1

Corporate Surveillance and the Need for Transparency

The term *corporate surveillance* describes the ubiquitous monitoring of the behavior of individuals by corporations. Corporate surveillance has become commonplace today, and most internet users are subject to it every day. Corporate surveillance powers the dominant business models on today's Internet, but it also has the power to curb individuals' rights and create a fundamentally unequal surveillance society. This power, coupled with corporate secrecy surrounding their methods of corporate surveillance, motivates the need for more transparency.

This book is a guide to performing systematic experiments to create more transparency for corporate surveillance and its algorithms. This chapter begins with a high-level overview of corporate surveillance and how it has evolved over time, including the players in the corporate surveillance ecosystem, the reasons why corporations conduct surveillance, and the negative effects caused by it. Finally, the chapter will argue why these characteristics of corporate surveillance mean that there is an urgent need to subject it to greater transparency and audit the algorithms it relies on.

1.1 Evolution of Online Corporate Surveillance

To illustrate how online corporate surveillance evolved, we can follow the history of Google, which, according to Zuboff (2019), was the first to use data not just to improve their service but for monetization. In 1998, Google launched their first service, Google Search. In their drive to improve search results for users, Google used the *collateral data* that accumulated as a by-product of people using the search engine, including the "number and pattern of search terms, how a query is phrased, spelling, punctuation, dwell times, click patterns, and location" (Zuboff, 2019). Initially, this data was used solely to improve the search engine, for example, to improve the relevance of search

results shown to users. During this time, users and search engine relied on each other: the search engine needed users to improve its functionality, and users needed the search engine to find resources on the web.

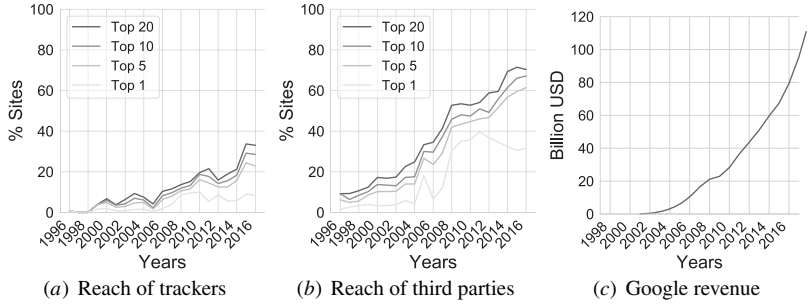
With the burst of the dot-com bubble in 2000, however, Google began to focus on advertising as a source of revenue. Initially, keyword-based ads were shown to users who searched for specific keywords. Later, Google's strategy shifted to targeting *individuals* instead of keywords. The collateral data that Google already collected for optimizing the search engine was used to create and analyze user profiles, which allowed Google to display ads related to the user's interests, not just their current search term. Google's 2003 patent for "Generating User Information for Use in Targeted Advertising" describes how data points that form a user profile, such as gender, age, and interests, can be inferred from collateral data. This approach allowed Google to optimize ad placement: instead of prioritizing the placement of ads with a high price per click, Google could prioritize ads that had a higher likelihood of a user actually clicking on them. This approach was then refined and expanded, for example, by tracking users on websites other than Google Search, which allowed Google to collect more profiling information, offer more precise targeting, and further optimize ad placement.

This increase of Google's trackers on non-Google websites was confirmed by longitudinal studies of tracking on the Internet. In 2005, trackers from `doubleclick.net` and `google-analytics.com` were embedded on 18% and 0% of first-party websites, respectively. In 2008, these numbers had already grown to more than 30%. Taking into account all companies acquired by Google, such as its acquisition of DoubleClick in 2007, Google's *reach*, i.e., the number of websites that embed Google's trackers, grew from under 10% in 2005 to almost 60% in 2008 (Krishnamurthy and Wills, 2009).

As a result of its advertising innovations and its increasing market dominance, Google's advertising revenue increased from $1.4 billion in 2003 to $134.8 billion in 2019 (Clement, 2020).

