Carnegie Mellon University

REFRAMING PRIVACY IN THE DIGITAL AGE:

THE SHIFT TO DATA USE CONTROL

María Alejandra Arciniegas Gómez

Carnegie Mellon University

Summer 2023

Thesis Committee:

David Danks | Alex London | Lorrie Cranor | Alan Rubel

To the people who made this possible,

you know who you are…

From the bottom of my heart, thank you.

Table of Contents

INTRODUCTION

This dissertation is concerned with *privacy*: specifically, vulnerabilities to harm that stem from infringements on individuals' privacy that are unique to or exacerbated by the modern power of predictive algorithms in the contemporary digital landscape. For one ubiquitous example, consider how facial recognition technology has evolved over the past decade and fundamentally altered the sense in which our personal image is exposed, searchable, and manipulable in digital spaces. Modern algorithms are capable, based on relatively few data points (often in the form of photos freely uploaded to online albums), of identifying individuals in pictures in a variety of contexts—in different settings, from different angles, even at vastly different ages, sometimes including baby pictures! Relatedly, reverse image search is now quite an effective tool, in many cases allowing anyone with access to easily ascertain someone's identity from a single photograph. And of course, image manipulation has progressed by leaps and bounds in recent years, approaching a point where predictive algorithms can, for instance, generate false but eerily accurate portrayals of people in situations they may never have actually been in.

Our point of departure is a conceptual argument centered on the "blurring" of two previously distinct conceptual categories—observations vs. guesses—and the moral ramifications of this blurring. The arguments we make are intended not only to be responsive to the current state of technological development, but also to be forward-looking, taking into account the possible trajectories of these new technologies and how they may impact our lives as they become ever more accurate and widespread. For another small example: it is already the case that algorithmic tools exist to compile "personality profiles" of, say, job applicants based on their publicly available digital footprint—social media posts, images, "likes", purchase histories, etc. Is this an invasion of privacy? Is the answer to this question dependent on how accurate they are? It has, of course, always been the case that job applicants are assessed and appraised by their (human!)

interviewers; however, there are moral norms that divide fair appraisals from invasive ones. Even if this division is not entirely sharp, there are clear cases of transgression—for instance, few would support a potential employer's "right" to hire an investigator to follow an applicant around for weeks, or to access their private records (personal correspondence, medical history, diary entries, etc.). What happens as algorithmic predictive tools approach the point of functionally simulating such morally transgressive cases? This is just one example of the type of "blurring" that concerns us in this dissertation. Our analysis extends from here to explore and clarify the nature and scope of the corresponding moral ramifications, providing a framework for tackling questions of design, implementation, and policy that ought to govern these increasingly ubiquitous technologies.

In Chapter 1 we begin by setting the stage for the special problem of privacy in the digital age. As noted, we argue that this stems primarily from the modern power of predictive algorithms to "blur" the distinction between *guessing* and *observing*. At human-scale levels of reasoning, this distinction is sound and morally salient, typically marking the difference between (mostly) harmless (and inaccurate) speculation versus direct invasion of privacy. But what happens when machines are capable of making highly sophisticated guesses that can functionally (in terms of accuracy, detail, etc.) emulate direct observation? The moral relevance of this distinction *as a binary* is thereby upended, and consequently, the practical, social, and legal standards for what counts as "privacy" must be re-examined.

Privacy is a fundamental concept supporting individual autonomy and well-being along multiple dimensions; ignoring paradigm shifts in our notion of privacy therefore risks deep and widespread harms. Deepfake pornography, for instance, can cause severe emotional distress and even economic damage to its victims. And how should potential job applicants adapt to a landscape in which their every action may contribute to a "personality profile" that determines their suitability as an employee? Perhaps that social media post asking for therapist recommendations was a bad strategic decision, in this light.

Our analysis in Chapter 1 continues by challenging the suitability of "data control" approaches to safeguarding privacy. Briefly: this is the idea that the proper way to protect individuals from digital abuses of the sort described above is by requiring their explicit consent to release various

data—clicking "agree" to the privacy policies of your social media platforms, for example, and explicitly managing what aspects of your data you are willing to share and what you aren't. Essentially, we argue that the blurring described above more and more makes the prospect of this sort of protection-through-data-control impractical. The natural fix—and the line of attack that we pursue throughout the thesis—is a shift to constraining/regulating *uses* of data rather than its mere possession. This is a complex topic that we return to again and again; at this stage we lay out an important distinction between "question-answering" versus "action-guiding" senses of data use, discuss some of the challenges and complexities of regulating use in this context, and provide some concrete examples of policy and regulatory structures that might support such an approach.

In Chapter 2 we dig deeper into the "blurring" discussed in Chapter 1, specifically focusing on cases where the collision between guessing and observing is only "partial" or "intermediate"—we have many algorithms that are *better* at guessing than humans, but still fall short of functionally emulating direct observation. To what extent do the arguments and policy implications considered in Chapter 1 still apply in such intermediate collision cases? Naturally, we contend that our arguments are still highly applicable, though our reasoning must be refined. We begin by distinguishing a spectrum of intermediate cases based, loosely speaking, on different ways that an algorithm's "actual" accuracy may or may not match its "perceived" accuracy. Of course, these concepts are fleshed out in the chapter itself. These distinctions in turn help us frame the fundamental concepts of *informational harm* versus *presentational harm* that we introduce subsequently: that is, roughly, harms that stem from some factual information becoming known, versus harms that are rooted in being presented in some way (whether it is factual or not), respectively. This framework allows us to clearly articulate the privacy-related harms that come along with even "intermediate" cases, and thus broaden and generalize our argument to this much larger domain.

Finally, in Chapter 3 we return to the deep but crucial challenge of actually addressing the many privacy concerns we have raised previously. The concept of *consent* is key here since it often plays a morally transformative role in such cases—making the difference between intrusion and invitation. Unfortunately, "individual consent" is extremely tricky in this context, for several

overlapping reasons. Following a long tradition in bioethics, we contend that the value and transformative nature of consent relies on it being both *informed* and *uncoerced*, and there are myriad ways for these necessary conditions to fail in modern digital landscapes. We explore many aspects of these failures and conclude that highly individualized conceptions or practical implementations of consent are not truly viable here. This, combined with the fundamental shift to use control motivated in the previous chapters, leads us to an exploration of regulatory structures that are both socially distributed *and* focused on data uses rather than data ownership, shifting the burden of responsibility from the individual to social structures of governance, with which we conclude.

Privacy underpins our ability to function, grow, and thrive as human beings. Many of the major technological advances we are witnessing in our lifetime are or have the potential to drastically impact our most fundamental conceptions of what privacy is and our ability to preserve it. Of course, this does not mean that these advancements are "bad" in themselves—on the contrary, some may support unprecedented improvements to our quality of life. However, in order to reap such benefits, we cannot disregard the dangers. My thesis is that a shift to a "use control" paradigm for privacy protection and management, paired with regulatory structures that shift the burden from individuals to institutions, will be essential to adapting and preserving our core notions of privacy. The framework I argue for intersects many social and policy dimensions; my hope is that this work advances the discussion and helps lay some of the crucial conceptual groundwork for a thoughtful and effective implementation.

CHAPTER 1

Reframing Privacy: **When Guessing and Observing Collide**

Machine learning algorithms are essentially mathematical systems that can analyze large amounts of data to identify patterns. This makes them very good at predicting. For decades, these sorts of algorithms had two major limitations. The first was a scarcity of data, since algorithms like these need to be trained on appropriate data to make their predictions. Some of them require massive amounts of data training to be useful, and this data wasn't available. The second limitation was about the computing power of the machines: not only how much of those huge data sets could be processed, but also the computational complexity of the learning algorithms themselves. In the last decade, first "Big Data" and nowadays "AI" have emerged as blanket terms referring to the solution for both limitations. Data scarcity has been partly improved due to the increasing monitoring, surveillance, and data collection that is happening today, including identifiers on the web, mobiles devices, and in the real world.[1]. That is, pervasive monitoring and data collection has increased the *breadth of information gathering,* leading naturally to better predictions.[2] The second limitation has been partly circumvented thanks to the far-reaching effects of predictive analytics, which rely heavily on opaque but powerful computational tools like deep neural networks. Nowadays many known technologies rely on neural networks, such as driverless cars,

[1] Monitoring technology is founded in identifiers on the web, on mobile devices, and in the real world that are shared automatically with companies and linked with each other to form growing "shadow profiles" of users over time. Identifiers on the web include cookies, IP addresses, TLS states, local storage super cookies, browser fingerprints; identifiers on mobile devices refer to: phone numbers, hardware identifiers (IMSI and IMEI number), advertising IDs, MAC addresses; finally, some real-world identifiers include: license plates, face prints (face biometrics) and credit card numbers. These identifiers are used to create a tracking network that combines in-software tracking (in websites and apps) and passive, real-world tracking. Tracking in-software includes practices such as: analytics and tracking pixels, embedded media players tracking, social media widgets, CAPTCHAs, session replay services, and the massive ad networks (which involve activities like real-time bidding, where large amounts of information stored in cookies are shared with multiple third parties and data brokers, without the knowledge of the user). Passive, real-world tracking includes: WiFi hotspots and wireless beacons tracking, vehicle tracking and ALPRs, face recognition cameras, payment processors and financial technology (Cyphers and Gebhart 2019).

