed about how their data is accessed and processed, developers can mitigate the harms regarding these issues.

Second, regarding utilization concerns. Uncertainties may arise regarding users' willingness to use AML to improve their privacy by altering social media data. They may feel constrained in their ability to use social media according to their own preferences if they permit a tool to alter their social media behavior, even if the goal is to improve their social media privacy. One way to increase their willingness to permit the use of these kinds of instruments is by giving them control over the extent to which they alter their social media data. As in a tool like Grammarly, enabling individuals to determine the degree of alterations fosters a sense of control and autonomy. By providing this control, developers can make users feel more in control of their online presence. Moreover, to prevent manipulation *by the tool itself*, users should be informed about the consequences of each option provided. In Sect. 3.1., we explained that manipulations occur when the influencers lack the care to help the target to understand the consequences of their action. Thus, to avoid becoming a manipulative tool, the tool that implements AML in our proposed direction should actively inform users of the repercussions of their decisions. For instance, users should be informed that they will be able to fool AI systems that want to profile their political preferences by 10% *but* confuse their human followers by 5% if they allow six alterations to their social media data, and that both percentages will double if they allow ten additional alterations. So that users' decisions are based on conscious consideration after being informed of prospective pros and cons.

Third, regarding social media environment concerns. The potential consequences of widespread adoption of AML may raise concerns about its impact on social media communities. As an illustration, consider the algorithms proposed by Makazhanov and Rafiei, wherein their profiling algorithms predict political preferences based on users' Twitter interactions with political party representatives [18]. In their algorithms, the more frequently a user's tweets contained words (keywords) from a ranked list of weighted party-specific topics, the more likely it was that the users supported the associated political party. If, for example, their algorithms correlated "renewable energy" and "technology" with Party A, then *John*, who frequently tweets about renewable energy and technology, would be identified as a supporter of Party A. Thus, if AML-based intervention suggests John stop tweeting about these keywords and instead start posting something about "mother," "family," and "peace" (the top keywords of opposing Party B), it may significantly change the user profile of John. However, if John started changing his social media behavior dramatically, there is a possibility that he would no longer be recognizable by his online friends who notice John based on his constant interest in renewable energy and technology. If many individuals begin altering their tweets with the intention of deceiving the AI profiler on Twitter, the platform will also become flooded with phrases that were not meant by the users themselves, and possibly even incoherent statements generated by the machine learning-based intervention. To address this concern, a human-centric approach to design and develop the tool is imperative. Some mechanisms for resolving the issue may also compete with other factors, such as technical evaluations. For example, instead of making substantial alterations that might raise a concern about identity manipulation, the designer of the mentioned AML-based intervention may consider a minor change in the form of a typo (for example, "transpurtation" instead of "transportation") [86]. Undoubtedly, making substantial alterations could be more effective to fool AI profiling systems than only suggesting minor typos. But, not only the essence of AML is

the implementation of small and barely detectable changes, balancing between technical efficacy, ethical consideration, and user values can mitigate the risk of unintended negative consequences of this technology that can negatively affect how humans interact with their worlds, while still achieving the intended objectives.

Fourth, regarding the effectiveness concerns. One may pose a concern about the efficacy of changing social media data in facilitating deliberative decision-making in elections. Even if people change their social media data to obscure their political profile, in the current digital era, it is indeed unrealistic to expect them to completely avoid manipulative political messaging while remaining online. Nevertheless, our proposal does not claim to eliminate exposure to such content. Rather, it seeks to diminish the degree to which AI-generated political profiles accurately infer the correct profiles of Internet users. For instance, *John* frequently posts on Twitter about his support for renewable energy, which makes him particularly susceptible to political messages related to this topic. Aware of this, Candidate A may target and bombard him with messages regarding this candidate's support for renewable energy, potentially influencing *John*'s election decisions. While these messages may indeed enhance John's knowledge and trigger him to deliberately consider voting for Candidate A in a reflective manner, it becomes problematic when he is intentionally left in the dark by Candidate A about the impacts of these messages in influencing his election decisions. As discussed in Sect. 3.1, educating John alone about these impacts is insufficient to ensure deliberative decisions, as AI-generated political influences can be subtle, massive, and less fully understood by most people. It is also unreasonable to expect John to remain vigilant against all influences all day. Therefore, our proposal centers on the idea that AML-based technology can harden political actors to exploit issues that John is particularly susceptible to by distorting the profiler AI's comprehension of his accurate profiles. John will continue to receive political propaganda that targets him. However, the propaganda will have a lesser impact on him due to its lack of relevance to his susceptibility, leaving him with more space to reflect before making his deliberative voting decisions.