Even though Google is the most prevalent tracker on the Internet, it is not the only corporation to track users, and the reach of other trackers has also increased. For example, the ten third parties most commonly embedded in websites reached approximately 10% of websites in 1996–2000, 20% of websites in 2005, and over 60% in 2016 (Lerner et al., 2016). Figure 1.1 shows how the reach of confirmed trackers, the reach of embedded third parties, and Google's advertising revenue have evolved between 1996 and 2016.

To illustrate the extent of what these trackers know about users, three examples from a 2020 data leak from the tracking company BlueKai are illuminating (Whittaker, 2020):



(*a*) Reach of trackers (*b*) Reach of third parties (*c*) Google revenue

Figure 1.1 Three time series illustrating the evolution of corporate surveillance between 1997 and 2016. (*a*) The percentage of websites that are reached by the top trackers, (*b*) the percentage of websites reached by the top third parties, and (*c*) Google's advertising revenue. (*a,b*) adapted from Lerner et al. (2016), (*c*) adapted from Clement (2020).

- "One record detailed how a German man, whose name we're withholding, used a prepaid debit card to place a €10 bet on an esports betting site on April 19. The record also contained the man's address, phone number and email address."
- "The record detailed how one person, who lives in Istanbul, ordered $899 worth of furniture online from a homeware store. We know because the record contained all of these details, including the buyer's name, email address and the direct web address for the buyer's order."
- "A record detailing how one person unsubscribed from an email newsletter run by an electronics consumer, sent to his iCloud address. The record showed that the person may have been interested in a specific model of car dash-cam. We can even tell based on his user agent that his iPhone was out of date and needed a software update."

Even though these records are highly detailed and specific, it is important to keep in mind that BlueKai is not the largest tracker: it tracks only about 1% of all web traffic. Google and Facebook have greater reach, and the content of their databases can be expected to be at least as fine-grained and most likely encompass more individuals.

1.2 Ecosystem of Corporate Surveillance

The ecosystem of companies participating in corporate surveillance has steadily expanded during the last decade. In 2012, there were six main

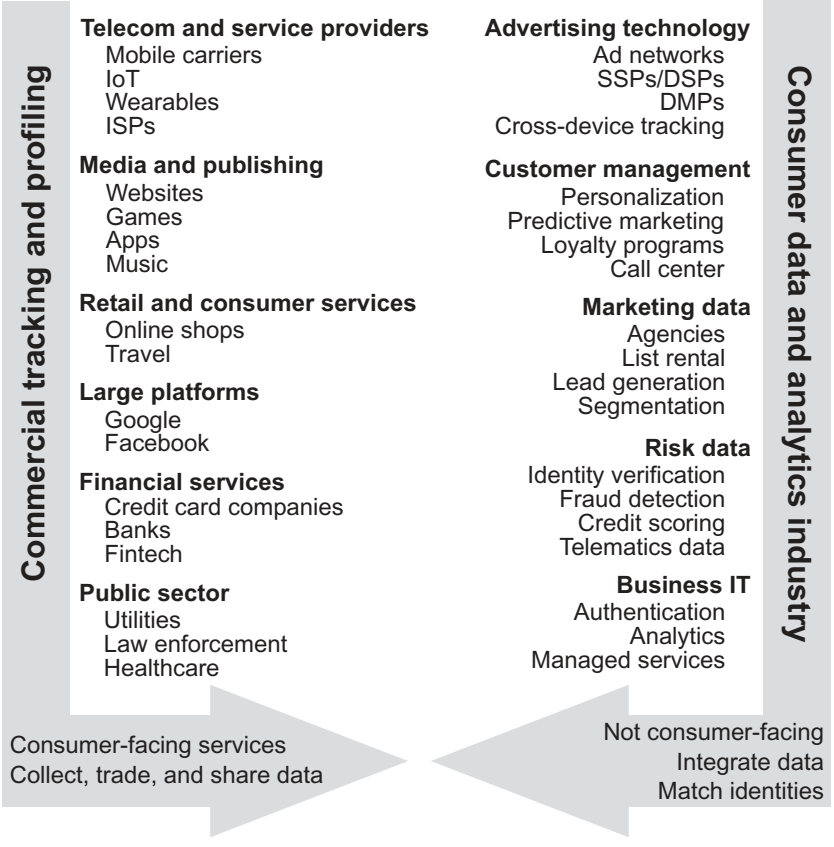


Figure 1.2 Tracking and profiling landscape. Adapted from Christl (2017).

business models for third parties on websites: advertising, analytics, social networks, content providers, front-end services, and hosting platforms (Mayer and Mitchell, 2012). Through their being embedded on websites, these third parties were generally in a position to track and profile users, although not all of them actually would do so.