[2] The digital world has an incredible wealth of data, such as social media data, mobile data, data from the Internet of Things (IoT), cybersecurity data, business data, health data, etc.

chat bots, GPT-3 and ChatGPT, digital assistants (like Alexa), and many others.[3] This all affects the *depth of reach we get from the data that is collected*. By depth here, we mean that the whole is greater than the sum of the parts: in many cases, from a relatively trivial dataset, we can extract valuable information that wasn't evident by observing the data points by themselves. This, we will argue, fundamentally changes how we ought to think about privacy.

Section 1 | The Blurring Between Guesses and Observations

This new "depth" of data analysis has brought about a major *epistemic* change in our society, namely, a blurring of the distinction between "observing" versus merely "guessing".[4] Usually by observation we mean "observing as noticing and attending to interesting details of things perceived under more or less natural conditions" (Boyd and Bogen 2021), through any of our senses and measurement instruments, and it is usually understood to be good basis for knowledge. On the other side of the spectrum, we have mere guesses, that can very easily be wrong (like trying to guess a lottery ticket vs seeing the numbers that came out after the raffle). Of course, science moves in the realm of inferences, taking data points (observations) to make educated guesses and predictions. But technological development has blurred the line between both extremes in two distinct ways. I will refer to them as *Leap Cases* and *Intermediate Cases* (described below). This blurring brings to center stage a philosophical question: what is the difference between observing and guessing? As it happens, these two concepts have strikingly different moral implications. With AI algorithms rapidly blurring the distinction between the two, crucial questions of policymaking are going to depend on how we respond to this blurring, especially relating questions about privacy. I intend for my work to help frame and inform these decisions. In this chapter I will examine the nature of Leap Cases and their moral significance, formulating a conceptual argument regarding the practical implications of this blurring for how

[3] Neutral networks are generally used to deal with audio, video, text, and images. A non-comprehensive list can be found at https://en.wikipedia.org/wiki/Artificial_neural_network#Applications. ("Artificial Neural Network" 2023)

[4] By "guessing" here I mean to include both purely "random" guesses but also, more typically, *educated* guesses, what in some cases might naturally be referred to as "inferences" (though they may still, in theory, be false).

we ought to protect individuals from the vulnerability to harm that infringements to privacy entail.

1.1. LEAP CASES

One of the features of machine learning algorithms is that the better the models are trained, the more accurate predictions they can make, with fewer data points. This means that with the data that in the past would only allow us to have a rough guess or inference, nowadays an algorithm can give us a highly accurate prediction. An example of this are *deep neural networks*, which are becoming ever more standard use in many fields. A stark example is *facial recognition* technology, which allow us to identify an individual from a single photograph of them. Thus, information that before only allowed for a guess now allows for something much more accurate, akin to an observation, thus blurring the distinction between the two concepts of "guessing" and "observing". Concretely, we are referring here to cases in which (*i*) the inputs are what we usually associate with "guessing" (since they are sparse); but (*ii*) the outputs are what we usually associate with "observing" (since there is high accuracy). Therefore, we are presented with a *leap* where guesses effectively become observations, eliminating the distinction we previously had. This type of blurring will be the focus of this chapter, with the goal of examining the moral and policy implications for thinking about privacy in the context of such inferential leaps.

1.2. INTERMEDIATE CASES

Even though some algorithms are reaching an incredibly high accuracy as to basically count as observations, others produce results that fall into a grey area. In this sense, the blurring between guessing and observing represents moderately accurate prediction that are too accurate to count as a guess, but not necessarily as accurate as an observation. Nonetheless, they are sometimes still treated as "sufficiently" accurate (whether they are or not). This type of blurring will be considered in Chapter 2. The question here is how much these cases differ from the 'leap cases' in their moral implications. Can we apply the same framework of harms?

To the same degree? These questions might depend on how the algorithms are treated, independently of how accurate they *really* are.[5]

1.3. EXAMPLE OF A LEAP CASE IN NATURAL SCIENCES:

A useful example is the picture of a black hole that saw widespread circulation in 2019.[6] The image is not a true photograph; rather, it is a reconstruction based on data gathered over time by many different telescopes around the globe, in total capturing only a small part of the full image. An algorithm was created to analyze this very limited information and "fill in the gaps": to infer, from an infinite number of possibilities, what the true image of the black hole would most probably look like. Nonetheless, the astrophysics community presented this inference (some sophisticated "guesswork") as essentially equivalent to an actual observation, ascribing to it a very high accuracy. This image was received differently from all previous black hole simulations or artistic renderings. The difference? It was created based on real information using an algorithm that was designed to infer the reality, based on partial information. Thus, as a society, we are already accepting these predictive algorithms as carrying a degree of epistemic weight higher than "common" guesswork. Moreover, notice that without the current machine learning power, the information gathered would have not been sufficient to have more than a rough estimation or guess. The power of the new technology allowed to what in the past would only add up to a guess, to be considered a highly accurate representation.

This technology is of course not limited to the natural sciences. It is being widely deployed in the social sciences, from the private sphere, such as mega corporations like Amazon, Meta (Facebook), Apple or Alphabet (Google), and also any other small business or startup that offer products or services; to public entities, law enforcement and governments all around the world.

---

[5] One might argue that in virtue of being treated *as if* they are accurate observations, such intermediate cases carry many of the same moral implications as high accurate predictions.

[6] You can see the image here: https://www.nationalgeographic.com/science/2019/04/first-picture-black-hole-revealed-m87-event-horizon-telescope-astrophysics/ (Drake 2019)

Several questions arise related to the blurring, from the psychological (e.g., do humans interpret AI predictions as true facts/observations?), to the ethical (e.g., is it right to predict characteristics of individuals without their consent?), to the sociological (e.g., how do we deal as a society with the loss of privacy that highly accurate predictions entail?), to those questions pertaining to the philosophy of science (e.g., what counts as a "true" observation vs. a mere inference?). Here I focus on the ethical/moral questions. Moral questions arise inevitably when we are contemplating algorithms applied to individuals and social issues: algorithms that predict psychological traits, health, ability for a job, capability to pay a loan, capacity to complete a degree program, probability of recidivism after a crime, etc., intersect with individuals' fundamental rights to pursue good lives an avoid harm. From a moral perspective, we focus specifically on the *vulnerability to harms* that these algorithms can bring.

Consider these examples of cases were 'guesses' and 'observations' have different moral implications, different degree of vulnerability to harms. In these cases, the epistemic and moral distinction between them should be obvious. I'll outline the individual cases and intuitions first, and afterwards discuss broad commonalities and differences amongst them.

Nakedness: Anyone is free to imagine how someone else might look while naked. But this is very different from secretly peaking at (or taking a picture of) someone in the changing room without their consent. Imagining ("guessing") is an internal, private affair, whereas by contrast observing without consent constitutes an intrusion that we deem morally wrong.

ATM PIN: If someone were to try to figure out someone else's ATM PIN, they might simply try to guess the 4-number combination (perhaps using their birthday), or they could stand behind the unsuspecting person at the ATM and look over their shoulder. In the first case, we might assume that the chance of a person correctly guessing the numbers is too low to constitute any serious harm: after a few wrong attempts they will be locked out of the system. While being locked out of one's account may be inconvenient, it pales in comparison to the second case where the threat of harm is

very high and can include economic ruin. Though we might think of both approaches as reprehensible, the second one is clearly worse in this sense.

Job application: During an interview or job application, an applicant who is in a precarious financial situation may not want to disclose this to their prospective employer, so as not to lose their bargaining power. Or, similarly, a soon-to-be parent may not want to disclose their situation, fearing it may hurt their chances of getting hired. The employer might take a guess about the personal circumstances of their interviewee, but such private musings don't themselves seem morally culpable.[7] On the other hand, actually having access to (observing) personal information can seriously erode bargaining power or lead to discrimination.