However, as was previously mentioned in Sect. 5, individuals' election decisions are not solely influenced by their direct exposure to political messages but also by their interpersonal and social constructions. Thus, technological solutions that are based on AML will not serve as a panacea for achieving deliberative decision-making in elections. In this direction, the concept of AML should be integrated with other pertinent approaches, such as enhancing digital literacy and increasing the transparency of algorithms, which may be relevant to future research exploring this field.

6 Conclusion

The development of AI opens up vast opportunities to both do harm and to improve human life, raising the possibility of an end to the zero-sum game between humans and AI. In this paper, we have critically explored the possible objection of using AML—which has been previously considered only adversarial—to benefit society instead. Specifically, we investigated the possibility of using AML to alter the social media behaviors of users in order to deceive the AI systems that help external entities (such as politicians) generate powerful personalized influences impeding people's privacy and deliberative decision-making in elections. As people would have more time to think and reflect on their own political stance without being bombarded by manipulative influences tailored to exploit their individual susceptibilities learned from their social media engagement, we expect this approach to benefit both people's privacy and deliberative decision-making, while allowing people to continue to engage in social media and participating in today's democratic practices.

We provided some suggestions to address four categories of potential concerns that might arise in the implementation of AML in the aforementioned direction: ethical, utilization, social-media environment, and effectiveness concerns, underscoring the importance of responsible technology development. Recognizing that technology is not neutral, we advocate the integration of values throughout the design process during the development of the technology, such as by approaches like VSD.

Finally, we would like to emphasize that the approach suggested in this paper is not a permanent fix to the problem of "influencers" using social-media data to obtain information about individuals and exert personalized influences. For instance, the AI systems that *have* been fooled could change their algorithms after learning about the AML used in this regard, forcing it to relearn from scratch in a cat-and-mouse game. Nonetheless, the approach presented could be one potential remedy to the current state of affairs, in which humans have few countermeasures against AI's attempts that learn about them using their social media data.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43681-024-00588-2.

Funding The Indonesia Endowment Fund for Education (LPDP) (grant number: KEP581/LPDP/LPDP.3/2022) supported the first author of this research.