In 2017, however, a large number of additional participants in the ecosystem were identified (Christl, 2017). These additional participants include both first-party sectors – sectors that offer consumer-facing services, shown on the left-hand side of Figure 1.2 – and third-party sectors – sectors that collect and analyze consumer data without offering consumer-facing services, shown on the right-hand side of Figure 1.2.

First parties that were traditionally associated with tracking and profiling were large platforms as well as media and publishing companies. However, the range of first parties has expanded so that now nearly every sector that has some online presence engages in tracking and profiling, including retail, internet and telecommunications providers, device manufacturers, financial services, and even the public sector. First-party participants in the corporate surveillance ecosystem typically collect data about individuals or allow third parties to do so; process and analyze the collected data for their own use; and share or sell the collected data, or access to it.

Third parties, in addition to advertising and analytics companies, now include companies that offer customer management, such as providers of call centers and loyalty programs; companies that provide marketing data; companies that offer risk management, such as background checks, screenings, or credit scores; and companies that offer business services such as authentication, identity verification, or fraud detection. Third-party participants in the corporate surveillance ecosystem focus on integrating data and matching user identities from multiple first parties. The collected data is also increasingly accessed for the purpose of government surveillance (Christl, 2017).

1.3 Motives for Corporate Surveillance

Various reasons have been brought forward for why companies engage in corporate surveillance, including economic motives, arguments from a business intelligence and optimization perspective, convenience, and the drive to achieve and maintain monopolies.

The economic motives for corporate surveillance rest on advertising-funded business models. These business models differ between the participants in the corporate surveillance ecosystem. Publishers can offer free web services if they have enough advertising revenue to support their operation (Couldry, 2016). Advertising revenues are higher if individuals with specific characteristics can be targeted. Advertisers can save money by targeting those individuals who will then go on to buy their product (Couldry, 2016). In addition, tracking allows advertisers to evaluate the effectiveness of their targeting (Yu et al., 2016).

The business model of large advertising platforms such as Google and Facebook consists of enabling – and profiting from – the publishers' and advertisers' business models. They collect comprehensive digital dossiers about individuals, including personally identifiable information, behaviors, and interests. Instead of selling these dossiers directly, which would give the buyer

the ability to make unlimited use of it after the sale, corporations sell the ability to make limited use of the data without transferring the data itself. For example, corporations may sell the ability to target an individual without revealing who the individual is. This gives corporations with large databases the ability to control how everyone else can use their data, giving rise to the claim that data is "the new oil" (Amnesty International, 2019).

Google and Facebook employ this business model by first offering useful digital services to consumers. The data collected from people who use these services can be mined to create detailed profiles about individuals that can be used to predict behaviors. The platforms then sell access to these profiles to other entities, such as advertisers, who want to target groups of people with specific attributes (Amnesty International, 2019). The valuation of Google and Facebook shows that this can be a very profitable business model, so much so that surveillance – not advertising – may be the true "business model of the internet" (Rashid, 2014).

To maintain and expand the viability of this business model, the corporations strive to collect ever more data to make ever more precise predictions about individuals. To do so, they intensify surveillance on products they already offer and branch out to new products and markets. For example, Facebook's Free Basics service attempts to dominate markets in the Global South, and Google's Nest and Assistant products attempt to dominate the smart home and smart city markets (Amnesty International, 2019). In addition, these corporations use their surveillance capabilities to identify potential competitors with the goal of either acquiring them as an additional source for data or forcing them out of the market by introducing a product with similar features. For example, Android, Maps, and YouTube were competitors acquired by Google; and Instagram was a competitor acquired by Facebook with the aim of crushing another competitor, Snapchat. In this way, big tech corporations are continually striving toward becoming monopolies and maintaining their monopoly status (Doctorow, 2020).