Discussing a job candidate: During the discussion phase for hiring, say, in a university, the committee relies on their conversation not being part of the public record in order to freely speak their minds. If they knew a transcript of the conversation would be shared with the applicant, it might change what they say. The applicant can always guess what criticisms may have been discussed (and, arguably, they *should* try to make such guesses!), but obtaining a transcript feels more like a violation, and it generates chilling effects.

Copy of a test: Students finding and studying a copy of a test before it is released counts as cheating. In contrast, simply making educated guesses about what questions are likely to appear on the test feels more like being a diligent student.

In these examples, at a high level we see a pattern in which "actually observing" seems to be crossing a line that "just guessing" is not. They represent cases in which an *agent A* does not want a certain piece of *information I* to be known (their naked appearance, PIN, socio-economic/personal situation, criticisms, or test questions). In these cases,

[7] Though we might disapprove of an employer who bases their decisions on such musings; see below.

observing seems morally wrong, whereas guessing seems less so (and even praiseworthy in some cases, as in the case of a student trying to guess the questions on the test).

Of course, in these examples the harmful aspect of observation is predicated on it being against the will of the subject. Obviously, someone who *intends* to share a naked picture, or to tell someone their PIN, or send a copy of a test (for review maybe) is not being harmed by the corresponding observations. Therefore, we implicitly focus on cases in which the observations in question are *not* welcomed in this sense: i.e., where they occur without the person's consent.

It's also worth highlighting the fact that the moral status of "guessing" can range fairly widely, from being praiseworthy in some cases as noted, to being problematic in its own right (it might be already shady wanting to guess someone's PIN). For our purposes, however, the central point is that observing is significantly *worse* than guessing: it is more impactful, and more extreme in its consequences (i.e., it increases the vulnerability to harm for the individual affected). True, trying to randomly guess someone's PIN reveals bad intentions, can cause them minor inconveniences by locking their account, and may make them feel broadly unsafe. However, all of these harms persist and are compounded in the observational case. Similarly, an employer who tries to guess whether a prospective employee may soon get pregnant may be attempting to engage in a kind of discrimination, which is blameworthy in itself; but if in addition they have access to their medical records, their attempts to discriminate will actually succeed.

A big part of the differences in harms observed above comes from the fact that guesses will probably be wrong; for example, after three wrong PIN guesses the system will block further attempts. These cases also illustrate how observing can reduce or damage the autonomy of an agent in a way that guessing does not. When tracking the harms that can be identified in these and other cases, concepts such as individual integrity, autonomy, freedom, social relationships, control over information, chilling effects, etc., arise. These concepts are commonly tied to the moral analysis of privacy. (Solove 2005; Citron and Solove 2021). So, it seems our basic intuitions about this distinction closely track the ideas behind *privacy* as a moral right. This is expected,

since the concept of privacy, historically, revolves around the basic idea of differentiating a "private sphere" from a "public sphere", in which we can describe the loss of privacy as the individual having more aspects of themselves being *observed* by others.

Following this line of thought, the distinction between guessing and observing can be seen to track *infringements of privacy*, where observing crosses the line into a violation of privacy and the moral harm that comes with it. The transition into a moral harm is what we want to focus on when examining the leap from one end of the spectrum between the two concepts to the other, that is, where "guessing" becomes "observing". Specifically, the predictive capabilities of machine learning algorithms can reach the point where their "guesses" are in fact *highly accurate inferences*. So, it is worthwhile to go back to the examples and ask what happens when we replace "guessing" with algorithmic highly accurate inferences. This will allow us to more clearly identify some of the different dimensions of privacy concerns that are at play.

Nakedness: Someone's desire to maintain control over the image of their naked body can come from several sources, including a general sense of human dignity and self-determination, but also involving the idea of intimacy and consent. Intimacy is an essential component for healthy development as an individual and is crucial for developing a social and moral personality which requires the ability to maintain several degrees of intimate relationships with others. Only then can we develop interpersonal relationships involving love, trust and friendship, all of which are at the core of human experience.

Now imagine a situation where an algorithm can piece together, from publicly available (fully clothed) photos of you on the internet an extremely accurate "guess" about what you look like naked. Here, although no one has literally observed your

naked body, your dignity and control over intimacy is severely impacted. In other words, the harms of an accurate guess mirror those of an actual observation.[8]

ATM PIN: This case exemplifies a very concrete harm, namely, financial harm (losing one's money), as the result of losing control of one's information. We might imagine an algorithm that can make very good guesses about your PIN based on innocuous data (like how quickly you are able to enter different numbers or even words on a keypad). Clearly, even without the act of peaking over your shoulder, this leads to the very same tangible harms. As mentioned, just the intention to guess might be problematic, but the vulnerability to harms tangibly worsens as high accuracy inferences serve as observations. Adding to the previous problematic case caused by bad intentions (causing possibly loss of trust, feeling of un-safeness, etc.), we have now a very tangible financial harm.

Job application: As discussed, this case makes salient the problem of becoming vulnerable through the loss of bargaining power or being the subject of discrimination. Imagine if the employer could run a quick algorithm to scrape publicly available data from social media accounts and other public records, and receive a report including "Likelihood of pregnancy within 3 years", or a range of the probable total value of your assets, or your productivity score, or chances of you switching jobs in the near future. Any of this could severely undercut bargaining power and leave one vulnerable in the same way that direct observation did. Once again, we emphasize here that an

[8] This is connected to an area in which neural networks have already been used, in what we call *deepfakes*: a technology that can create synthetic media in which a person in an existing image or video is replaced with someone else's. Unsurprisingly, this technology has been mostly used in porn, to create videos and images of unwilling participants (mostly women). In this case the representation of the unwilling subject cannot be called accurate (it might be the subject's face on a random body). As above, this illustrates the fact that even "guesses" can be harmful—the mere fact of being associated with pornography might mortify some people, similar to a false rumor. But again, the point is not that guesses don't themselves cause harm, but that *observations or highly accurate guesses substantially compound those harms*. Relatedly, if the person in question happens to be an exhibitionist, and doesn't care about the disclosure of any nude representation of themselves (accurate or not), then this case this violates the established premise that the agent does not want or intends this information to be shared, and falls outside of the scope examined here.

employer who is *trying* to discriminate may already be blameworthy for this, regardless of whether they have the information to act on it. We are here concerned with the *additional* harms created by the availability of highly accurate guesses.

Discussing a job candidate: This case hints at the problems that generalized surveillance might create for how we behave, like the chilling effects that come from knowing that what we say will be observed, recorded, and disclosed. Surveillance, eavesdropping, tracking, or dissemination of information pose a threat to our self-determination and autonomy. Are these threats less if the "surveillance" is probabilistic, but with very high accuracy? If my private comments can be simulated based on a sophisticated "profile" of me built up from every scrap of publicly available information?[9] This may seem like science fiction, and thankfully at the moment it still is (though in some ways we are getting worryingly close to such a scenario), but we can (and must) still grapple with the potential moral implications, even if the technology is not there yet.

Copy of a test: A case like this shows how observing is equated to cheating, in environments when secrecy is important to maintain the integrity of the process. This specific case is interesting because here, the teacher is okay with the students correctly inferring the test questions, but *only* if that knowledge was arrived at through certain means (e.g., reasoning about the important topics covered in the lectures). By contrast, a student who simply runs an algorithm that collates thousands of previous test questions from similar courses and information about the specific teacher to arrive at an accurate "guess" of the upcoming test is clearly not doing something praiseworthy. The purpose of the course is to teach the students something or give them some skills. If they learn those things/skills and apply them to get a good grade, everything is working as desired. If they use some software instead, they are

[9] Here, even *somewhat* accurate predictions can carry harms similar to observations: e.g., an algorithm that predicts who on the committee was most critical/negative of the candidate.

circumventing the process to achieve a good grade without having learned the target skills.

1.4. THE IMPORTANCE OF PRIVACY

Privacy is a core value supporting the safeguarding of individuals' interests. A proper understanding and implementation of privacy allows us to strike an ethical balance between secrecy and the free flow of information. The protection of data can serve to prevent harm, allowing individuals the power to negotiate and establish fair conditions of engagement; it provides checks and balances, guarantees for redress and accountability, protects us from informational injustice, bias, and discrimination, prevents pernicious chilling effects caused by surveillance, and thereby allows us to act based on our own moral autonomy and human dignity. In a nutshell, privacy is the main line of defense we have at the present moment to protect individuals' interests, autonomy, and self-development, and in so doing, our wellbeing. If we are not careful to preserve privacy, we are left at the mercy of the interests and goals of those who own the flow of information, and as the last decade has shown, they will exploit that power, acting in ways that preserve and enhance the asymmetry of power and control in pursuit of their own interests.