Data availability In this article, we do not analyze or generate any datasets, because our work proceeds within a theoretical and ethical approach.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Manheim, K., Kaplan, L.: Artificial intelligence: risks to privacy and democracy. **21**, 83 (2019)
2. Brkan, M.: Artificial intelligence and democracy: delphi. Interdiscip. Rev. Emerg. Technol. **2**, 66–71 (2019). https://doi.org/10.21552/delphi/2019/2/4
3. Helbing, D., Frey, B.S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R.V., Zwitter, A.: Will democracy survive big data and artificial intelligence? In: Helbing, D. (ed.) Towards Digital Enlightenment, pp. 73–98. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-319-90869-4_7
4. Ünver, H.A.: Artificial intelligence, authoritarianism and the future of political systems. **22** (2018)
5. Zittrain, J.: Engineering an Election. **127**, 7 (2014)
6. Rouvroy, A., Poullet, Y.: The right to informational self-determination and the value of self-development: reassessing the importance of privacy for democracy. In: Gutwirth, S., Poullet, Y., De Hert, P., de Terwangne, C., Nouwt, S. (eds.) Reinventing Data Protection? pp. 45–76. Springer Netherlands, Dordrecht (2009). https://doi.org/10.1007/978-1-4020-9498-9_2
7. Schwartz, P.M.: Privacy and democracy in cyberspace. SSRN Electron. J. (2000). https://doi.org/10.2139/ssrn.205449
8. Christiano, T., Bajaj, S.: Democracy. In: The Stanford Encyclopedia of Philosophy (2022)
9. Dahl, R.A.: On Democracy. Yale University Press (2020)
10. Elster, J. (ed.): Deliberative Democracy. Cambridge University Press, Cambridge, U.K.; New York (1998)
11. Cadwalladr, C., Graham-Harrison, E.: Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election
12. Rosenberg, M., Confessore, N., Cadwalladr, C.: How trump consultants exploited the Facebook data of millions. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html, Accessed 15 Nov 2022
13. Baran, J., Kajstura, M., Ziółkowski, M., Rajda, K.: Does Twitter know your political views? POLiTweets dataset and semi-automatic method for political leaning discovery (2022). http://arxiv.org/abs/2207.07586
14. Belcastro, L., Cantini, R., Marozzo, F., Talia, D., Trunfio, P.: Learning political polarization on social media using neural networks. IEEE Access **8**, 47177–47187 (2020). https://doi.org/10.1109/access.2020.2978950
15. Campanale, M., Caldarola, E.G.: Revealing political sentiment with Twitter: the case study of the 2016 Italian constitutional referendum. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 861–868. IEEE, Barcelona (2018). https://doi.org/10.1109/asonam.2018.8508243
16. Conover, M.D., Goncalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of Twitter users. In: 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing, pp. 192–199. IEEE, Boston, MA, USA (2011). https://doi.org/10.1109/passat/socialcom.2011.34
17. Kitchener, M., Anantharama, N., Angus, S.D., Raschky, P.A.: Predicting Political Ideology from Digital Footprints (2022). http://arxiv.org/abs/2206.00397
18. Makazhanov, A., Rafiei, D.: Predicting political preference of Twitter users. In: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE, Istanbul, Turkey (2012)
19. van de Poel, I., Kroes, P.: Can technology embody values? In: Kroes, P., Verbeek, P.-P. (eds.) The Moral Status of Technical Artefacts, pp. 103–124. Springer Netherlands, Dordrecht (2014). https://doi.org/10.1007/978-94-007-7914-3_7
20. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks (2014). http://arxiv.org/abs/1312.6199
21. Friedman, B., Hendry, D.G.: Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press (2019). https://doi.org/10.7551/mitpress/7585.001.0001
22. Verbeek, P.-P.: Cover story beyond interaction: a short introduction to mediation theory. Interactions **22**, 26–31 (2015). https://doi.org/10.1145/2751314
23. Pitt, J.C.: Guns don't kill, people kill; values in and/or around technologies. In: The Moral Status of Technical Artefacts. Springer Netherlands, Dordrecht (2014). https://doi.org/10.1007/978-94-007-7914-3
24. Kapp, E.: Elements of a Philosophy of Technology: On the Evolutionary History of Culture. University of Minnesota Press, Minneapolis (1877)
25. Verbeek, P.-P.: Ambient intelligence and persuasive technology: the blurring boundaries between human and technology. NanoEthics **3**, 231–242 (2009). https://doi.org/10.1007/s11569-009-0077-8
26. Collingridge, D.: The Social Control of Technology. St. Martin's, New York (1980)
27. Hutiri, W.T., Ding, A.Y.: Bias in automated speaker recognition. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 230–247. ACM, Seoul Republic of Korea (2022). https://doi.org/10.1145/3531146.3533089
28. de Reuver, M., van Wynsberghe, A., Janssen, M., van de Poel, I.: Digital platforms and responsible innovation: expanding value sensitive design to overcome ontological uncertainty. Ethics Inf. Technol. **22**, 257–267 (2020). https://doi.org/10.1007/s10676-020-09537-z
29. Kudina, O.: Ethics from within: google glass, the collingridge dilemma, and the mediated value of privacy. Sci. Technol. Hum. Values **44**, 291–314 (2019). https://doi.org/10.1177/0162243918793711
30. Friedman, B., Kahn, P.H., Borning, A.: Value sensitive design and information systems. In: Human-Computer Interaction in Management Information Systems: Foundations. pp. 348–372 (2006)
31. Ryfe, D.M.: Does deliberative democracy work? Annu. Rev. Polit Sci. **8**, 49–71 (2005). https://doi.org/10.1146/annurev.polisci.8.032904.154633