Business intelligence is another argument for corporate surveillance. Information about how users use an online service can be used to improve it, as Google did with its search engine (Couldry, 2016). This may also justify the use of analytics services by websites, because analytics services allow website owners to acquire this information without having to implement the data collection themselves (Yu et al., 2016).

Finally, some have argued that corporate surveillance can be convenient and even desirable for users. For example, social buttons and social logins are often seen as convenient because they can save time (Yu et al., 2016). In addition, many users are keen to optimize aspects of their own lives, for example

through health monitoring, the convenience of smart home applications, or the *quantified self* movement (Couldry, 2016).

1.4 Undesirable Effects of Corporate Surveillance

However, many authors have argued that corporate surveillance exhibits a range of undesirable effects, both for individuals and for society as a whole. For example, Zuboff (2019) wrote that "surveillance capitalism is a rogue force driven by novel economic imperatives that disregard social norms and nullify the elemental rights associated with individual autonomy that are essential to the very possibility of a democratic society."

Amnesty International (2019) frames these undesirable effects as *human rights risks* that can negatively affect the right to privacy, the right to freedom of expression, and the right to equality and nondiscrimination. In addition, the scale of platforms such as Google and Facebook can amplify these human rights risks because they allow manipulation of users at scale and feature addictive platform designs that aim to maximize the time users spend on the platform.

1.4.1 Right to Privacy

The right to privacy is often dismissed with the argument of having "nothing to hide." However, as Snowden (2019) memorably wrote, "saying you don't care about privacy because you have nothing to hide is like saying you don't care about free speech because you have nothing to say."

In fact, privacy is often argued to be a prerequisite to the enjoyment of other human rights, such as freedom of speech, freedom of thought, and freedom of assembly (Wachter, 2017; Rachovitsa, 2016; Buttarelli, 2017). This is because privacy is essential to develop a sense of our own identity and is therefore essential to human autonomy and self-determination (Amnesty International, 2019). Surveillance, however, curtails this right by creating a pressure to conform. In addition, corporate surveillance can define individuals' identities for them by creating profiles of individuals and using these profiles for targeting.

Big tech corporations have repeatedly made public promises to protect user privacy. However, they have also repeatedly failed to uphold their promises. Examples include Google's location tracking on Android, undisclosed microphones in Google's Nest devices, and Facebook's knowledge of the Cambridge Analytica data misuse (Amnesty International, 2019).

In addition, storage of large amounts of potentially sensitive data about individuals presents not just a privacy risk in the context of the company storing the data. For example, there are many known cases where privacy was breached by hackers or disgruntled employees, where data was sold after bankruptcy, and where data was subpoenaed by government agencies (Yu et al., 2016).

1.4.2 Right to Freedom of Expression

Freedom of expression includes not just freedom of speech but also the "freedom to seek, receive and impart information and ideas of all kinds" (United Nations, 1948). Corporate surveillance can threaten freedom of expression, for example by biasing search result rankings and by limiting the right to read anonymously.

Search result rankings are important because humans predominantly focus on the top of a search result list, and only a very small portion of search results beyond the first page are ever accessed (Hannak et al., 2013). Biases in these rankings, if they exist, would affect the ability of individuals to find information. For example, search results could be ordered in such a way that the top results always reinforce a user's existing worldview, thereby trapping users in filter bubbles. Even though existing studies suggest a limited extent of personalization (Hannak et al., 2013; Cozza et al., 2016), platforms are not transparent about their search result–ranking algorithms.

The right to read anonymously was affirmed by the US Supreme Court in 1962 because surveillance of reading habits is likely to create a chilling effect that makes individuals self-censor what they read. However, e-readers show how this right may be under threat from corporate surveillance. For example, Amazon reserves the right and claims the technical capability to record information regarding a user's interaction with Kindle content, including which books a user reads, which sections of a book have been read, what annotations were made, which geographical locations the e-reader was used at, how quickly a user is reading, and what the user's preferred complexity of reading material is (Wicker and Ghosh, 2020). However, Amazon does not confirm that it is indeed collecting these comprehensive records of Kindle users' reading habits, nor does it give users access to data collected about them.

1.4.3 Right to Equality and Nondiscrimination

The right to nondiscrimination can be violated by online advertising platforms when they allow the targeting of individuals based on protected characteristics.