Privacy is a complex and multi-faceted concept that plays an important role in social, economic, legal, technical and policy making perspectives. Though there is not a single, clear taxonomy in the landscape of the study of privacy, generally speaking defenders of privacy approach the subject by appealing to its relation to any of the following: control over information, human dignity, intimacy, self-concept, development of personhood, autonomy, social relationships, and restriction of access. It is useful to see how these cases bring out different dimensions of privacy. Understanding some aspects of the value of privacy and its different conceptions will clarify the moral importance of this matter.

Some theories of privacy claim its value is grounded in its connection with the protection of *human dignity* (individual integrity and freedom, autonomy, independence). Without privacy, a person is open to scrutiny and therefore vulnerable, which hinders their autonomy, freedom and

sense of self. On this view, privacy infringements such as surveillance, eavesdropping, tracking, or dissemination of information pose a threat precisely to the degree that they harm human dignity (Bloustein 1964).

Yet another focus highlights the close relation between privacy and *intimacy.* In short, without privacy, intimacy would not be possible (Fried 1970; Gerety 1977; Gerstein 1978; Cohen 2009). But intimacy is an essential component for healthy development as an individual. Developing a social and moral personality—an inner self or self-concept—requires the ability to maintain several degrees of intimacy with others. Only then can we develop interpersonal relationships involving love, trust and friendship, all of which are at the core of human experience.

The value of privacy can also be framed in terms of the importance of *social relationships* more generally. On this view, privacy is necessary to developing not only close relationships with others, but any sort of relationship (Rachels 1975). Privacy is required to protect one's or interests or assets, not to mention to save us from psychological (shame, embarrassment, etc.) and physical harm. Importantly, it allows us to navigate and perform the different roles we are required to take on in society, which call for different behaviors in different contexts, highlighting its connection to behavior and activities.

Lastly, privacy can be framed in terms of *accessibility by others,* that is, protection from unwanted *access* by others (Bok 2011). Here, privacy is achieved in three interrelated ways: anonymity (no one has information about me), solitude (no one has physical access to me) and lack of attention (no one pays attention to me) (Gavison 1980). This focus emphasizes the importance of concepts like secrecy, confidentiality, seclusion, and anonymity as forms of privacy.

It's worth noting that all of these different focuses of the value and importance of privacy overlap and complement one another, rather than competing. Many of these approaches go back once and again to ideas about the correct development of personhood, autonomy, human relations, and participation in society; ultimately, they share a common core of showing that privacy is essential for human flourishing and well-being. After all, privacy is essential for human

development and well-being as it is a core value that supports and safeguards the interests of an individual.

We can see that the array of harms caused by privacy infringements are varied. We have explored some of them in the scenarios above. Solove and Citron (2021) do a good job of laying out a taxonomy of privacy harms:

- Physical Harms
- Economic Harms
- Reputational Harms
- Emotional Harms
- Relationship Harms
- Chilling Effect Harms
- Discrimination Harms
- Thwarted Expectations Harms
- Control Harms
- Data Quality Harms
- Informed Choice Harms
- Vulnerability Harms
- Disturbance Harms
- Autonomy Harms

Privacy is required to protect ourselves from physical harms or economic harms (such as to protect one's interests or assets), but also to save us from psychological harms and emotional harms, such as shame, embarrassment, or chilling effects. It protects us from discrimination (that can produce physical, material, and psychological harms), and from disturbance from our goals and self-determination, it gives us control over our own lives. These examples demonstrate the various ways in which the observing versus guessing distinction tracks different types of privacy concerns and show that as accurate inferences get closer to observations, they tend to have the same morally problematic consequences.

Now, one might argue that even very good inferences are different from observations because of the "legitimate ways" they were obtained. For example, a person could argue that no matter how highly accurate an algorithmically generated naked picture is, generating it is not morally equivalent to *taking* the picture, since nothing ever actually crossed the boundary of literally seeing the person naked. But I want to argue that what really matters here from a moral perspective are the harms that come from such actions, and not the specific way the information was obtained. The crucial point is *vulnerability to harms*: predictive algorithms can function to make you equally vulnerable to the harms that come from direct observations, and therefore can be morally wrong.[10]

We have examined scenarios that exemplify how we are made more vulnerable to various harms by high accuracy inferences that, functionally, serve as observations. We have also noted that the field of harms seem to correspond to the harms associated more generally with privacy infringements. One might wonder then, what is the use of framing privacy concerns in term of observations vs guesses and the blurring, in this case *leap*, between the two that I have portrayed. In essence, the answer to that question lies in the solution. The usefulness of this framing is that it can highlight where our efforts should focus when trying to mitigate these harms.

[10] Now, what happens if the algorithm is wrong? I mentioned that the reason guesses are morally neutral (or in some cases at least less wrong) is tied to the fact that they tend to be off the mark. This could make one think that by the same reasoning, predictive algorithms are only a threat if they are factually correct and would be "morally neutral" if they happened to be wrong. This is not the case: incorrect predictions can cause a lot of harm if they are *treated* as correct. What I want to highlight is that *even if the predictions end up being wrong, their consequences can still be far-reaching*. Thus, the conversation around predictive analytics does not necessarily hinge on the predictions being correct, but rather on people *believing* that they're correct, or at least highly accurate (see Chapter 2 for more on this). In fact, in some cases it could be even more harmful when they are wrong; regardless, the moral challenges persist either way.

Section 2 | **Reframing Privacy: Why Data Control Approaches Won't Work**

Thus far I have made the case that there is a moral distinction between the concepts of guessing and observing: guessing (inferring) tends to be more morally permissible than observing without consent. As discussed, some cases of guessing can themselves be morally dubious, like trying to guess someone's ATM PIN, whereas in other cases the guesses might be morally neutral, like when trying to imagine what someone might look like naked. In both cases, though, the moral permissibility of the action degrades when it is an actual observation. I've argued this is because the vulnerability to harm is greatly increased with observations (e.g., money can be actually taken out of your bank account; someone can violate your wish for privacy by observing you naked against your will, etc.).

Now, we live in a world that is completely permeated by digital spaces. In a very tangible sense, our digital lives have become completely essential and intertwined with our "analog" lives. And in the digital space, monitoring and tracking is ubiquitous. Data collection is the basic standard under which all our interactions with the digital space work. In many cases this is for good reasons, while in others it is pushed to the extreme because of the pervasiveness of the data economy that has made overreaching data collection the default.

When trying to tackle violations of privacy in the digital space, a very common approach has been to focus on *data control*. This includes a variety of efforts (including the GDPR and other privacy policies) that focus on trying to decrease the amount of data that is being collected. The idea is that one's privacy will be protected to the extent that each individual controls more and more of the information about themselves that is shared with the world. My argument here is that, though laudable, focusing on data control will not work. Instead, we have to focus on regulating the *uses* of the data.

To understand this, let us first understand why data control[11] will not work. The argument for this is two pronged. On the one hand, the advancement of machine learning has made it so that fewer and fewer data points are needed to go from what previously would only be sufficient for a rough guess (inference), to nowadays a highly accurate inference (i.e., effectively an observation). The second prong of the argument is that it is the easy availability of data—not only the raw number of data points—that makes this approach untenable. The truth is that we are constantly hemorrhaging data throughout our daily lives, as there is no feasible way to limit our interaction with digital spaces; in other words, leaving a rich digital footprint is now unavoidable. This easy availability of information makes it completely impossible to exercise enough data control to really protect users from the consequences of violations to their privacy (in other words, the vulnerability to harms that their loss of privacy entails).

Two-pronged argument:

The general structure of my argument is as follow:

Historically, it was easy to know what information ***I*** could be predicted from data ***D***.

Hence, if I wanted to ensure that someone did not have access to ***I***, then I "just" had to prevent them having access to any data ***D*** that could be used to predict ***I*** (i.e., I just had to exert control over my data).

[11] An assumption behind the data control approach is that data ownership is what gives the holder of the data the freedom to use it in whatever way they see fit (i.e., to use it to train or develop any algorithm they—a company, usually—might be developing). Since the focus there is on legally owning data, the mitigations considered tend to focus on regulating who can sell data sets and to whom they can be sold. But once someone owns a dataset, they can use it to train or as input to any algorithm they desire. Some existing privacy policies such as CCPA are a mixture between data control and data use; though arguably CCPA mostly focuses on data control, there are some places where it restricts the uses that can be made of the data (for example, you can't use data to infer various kinds of PII (personal identifiable information).

These $D \rightarrow I$ relationships were easy to manage in the sense that one could control, with relative precision, the "prediction set I" by controlling what data to share/reveal.