32. Chambers, S.: Deliberativedemocratic theory. Annu. Rev. Polit Sci. **6**, 307–326 (2003). https://doi.org/10.1146/annurev.polisci.6.121901.085538
33. Goodin, R.E.: 6 First talk, then vote. In: Innovating Democracy. pp. 108–124. Oxford University PressOxford (2008). https://doi.org/10.1093/acprof:oso/9780199547944.003.0006
34. Chambers, S.: Deliberation and mass democracy. In: Deliberative Systems: Deliberative Democracy at the Large Scale. Cambridge University Press, Cambridge (2012) https://doi.org/10.1017/cbo9781139178914.004
35. Manin, B.: On legitimacy and political deliberation. Polit. Theory **15**, 338–368 (1987). https://doi.org/10.1177/0090591787015003005
36. Thaler, R.H., Sunstein, C.R.: Nudge: Improving Decisions about Health, Wealth, and Happiness. Yale University Press, New Haven (2008)
37. Arno, A., Thomas, S.: The efficacy of nudge theory strategies in influencing adult dietary behaviour: a systematic review and meta-analysis. BMC Public Health **16**, 676 (2016). https://doi.org/10.1186/s12889-016-3272-x
38. Fong, G.T., Hammond, D., Hitchman, S.C.: The impact of pictures on the effectiveness of tobacco warnings. Bull. World Health Organ. **87**, 640–643 (2009) https://doi.org/10.2471/blt.09.069575
39. Rithalia, A., McDaid, C., Suekarran, S., Myers, L., Sowden, A.: Impact of presumed consent for organ donation on donation rates: a systematic review. BMJ **338**, a3162–a3162 (2009). https://doi.org/10.1136/bmj.a3162
40. Schmidt, A.T., Engelen, B.: The ethics of nudging: an overview. Philos. Compass. **15** (2020). https://doi.org/10.1111/phc3.12658
41. Klenk, M.: (Online) manipulation: sometimes hidden, always careless. Rev. Soc. Econ. **80**, 85–105 (2022). https://doi.org/10.1080/00346764.2021.1894350
42. Ballew, C.C., Todorov, A.: Predicting political elections from rapid and unreflective face judgments. Proc. Natl. Acad. Sci. **104**, 17948–17953 (2007). https://doi.org/10.1073/pnas.0705435104
43. Buchanan, L., O'Connell, A.: A brief history of decision making. Harv. Bus. Rev. **84**, 32–41 (2006)
44. Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C.R., Hertwig, R.: How behavioural sciences can promote truth, autonomy and democratic discourse online. Nat. Hum. Behav. **4**, 1102–1109 (2020). https://doi.org/10.1038/s41562-020-0889-7
45. Olivola, C.Y.: Elected in 100 milliseconds: appearance-based trait inferences and voting. J. Nonverbal Behav. **34**, 83–110 (2010). https://doi.org/10.1007/s10919-009-0082-1
46. Parsons, T.: On the concept of influence. Public Opin. Q. **27**, 37 (1963). https://doi.org/10.1086/267148
47. Spohr, D.: Fake news and ideological polarization: filter bubbles and selective exposure on social media. Bus. Inf. Rev. **34**, 150–160 (2017). https://doi.org/10.1177/0266382117722446
48. Todorov, A., Mandisodza, A.N., Goren, A., Hall, C.C.: Inferences of competence from faces predict election outcomes. Science. **308**, 1623–1626 (2005). https://doi.org/10.1126/science.1110589
49. Willis, J., Todorov, A.: First impressions: making up your mind after a 100-Ms exposure to a face. Psychol. Sci. **17**, 592–598 (2006). https://doi.org/10.1111/j.1467-9280.2006.01750.x
50. Christiano, T.: Algorithms, manipulation, and democracy. Can. J. Philos. **52**, 109–124 (2022)
51. Sunstein, C.: #Republic: Divided Democracy in the Age of Social Media. Princeton University Press (2018). https://doi.org/10.1515/9781400890521
52. Vaidhyanathan, S.: The politics machine. In: Antisocial Media, pp. 148–176. Oxford University Press (2022). https://doi.org/10.1093/oso/9780190056544.003.0007
53. Yeung, K.: 'Hypernudge': big data as a mode of regulation by design. Inf. Commun. Soc. **20**, 118–136 (2017). https://doi.org/10.1080/1369118x.2016.1186713
54. Lodge, M., Taber, C.S.: The rationalizing voter: unconscious thought in political information processing. SSRN Electron. J. (2007). https://doi.org/10.2139/ssrn.1077972
55. DeCew, J.W.: In Pursuit of Privacy: Law, Ethics, and the Rise of Technology. Cornell University Press (1997). https://doi.org/10.7591/9781501721243
56. Solove, D.J.: Conceptualizing privacy. Calif. LAW Rev. **90**, (2002)
57. Fried, C.: Privacy. Yale Law J. **77**, 475 (1968). https://doi.org/10.2307/794941
58. Westin, A.F.: Privacy And Freedom. Washington and Lee Law Review (1968)
59. Union, E.: Charter of Fundamental Rights of the European Union. (2012)
60. Allen, A.: Uneasy Access: Privacy for Women in a free Society. Rowman & Littlefield (1988)
61. Warren, S.: Louis, Brandeis: The right to privacy. Columbia Univ. Press. 1–21 (1989)
62. Moore, B.: Privacy: Studies in Social and Cultural History. Routledge, Abingdon (2018)
63. Bloustein, E.J.: Privacy as an aspect of human dignity: an answer to dean prosser. In: Schoeman, F.D. (ed.) Philosophical Dimensions of Privacy, pp. 156–202. Cambridge University Press (1984). https://doi.org/10.1017/CBO9780511625138.007
64. Wylie, C.: Cambridge Analytica whistleblower: we spent $1m harvesting millions of Facebook profiles (2018). https://www.youtube.com/watch?v=FXdYSQ6nu-M
65. Auliya, S., Nugroho, L.E., Setiawan, N.A.: A review on smartphone usage data for user identification and user profiling. Commun. Sci. Technol. **6**, 25–34 (2021). https://doi.org/10.21924/cst.6.1.2021.363
66. Brown, B.: Studying the internet experience. (2001)
67. Kokolakis, S.: Privacy attitudes and privacy behaviour: a review of current research on the privacy paradox phenomenon. Comput. Secur. **64**, 122–134 (2017). https://doi.org/10.1016/j.cose.2015.07.002
68. Norberg, P.A., Horne, D.R., Horne, D.A.: The Privacy paradox: personal information disclosure intentions versus behaviors. J. Consum. Aff. **41**, 100–126 (2007). https://doi.org/10.1111/j.1745-6606.2006.00070.x
69. Nissenbaum, H.: Privacy in Context: Technology, Policy, and Integrity of Social Life (2010)
70. Jiang, H., Pei, J., Yu, D., Yu, J., Gong, B., Cheng, X.: Applications of differential privacy in social network analysis: a survey. IEEE Trans. Knowl. Data Eng. 1–1 (2021). https://doi.org/10.1109/tkde.2021.3073062
71. Google/Ipsos: Privacy by Design: the Benefits of Putting People in Control (2022)
72. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings (2014)
73. Usynin, D., Ziller, A., Makowski, M., Braren, R., Rueckert, D., Glocker, B., Kaissis, G., Passerat-Palmbach, J.: Adversarial interference and its mitigations in privacy-preserving collaborative machine learning (2021). https://doi.org/10.1038/s42256-021-00390-3
74. Hathaliya, J.J., Tanwar, S., Sharma, P.: Adversarial learning techniques for security and privacy preservation: a comprehensive review. Secur. Priv. **5** (2022). https://doi.org/10.1002/spy2.209
75. Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I.P., Tygar, J.D.: Adversarial machine learning. In: Proc. 4th ACM Workshop Secur. Artif. Intell. (2011). https://doi.org/10.1145/2046684.2046692
76. Rosenberg, I., Shabtai, A., Elovici, Y., Rokach, L.: Adversarial machine learning attacks and defense methods in the cyber