In addition, discriminatory targeting can also be enabled by other targeting mechanisms and the ad delivery process.

Facebook's advertiser interface allowed advertisers to place ads that were targeted at protected characteristics such as age (Angwin et al., 2018), gender (Scheiber, 2019), and race (Angwin and Parris, 2016). This was possible even when the ads were for housing, finance, or jobs, where this discrimination is illegal according to US law. For example, Facebook allowed advertisers to target housing ads exclusively at white users, and ads for high-paying tech jobs exclusively at male users. Following lawsuits brought by the American Civil Liberties Union (ACLU), Facebook was forced to stop allowing discriminatory targeting on its platform (Scheiber and Isaac, 2019). As part of the settlement, Facebook agreed to create a separate advertising portal for housing, finance, and job ads, which would not offer the possibility to directly target protected characteristics. Advertisers would be mandated by Facebook's policy to use the new portal.

However, Facebook's other targeting mechanisms allow for discriminatory targeting even without direct use of protected characteristics (Speicher et al., 2018). For example, discrimination can be based on custom audiences, if the custom audience is constructed from audiences such as voter records for which the protected characteristics are known. Discrimination can also be based on attribute-based targeting, if attributes can be found that correlate well with a protected characteristic. Finally, discrimination can be based on look-alike targeting, if the advertiser knows identifiable information such as email addresses for a small group of people with the desired protected characteristic.

Discrimination can also result from the ad delivery process, where budget allocation and the use of stereotypical imagery can both lead to skewed ad delivery (Ali et al., 2019a). For example, the audience for ad campaigns with low budgets may consist of more than 55% men because women are more expensive to target. In addition, ads with stereotypically female images can reach audiences with more than 90% women. This skewed delivery affects all ads on Facebook, including ads for real-world employment and housing opportunities.

1.5 Need for Transparency

These undesirable effects emphasize the need for greater transparency of corporate surveillance systems and the algorithms they use. Big tech platforms themselves have recognized this need and provide a small set of transparency tools to their users.

Facebook, for example, gives users access to ad explanations that inform users why they are seeing a specific ad. However, these ad explanations are incomplete and misleading because they omit the most salient targeting attributes (Andreou et al., 2018).

In addition, Facebook provides a library of political ads to make it easier to analyze political advertising including how political ads are targeted. However, this library of political ads is incomplete, does not show microtargeting attributes, and requires advertisers to self-report their political ads (Silva et al., 2020). These shortcomings make Facebook's ad library unsuitable as a transparency tool.

Transparency tools provided by other platforms have similar shortcomings, which indicates that big tech is at best an unreliable source of meaningful insight into the characteristics and extent of its surveillance operations. However, transparency is essential for four reasons: to hold big tech accountable, to ensure a functional democracy, to ensure human autonomy and individuality, and to collect evidence for the need and effectiveness of laws and regulations.

Transparency to hold big tech accountable is needed to bring to light when big tech uses corporate practices that are unlawful, whether intentionally or not (Scheiber, 2019). For example, the case of Facebook allowing advertisers to place discriminatory ads would not have been uncovered without transparency research conducted by investigative journalists. However, the concentration of power in big tech platforms is an obstacle to accountability (Amnesty International, 2019), which further emphasizes the need for transparency.

Transparency to ensure a functional democracy is needed, for example, to uncover biases in big tech that can influence elections. For example, Hargreaves et al. (2018) found bias at the top of Facebook's news feed that favored some parties over others in an Italian election. In addition, Facebook has repeatedly experimented with displaying social messages to motivate users to vote on election days (Bond et al., 2012; Jones et al., 2017). Even though Facebook showed that their interventions on the Facebook news feed led to an increase in voters, they did not provide evidence that they used this power to influence voters in an equitable way. Transparency is needed to assure the public that Facebook did not selectively display their intervention to people they estimated to vote in a certain way.

Transparency to ensure human autonomy and individuality is needed not just because of the negative impacts of pervasive surveillance described above but also because big tech platforms, in particular Facebook, have admitted to performing experiments on their users (Kramer et al., 2014). For example, Facebook experimented with *emotional contagion* by manipulating the news feeds of Facebook users to display more posts with positive or negative