However, advances in prediction algorithms have made it so that now it is very hard to predict which D might lead to which I, and moreover, in many cases,

Relatively sparse or innocuous data D can lead to very rich/accurate information I.

Such data D are readily accessible/available.

Hence, the strategy outlined in 1-2 no longer works.

So, we need to shift our thinking away from "control of data" (which is infeasible) and instead to something like "control of predictions" or "control of use".

The sophistication of modern algorithms is not only about "big data" or what large databases can be compiled; as algorithms get better trained, they are able to reach conclusions with fewer data points. As an example, as early as 2009, Acquisti and Gross' *Predicting Social Security numbers from public data* observed a correlation between individuals' SSNs and public information about them such as their place and date of birth, and this was enough to allow for reasonably accurate statistical inference of private SSNs. This was in 2009 and their algorithm wasn't even close to what we have today, as it merely relied on statistical correlations.[12] Even so, it was already showing a disturbing pattern which would only become more extreme: revelatory predictions based on comparatively few data points.

[12] "Algorithm Description: Our prediction algorithm exploits the observation that individuals with close birthdates and identical state of SSN assignment are likely to share similar SSNs. It employs the DMF as a public source of information about SSNs assigned overtime and across states. For each target individual, the algorithm proceeds by first predicting the target's ANGN, and then the SN associated with the predicted ANGN." (Acquisti and Gross 2009).

Nowadays, we have developed technology like facial recognition, which requires only one photograph to identify an individual. Some are trying more controversial approaches, like using facial recognition to identify sexual orientation[13] or "criminality". Even more recently, algorithms are being developed to recognize individuals based on their gait (the way they walk). Studies similar to these ones abound; machine learning techniques are claimed to outperform humans on a variety of tasks related to facial recognition, emotion detection, mental health assessment and so on. In line with the discussion above, even if these algorithms are not actually producing the results that people take them to be producing, the fact that people *treat* them as if they are often creates the very same vulnerabilities to harm (a company might decide to discriminate against LGBTQ people using an algorithm, even if, in reality, the algorithm is inaccurate. More on this in Chapter 2).

Notice here that in some cases highly accurate predictions are possible with only one data point when we have the *right type of data point*. So, take the example of trying to figure out if someone might be pregnant. The most accurate way to do it is to get a blood sample and test it. And we can agree that taking a blood sample without someone's permission and running a pregnancy test behind their back would constitute a violation of their privacy. By the same token, we also expect that blood samples are something we can reasonably have some expectation to remain private and not easily accessible to others (similar reasoning would apply to urine samples, saliva samples, and samples of other bodily fluids). This is where the second prong in our argument becomes important, since the information required from us to make accurate predictions is more

[13] Wang and Kosinski's *Deep neural networks are more accurate than humans at detecting sexual orientation from facial images* (2018). In it, they affirm that: "[w]e show that faces contain much more information about sexual orientation than can be perceived or interpreted by the human brain. Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 81% of cases, and in 71% of cases for women. Human judges achieved much lower accuracy: 61% for men and 54% for women. The accuracy of the algorithm increased to 91% and 83%, respectively, given five facial images per person (…)Additionally, given that companies and governments are increasingly using computer vision algorithms to detect people's intimate traits, our findings expose a threat to the privacy and safety of gay men and women." (Wang and Kosinski 2018)

and more easily available, ultimately making it completely impossible to control—recall the algorithms that require only one photo of an individual to work.

Consider the famous, if anecdotal, example of Target's pregnancy prediction algorithm. A company like Target, which has a vested interest in knowing when a possible customer might be pregnant. This is because research shows that first time parents are more likely to change their lifelong shopping behavior during that overwhelming period. To this end, Target instituted a practice of sending coupons for baby items to customers it guessed were pregnant, and started to develop a "pregnancy prediction score" powered by AI to improve the accuracy of these "guesses".

This case is described in Charles Duhigg's controversial[14] 2012 report of Target's customer data collection. The report describes an angry parent's complaint to Target about sending pregnancy coupons to her 15-year-old daughter, supposedly because it might influence her to get pregnant. But as it turned out, the daughter *was* already pregnant, and Target essentially knew about it before her family. How? The store assigns every customer a guest ID number linked to their credit card, name, or email address and collects their demographic information through various channels while tracking their purchases. Pregnant women tend to purchase different items during the gestation period, such as supplements like calcium, magnesium and zinc.

As the story shows, Target's algorithm was at least partially successful. But compare this case to one in which Target instead secretly acquired blood samples from their customers to assess pregnancy status. Most would agree that such a practice would be completely unethical. So, we see that Target is clearly using to their advantage the fact that the information that it used is in some sense "easily available" information, and they believe (or claim) that it being so somehow

[14] Since Duhigg's report, other journalists have cast doubt about the accuracy of this story (Piatetsky 2014; Fraser 2020). The main point of contention seems to be on whether the anecdote about the father complaining to Target really happened and whether the algorithm Andrew Pole was developing was really as accurate or precise as the media made it seem. What we can gather is that such algorithm was indeed created by Andrew Pole. He was the keynote speaker in the 2010 *Predictive Analytics World* conference, in which Pole describes a project to predict customer pregnancy. The question about its accuracy is fairly debated, but our argument presented here does not depend on the accuracy, but on the reaction it generated from people who started questioning the privacy violations it entailed.

morally transforms their actions into permissible ones. Nevertheless, the news about this pregnancy prediction score was not well received by the population: it was highly controversial and raised questions about privacy violations. In fact, Target was also not advertising it, and it was Duhigg's report that generated so much conversation around it. The attempt at secrecy suggests that Target's team knew that regardless of the data they used, they were infringing on their customers privacy (with the goal, of course, of furthering the company's interests—gaining more frequent customers—which translates to revenue for the firm).

These examples highlight how the nature and trajectory of predictive analytics must inform the practical actions we take to protect privacy. The issue is not merely privacy infringements that stem from data collection overreach (i.e., direct observations), but also those that derive from the algorithmic processing of "basic" information, i.e., information that is essentially impossible to contain. The idea of giving individuals better control over their information assumes that there is some core set of information that they can manage and control in order to protect their privacy and puts the onus of the responsibility on individuals to choose wisely how much to share. In other words, the underlying assumption is that sharing less will keep them safer. My argument here is that this intuition can no longer be upheld. It seems ridiculous, for example, to expect a person to never share a picture or video of themselves, in order to ensure that they are not "deep faked" in any way, or to ensure that their sexual orientation remains private, or, as per the previous example, to hide all their shopping habits to protect their pregnancy status (which would require not only never buying anything online, but also only using cash, and forgoing any discounts offered via a store card, coupon, etc.). The leap between guesses and observations is powerful precisely because it depends on having incredibly inaccessible information about the person; quite the contrary, it requires only highly available, impossible to contain information.

2.1. Public versus Private information

A question arises: Isn't readily accessible information effectively "public" information? And, if so, how can it be wrong to make predictions based on public information?

When thinking about privacy violations such as pregnancy predictions, the conversation often leads to the question of what data, what information, is "fair game", and what is not. As mentioned above, most people would consider something like blood samples to belong to a "private sphere" whereas other data, such as a person's buying habits, might belong to a "public sphere" and therefore be fair game for use (Rainie and Anderson 2014). In this manner, sometimes, the conversation around privacy boils down to what constitutes private vs. public data. Here, I argue that this is not the right way to frame the issue.

The idea that some data that is genuinely private and some is genuinely public used to make more sense before the technology revolution of the last century. After all, the understanding of privacy as *control over information* started with the attorneys Samuel Warren and Louis Brandeis in the late 1800s when they wrote their famous essay titled "The Right to Privacy" (Warren and Brandeis 1890), which opened the legal conversation about the right to privacy in the US. They refer to privacy as the right to be left alone, drawing a fundamental distinction between private vs. public.