security domain. ACM Comput. Surv. **54**, 1–36 (2022). https://doi.org/10.1145/3453158
77. Martins, N., Cruz, J.M., Cruz, T., Henriques Abreu, P.: Adversarial machine learning applied to intrusion and malware scenarios: a systematic review. IEEE Access. **8**, 35403–35419 (2020). https://doi.org/10.1109/access.2020.2974752
78. Alsmadi, I., Ahmad, K., Nazzal, M., Alam, F., Al-Fuqaha, A., Khreishah, A., Algosaibi, A.: Adversarial NLP for social network applications: attacks, defenses, and research directions. IEEE Trans. Comput. Soc. Syst. **10**, 3089–3108 (2023)
79. Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., Zhao, B.Y.: Fawkes: Protecting Privacy against Unauthorized Deep Learning Models (2020). http://arxiv.org/abs/2002.08327
80. Bail, C.: Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing. Princeton University Press, Princeton (2021)
81. Nguyen, C.T.: Echo chambers and epistemic bubbles. Episteme **17**, 141–161 (2020). https://doi.org/10.1017/epi.2018.32
82. Chen, P.-Y.: Adversarial Machine Learning for Good. In: AAAI Conference (2022)
83. Sablayrolles, A., Douze, M., Schmid, C., Jégou, H.: Radioactive data: tracing through training (2020). http://arxiv.org/abs/2002.00937
84. Umbrello, S., De Bellis, A.F.: A value-sensitive design approach to intelligent agents. Artif. Intell. Saf. Secur. (2021). https://doi.org/10.1201/9781351251389-26
85. Adomaitis, L., Oak, R.: Ethics of adversarial machine learning and data poisoning. Digit. Soc. **2**, 8 (2023). https://doi.org/10.1007/s44206-023-00039-1
86. Gao, J., Lanchantin, J., Soffa, M.L., Qi, Y.: Black-box generation of adversarial text sequences to evade deep learning classifiers. In: 2018 IEEE Security and Privacy Workshops (SPW), pp. 50–56. IEEE, San Francisco, CA (2018). https://doi.org/10.1109/spw.2018.00016