Certainly, providing a concrete, positive definition of privacy is important in legal contexts; however, the narrowness of such a definition might result in it failing to fully capture what we actually want to protect. For a pertinent example: we might understand privacy as the right of a person not to have personal information about themselves made public (e.g., facts about their health, income, weight, sexual orientation, personal tastes, and interests, etc.); however, on a narrow view, this right would not apply to information that is already public. So, any information that already appears in newspapers, court records, or any other public document would count as part of the public record, meaning that further uses of it would not technically constitute a privacy infringement—including sharing it with a drastically larger audience. In the era of the internet and social networks, this means that information shared in a smaller platform, which may be assumed to be seen by few, can therefore be collected, stored, sold, and used by any number of entities, some of which the individual might not even know exist (e.g., broker companies whose main purpose is to acquire, package, and sell user datasets). This suggests that a broader, more nuanced understanding of the "public record" is needed, as well as what

constitutes "control" over our information (e.g., does it mean legal ownership? Understanding how it is being collected and its reach? Being able to delete it at any moment? These are complicated matters).[15]

Back then, what constituted the "public record" was more clearly defined which for a better understanding of the differentiation between public and private. With the advent of social media platforms, the distinction became less clear cut. Is what people share on social media public or private? What if it is shared to a very small audience? What if the social media account is set to private? Can third party companies buy and resell my data if I was sharing it only to my social media friends? What if my profile is set to public but I know (the social media metrics tells me) that my posts are regularly seen by only a dozen of people? And if we move away from social media into buying habits, banking habits, use of GPS, etc., it's even less clear what we can be reasonably understood by "public information". Are my buying purchases public or private? What about my movements around the city on a day-to-day basis? My "directions" searches on Google Maps? On the one hand, it seems reasonable to think that my minute-to-minute location is a private affair. On the other hand, any time I am out in public, no one needs any special permission to notice I am where I happen to be and record that information, mentally or otherwise.

Some authors (Selbst 2013; Solove 2002; Nissenbaum 2009), have argued that the notions of private vs. public are *context dependent,* and trying to differentiate what constitutes one versus the other is no longer a useful framing. Easily available information about us is everywhere and therefore it being "public" in some sense constitutes a weak excuse for it to be exploited in any way desired. Andrew Selbst (2013) argues against a conception of privacy that

[15] But the import of privacy may not hinge solely on a suitable articulation of what control over information means. Other theories of privacy focus more on how its value is grounded in its connection with the protection of *human dignity* (individual integrity and freedom, autonomy, independence). Without privacy, a person is open to scrutiny and therefore vulnerable, which hinders their autonomy, freedom and sense of self. On this view, privacy infringements such as surveillance, eavesdropping, tracking, or dissemination of information pose a threat precisely to the degree that they harm human dignity (Bloustein 1964) (see Section 1.4 on The Importance of Privacy above for more detail).

hinges on a binary distinction between public and private (which includes most if not all legal formulations). He states that "[s]ome definitions rely on whether information is secret or not, whether conduct occurs inside or outside, or whether the kind of conduct is in some general sense normatively private/intimate or not. The one similarity between all these definitions is the reason they all fall short: they are all binaries. Each of these binaries has proven inadequate, unable to capture society's definition of 'privacy'" (ibid, p. 647). He concludes: "The problem with binaries is that to employ them is to attempt the impossible—to simplify privacy by abstracting away the context."

Selbst draws from other authors who have steered in a similar way. Daniel Solove (2002) proposes that we conceptualize privacy from the bottom up rather than the top down, this is, from particular contexts rather than in the abstract. Helen Nissenbaum has focused on developing her *'theory of contextual integrity*', and suggests in her book *Privacy in Context*, that: "[A] right to privacy is neither a right to secrecy nor a right to control but a right to appropriate flow of personal information (...) Privacy may still be posited as an important human right or value worth protecting through law and other means, but what this amounts to is a right to contextual integrity and what this amounts to varies from context to context" (Nissenbaum 2009).

Hence, if we must abandon a binary between public and private, we cannot claim that the leap from guessing to observing is problematic only to the degree that it involves private information, and that as long as it relies only on public information, everything is fine. What matters is the resulting (highly accurate) prediction, whether it was based on many or very few data points. If we accept this, the idea that we can have control over our privacy by carefully choosing what to share and what not to share becomes increasingly infeasible. So, in sum I am arguing that this approach is essentially doomed and what we have to focus on is how we control (regulate) the *uses* of the information that is collected. If the argument has been persuasive thus far, and the need to pivot into uses becomes apparent, the next question that arises here is, what do we mean by "use" of a technology?

Section 3 | **Reframing Privacy – Regulating Uses Rather Than Data Possession**

In summary, the approach here proposed argues that, considering the leap from guesses (inferences) to observations and the consequences that leap has on our understanding of privacy, trying to control the flow of data is untenable. Our constant hemorrhaging of data is simply too ubiquitous, and the training of the algorithms too advanced (and becoming increasingly so every day) to make it feasible to have any meaningful control of our information. If we want to enact solutions that can tangibly protect us from the harms that losing our privacy makes us vulnerable to, we must focus on controlling the *uses* that the data collectors/processors can make with our data.

To be clear, this focus on the uses of information is not simply the result of an empirical claim about the insurmountable challenge to regain control over our information in digital spaces. It is true that over the past decades, individuals have left increasingly high data records of themselves online, and there are few mechanisms in place to control access to those records. Some might argue that creating and strengthening such mechanisms should be the primary goal. But the argument presented here suggests a different conclusion: we are proposing that the very nature of information-extraction algorithms has reached the point where individuals *cannot*, even in principle, effectively guard their information in a meaningful way. This is a strong claim that states that data control is effectively impossible. However, for anyone who has not been convinced with the arguments presented thus far, I want to note that even a weaker claim gives enough reasons to consider the shift to data-use control. This claim says that overall, we can expect more practical success (in terms of reducing vulnerability to harm to people) by focusing on controlling use as opposed to controlling access. Both cases lead to the focus on uses of information rather than mere possession of it. Nevertheless, controlling data use brings its own set of challenges and concerns, which I will address in future sections. But first, what do we mean when we talk about "uses" of data?

3.1. What do we mean by "uses"?

The flow of information begins with *data collection*. This happens when an individual interacts with any sort of space, digital or not, that is collecting the data in the interaction: the user engages in an *activity* that gets monitored and data is collected.[16] Subsequently the data is *processed* in some way. By this, I mean that the data is prepared (labelled, organized, etc.) and used either to train an algorithm or as input to the algorithm or (often) both. In turn, an algorithm has been created to answer a specific question. In this sense, data collection is "passive", but data processing is not. Here I argue that this is a first sense in which we are *using* the data: *to answer a question*. A second sense is when we *use* the *result* of an algorithm: say, for hiring a person, assessing parole, etc. This is what do we do with the answer to the question that was posed, or in other words, how it guides our subsequent actions. I'll argue that both these senses of "data use" have ethical consequences and should be better regulated; instead of focusing on data ownership, it is data use (in these two senses) that must be the focus of policy regulation.

3.1.1. Data Use 1 - Data Processing: the question-answering sense

No algorithm is neutral; they are intrinsically value-loaded. More precisely, an algorithm is always trying to answer a question in the sense that it was designed or is being used to solve a certain problem. For example, data about someone's gait may just sit in a database somewhere, but once someone applies an algorithm/model to it, it produces an answer to a question: "Is this person pregnant? Are they in a hurry? Are they gay? What's their name?" This is already a use of the data. Examples of other questions might be: "What is this person's credit score? Their identity, their personality traits, their emotional status, their mental health", etc. Ultimately, the outputs of this algorithm (namely, the answers to the questions) will also be used in some way, typically to inform some decision: "Are we giving you the loan?

[16] One can potentially interact with a digital space without data collection occurring (rare as it might be nowadays), and one can have data collected about themselves without engaging in a digital space.

Are we hiring you? Are we giving you a shopping discount? Etc." Such decisions correspond to the second sense of data use, the action-guiding sense of use explored in the next section.

The point here is that even before an algorithm is used in any decision-guiding sense, the mere selection of the question we want to answer already embeds values. This occurs in two main ways: first, the choice of the question corresponds to a goal, and goals entail values; second, the assessment of the algorithm's success at answering that question depends on how we choose to measure success, namely, which variables we measure and how we measure them—this depends on what proxies we choose for the values we care about, and thus is also closely related to values. These points are discussed further below.

3.1.1.1. GOALS ENTAIL VALUES

The idea that goals entail values, and that these values can then become embedded in the tools we create to achieve those goals, applies much more broadly than just to the case of algorithmic tools. To better understand and contextualize this idea that tools can have moral values embedded in them, we can begin with an illustrative example presented by Langdon Winner (1980) in his paper *Do Artifacts Have Politics?* Broadly, Winner argues that artifacts[17] can instantiate political properties, whereby "political" he means "arrangements of power and authority in human associations as well as the activities that take place within those arrangements". He states that artifacts "are instances in which the invention, design, or arrangement of a specific technical device or system becomes a way of settling an issue in a particular community" and there are also "cases of what can be called inherently political technologies, man-made systems that appear to require, or to be strongly compatible with, particular kinds of political relationships." (ibid, p. 123) In essence, artifacts of all sorts can "embody specific forms of power and authority", such as authoritarianism, social equity, freedom, and cultural pluralism (he mentions nuclear power as an example of the first and solar power as an example of the latter).

[17] For him, 'technology' is understood to mean "all of modern practical artifice" (p. 123).

His most famous example is the bridges over the parkways on Long Island, New York, built over the span of several decades, 1920s-1970s. He points out that many of the overpasses are oddly low. At first glance that in itself may seem to be innocuous, but the reason behind it is very telling. The designer, Robert Moses, built over 200 low-hanging overpasses in the area specifically so that buses would not be able to go under them. His reasoning was rooted in racial and social-class prejudice: if buses could not go through, this would disincentivize poor people—who more heavily rely on buses for transport around the city—from accessing the parkways and public parks in the zone beyond (though people have argued that the truth around this story is more complex)[18]. But let's assume that for Moses, the goal was to make some recreational parts of the city accessible only to the wealthier population, and not the poor. This example illustrates the values that can be embedded in the creation and deployment of artifacts; technology and algorithms are a part of that.

Next consider an example that involves algorithms and AI, the case of Job Ads administered by Facebook (run by Meta). A recent report[19] by Global Witness (Global Witness 2021) showed that Facebook seemed to be violating UK equality and data protection laws with their ad targeting. The ads people see on Facebook rely on the data collected about their users on and off the platform, which is in turn used to create profiles of each individual with the purpose of ensuring "that ads are delivered to the people who are most likely to click on

[18] This is another very well-known story that has been re-examined. The story originates from a biography book about Robert Moses titled *The Power Broker* written by Robert A. Caro in 1974. In the book, "Caro reveals that Moses ordered his engineers to build the bridges low over the parkway to keep buses from the city away from Jones Beach—buses presumably filled with the poor blacks and Puerto Ricans Moses despised." (Campanella 2017) The story was told to Caro by Sidney M. Shapiro, a close Moses associate and former chief engineer and general manager of the Long Island State Park Commission. Nevertheless, the story might be more complex. It is not in question that Moses held bigoted and extremely racist views, but he also constructed infrastructure such as pools and parks in poorer parts of New York. Moreover, the initial construction plans for construction included bus stops, and there are other external factors that contributed to why the bridges ended up the way they did. Now, though the real intentions of Moses are debated here, it doesn't change the argument that Winner is making about artifacts embedding values. As the example of Facebook personalized job ads shows below, even if not consciously intended by the creator, once we construct technologies, the values embedded in them have real consequences. These can be either shifting society towards a more accessible world for everyone, or ossifying and reinforcing negative biases that we currently have in our societies.

[19] https://www.globalwitness.org/en/campaigns/digital-threats/how-facebooks-ad-targeting-may-be-in-breach-of-uk-equality-and-data-protection-laws/?utm_source=hootsuite&utm_medium=twitter (Global Witness 2021)

them". An algorithm is used for this job; Global Witness reported that 96% of the people shown ads for mechanic jobs were men; 95% of those shown ads for nursing/nurse jobs were women; 75% of those shown ads for pilot jobs were men; 77% of those shown ads for psychologist jobs were women. The problem is that Facebook profits from optimizing clicks per ad, and this entails placing value on showing people whatever they are most likely to click; thus, in their profiling, their algorithm ends up replicating the biases that we have in society. But the problem is not only that it replicates the biases, but that in doing so, it reinforces them too: as we can see with this case, men and women will miss out on job offers, and be more likely to apply to the ones they were shown. This example shows that even if Facebook's goal is simply to maximize clicks, which may seem "neutral" in some sense, this goal leads to problematic behavior in which the algorithms embed and reinforce societal sexist values. Discrimination in personalized job ads is not new; it is quite pervasive,[20] and shows how social values are embedded into these technologies in a way that can perpetuate discriminatory practices if they are not taken into account when designing the technologies.

Based on Facebook's job ads example, one objection might arise. One could argue that embedding "bad" values in algorithms can be mostly unintentional, and, if so, is it the same as Moses' bridges case? Or, on the contrary, does intent matter, so these cases have different moral statuses? To answer these questions, it is useful to think in terms of intention, but also *foreseeability*. Was the unintended consequence foreseeable? We might distinguish three cases here:

[20] This is not the only case that shows discrimination depending on gender and age on Facebook. As the report notes, "Our evidence tallies with the findings of Algorithm Watch and academics who have also shown that Facebook's ad delivery algorithm is highly discriminatory in delivering job ads in France, Germany, Switzerland and the US. In fact, recent investigations in the US have shown that Facebook's ad delivery system is skewed by gender, potentially excluding swathes of women from seeing job opportunities even when they are equally as qualified as the men that are being selected by Facebook to see certain ads. (…) The New York Times and ProPublica have shown that big companies, including Facebook, have excluded older people from their job ads on Facebook in the US and The Sunday Times has done the same in the UK. Civil rights organizations sued in the US and in 2019, as part of the settlement of five of the lawsuits, Facebook prevented advertisers from targeting housing, employment and credit offers to people by age, gender or ZIP code - but only did so in the US and, later, Canada." (Global Witness 2021)

1. Cases where the bad consequences were intended (such as Moses' bridges)

2. Cases where the bad consequences were unintended but foreseeable (often true of algorithmic bias nowadays)

3. Cases where the bad consequences were unintended and not foreseeable (true accidents)

Case (1) as we've seen is straightforwardly considered morally reprehensible. Most people would agree that (1) and (2) are worse than (3). It is possible to argue that Facebook didn't intend to replicate societal biases, and that might be correct. But the question here is if the consequences of their algorithm were foreseeable, which they were (any basic ethical analysis of the impact of this tech would easily reveal them). This means the lack of intent is insufficient to eliminate all blame, since the negative effects were clearly foreseeable. Wishful ignorance is not enough to shield us from moral responsibility. At the other end of the spectrum, there might also be cases in which truly the best intentions lead to unforeseeable bad consequences. In these scenarios the best we can do is learn from the mistake and work on solving it. We might conceive of case (1) as instances of Bad Intent, case (2) as instances of Culpable Negligence and case (3) as mere accidents.

Now it might be tempting to assume that the crux of the moral issue here rests on whether the negative consequences were truly unforeseeable, but this might lead to endless debates, and should not be the focal point. The most important takeaway is that, once the dangers of certain practices are discovered, there is a moral obligation to mitigate or correct the harms. The choice to keep using a technology that has been shown to be morally flawed once the dangers are discovered, constitutes an intention; the continued act of using it is a politic act—an algorithm might initially fall into case (2) or (3), but if a company continues using it even after the harmful effects are revealed, then it arguably moves into a case (1).

3.1.1.2. MEASURES AND PROXIES

It is worth briefly considering another, related value laden challenge, that is present whenever we attempt to measure some quality, which is exactly what we are trying to do with to algorithms: the *problem of translation*. Passi and Barocas (2019) talk about the complexities involved in the process of formulating a problem in data science and its impact on fairness. "It requires various forms of discretionary work to translate high-level objectives or strategic goals into tractable problems, necessitating, among other things, the identification of appropriate target variables and proxies" (ibid, p.29) The translation problem is relevant because when developing an algorithm, we are trying to answer a high-level question, but often the answer to that question must be inferred from available variables that only serve as proxies for the values we truly wish to capture; in particular, we have to figure out what all the relevant proxies are. If we are talking about finding a "good employee" what does "good" mean? What are the variables that will be included in the model to capture the "good"? Choosing proxies can already carry moral weight, as the chosen variables can have disparate impacts on different populations/individuals. For instance, will we consider their income? Their credit scores? Their zip codes? Their education level? The number of friends they have? Their shopping habits? Why or why not? Or how about using their personality traits? If we think a good employee is a charismatic employee, how do we measure it? In essence, how do we decide which variables are relevant to a given question and how do we measure them? Many algorithms fail because they are not properly capturing what the programmers hoped to be capturing and therefore the question at hand ("Is this a good prospective employee?") is not being really answered.

An example of this, given by Passi and Barocas (2019), is using "arrests" as proxy for crime when using machine learning for policing and criminal justice. The problem here is that the chosen target variable is not accurate, it is racially biased since more black people are arrested for the same crimes. Therefore, relying on the model will lead to "misleading reports about the model's real-world performance: these metrics would reflect the model's ability to predict the label, not the true outcome. Indeed, when the training and test data have been

mislabeled in the same way, there is simply no way to know when the model is making mistakes." They conclude "[c]hoosing a target variable is therefore often a choice between outcomes of interest that are labeled more or less accurately. When these outcomes are systematically mismeasured by race, gender, or some other protected characteristic, a model designed to predict them will invariably exhibit a discriminatory bias that does not show up in performance metrics" (ibid, p. 2)

For another example, consider healthcare and risk scores. Obermeyer et. al (2019) examine one class of commercial risk-prediction tools that is estimated to be applied to roughly 200 million people in the United States each year. They note that when developing a risk-score algorithm for illness, the proxy chosen was health care costs, assuming the sickest people require more expensive care. The problem, notably, is that "unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise" (ibid, p.1). This is incredibly problematic since healthcare systems rely on these commercial prediction algorithms to identify and allocate help and resources to their patients that need them the most.

To conclude this section, I want to clarify that, of course, not every design aspect of an algorithm is necessarily ethically relevant in every case. For example, the choice between coding in Python versus R might not have any practical impact on anyone's life (aside from the programmers!) in most contexts. Similarly, the choice between, for instance, using deep neural networks versus a linear model may be innocuous in some cases and not in others—in particular, *interpretability* or *transparency* will vary based on the model. The algorithm might be a black box, or it might be interpretable. How do we decide what's best? Again, this shows that the intention behind the algorithm matters, i.e., what question is it trying to answer? If we are developing a credit score, we might want to understand how the algorithm made the decision and an opaque algorithm might be problematic. But if we are only trying to find a way for a self-driving car to stay in the lane, it might not be relevant to understand

how it's doing so, as long as it works. Similarly, when seeking accountability, the transparency of the model might be important.

How can we go about this? Authors like Cynthia Rudin (2019) explore the use of black-box models in machine learning. Her view is that when algorithms are used in *high-stakes decisions* where the consequences can be severe (such as problems in healthcare, criminal justice or finance)[21], the interpretability of the model is of utmost importance, both for understanding the prediction and for accountability. She argues against the idea that there's a trade-off between accuracy and interpretability, and presents in detail the problems of trying to build what she calls "explainable AI" (namely, additional algorithms designed to explicate black-box algorithms), instead of building interpretable algorithms from the get-go. Contrasting "high-stakes" versus "low-stakes" decision making seems to be a good starting point for evaluating whether the technical modelling choices of the algorithm itself matter, ethically speaking. This connects with the second sense of usage that I will explain below. As Rudin's stance shows, ethical reflection about modelling choices (pertaining to the question-answering sense of use) is intrinsically connected to and depends greatly on the action-guiding sense of use. For now, the important thing to notice is that the whole creation of the algorithm is trying to answer a question and the question already implies an intended use of the algorithm: it is not neutral.

3.1.2. Data Usage 2 – Application: the action-guiding (action-informing) sense

The second sense of data usage relates to the use you make of the answer to your question. In other words, the algorithm, once it gives its output, is going to be applied in society in a certain way. For instance, suppose you now know your employee is pregnant—do you fire them? Pass them over for a raise? Send them ads for diapers? And so on. Similarly, once we have a credit score, are we using it to deny access to loans or for providing credit counseling?

[21] Specifically, she writes: "there have been cases of people incorrectly denied parole, poor bail decisions leading to the release of dangerous criminals, ML-based pollution models stating that highly polluted air was safe to breathe and generally poor use of limited valuable resources in criminal justice, medicine, energy reliability, finance and in other domains." (Rudin 2019, p. 206)

Or if we design an algorithm that diagnoses propensity to a specific ailment—are we using it for preventive care or to raise insurance premiums? So, to be clear, the action-informing sense of the use of an algorithm consists in the role it plays in a decision-making process i.e., the choice of when to apply it, and the role that the outputs of that algorithm play in guiding a decision. This is, we define "use" in the action-informing sense in terms of the functional role of the data/algorithm plays in decision making, the impact that it has on actions.

Crucially, the same algorithm can have many disparate applications; indeed, the exact same code can be used in different ways. Code alone does not determine use. Take facial recognition technology, which can be deployed to track the movements of individuals in a city. It might be used to discover missing people, but it can also be used to identify protesters. It might be used to find criminals, but also to track political activists. It is evident in these cases that the different uses of the algorithm may have different moral statuses.[22] Thus, when we focus on regulating "data use", the complexity and nuances in the differences between the various possible action-guiding uses means that designing effective regulation around them will be an endeavor that must, to some extent, be carried out on a case-by-case basis.

Take the case of AI in the medical sphere. We have seen examples in which an algorithm that can diagnose a specific ailment, say diabetic retinopathy, may have a clearly prosocial use (in that the use of it aims to better assist the patients), but also a negative (or at least not clearly pro-social) use (e.g., the algorithm could be used to deny insurance coverage for pre-existing conditions). How then do we answer the question of whether we "should" develop the algorithmic diabetic retinopathy diagnostic tool?

This is the fundamental question: when "should" we regulate the development or deployment of a given algorithm? Of course, the first question is whether we can, in fact,

[22] As mentioned, even parts the same code can be repurposed in the design of other algorithms: facial recognition tools can become a part of algorithms used to predict sexual orientation or criminality, or to identify individuals by their gait, or by their irises, etc.

develop an algorithm that accomplishes a given goal, that is, one that actually works.[23] This question of "success" or accuracy corresponds to the question-answering sense discussed above. Let's assume that it does, as the paper by Abramoff et al. (2018) seems to show in the case of diagnosing diabetic retinopathy. In this case, the core question arises: *should* we develop such an algorithm? So, we see that to answer this question we need to know more than just what question the algorithm is answering. In regulating its use, we have to ponder, who benefits? The patient or the insurance companies? Maybe both? Examining the stakeholders and the moral values involved we can regulate the use, for example, by accepting its use for patient treatment but not as a tool for assessing insurance costs or access. If the intended (morally acceptable) use of an algorithm is concrete and specific enough, we can have some hope of crafting effective regulation for it. But as we will see, part of the problem is that the action-guiding uses of an algorithm are not always clear cut, or even known in advance to those who develop it.

The diabetic retinopathy diagnosis algorithm might be an example of a technology we can more easily regulate for good uses. But one might wonder whether some algorithms, by the very nature of the questions they are answering, are so predisposed to ethically questionable action-guiding uses that we should heavily restrict or outright ban their development in the first place.

In this line of thought, we might conceive cases of clearly immoral or socially destructive uses of algorithms. For instance, as noted, some people have tried to predict (a la "Minority Report") probable criminality based on facial structure—a usage which seems like this century's version of phrenology. It is similarly very hard to imagine good uses of algorithms that aim to predict sexual orientation, since the most obvious applications involve exploitation by discriminatory governments or corporations. Of course, in these cases, we might suspect that the algorithms are simply failing at what they promise to do, that is, they are taking the

[23] Take for example the cases of algorithms meant to detect Covid-19 using chest radiographs and CT scans which have failed (See the paper by Roberts et al. (2021) as an example and analysis of why these algorithms have not been properly designed nor trained).

wrong proxies, or drawing the wrong conclusions. But let's bracket this issue and assume for the moment that they did successfully answer the question posed: they accurately predict sexual orientation or criminality, for example. In these cases, the action-guiding (action-informing) uses are so problematic that the proper policy response might be to restrict the development of those algorithms altogether. Indeed, any algorithm whose principle use in society is to violate human rights falls into this category. In other words, when examining a new technology, we must evaluate their power to oppress. This has to be contextual to the rules of the current system.

Naturally, many technologies will be harder to evaluate because they have several possible uses. Facial recognition technology (FRT) is perhaps a particularly hard case, and therefore worth exploring. In what follows, I will emphasize three different points:

I. The wide variety of potential uses of the same question-answering algorithm is another reason why focusing on "data control" is futile (specifically: the issue is not whether someone *has* my facial markers but rather what they will do with them).

II. The ethical status of FRT depends on its action-guiding role, not only its question-answering role. [24]

III. However, if control based on action-informing usage is not feasible, we need to consider whether we should ban this entire *type* of question-answering technology (or relatedly, whether we should change the exact question that the tech answers).

So, should we develop FRT technology? As discussed, its uses are far reaching and varied. From identifying previous shoplifters or criminals, to finding missing people, to tracking activists or social leaders. Authoritarian regimes can exert an incredible control over their

[24] At the very least, we ought not use FRT if it is bad at its question-answering role.

citizens with it, as can be seen in China. Countries like the US are not that far behind. The Clearview scandal in 2019 revealed how many law enforcement agencies around the country are using the private company Clearview's technology to assist in their searches (Hill 2020; 2022; Devany 2022). We are presented here with a technology that has both good and bad uses. How do we reason about this?

The first point to notice is that FRT is an exemplar case in which data control seems completely futile. For one thing, the individual is completely powerless to exert any meaningful control over their privacy because, as explained in my analysis of the blurring between guesses (inferences) and observations, there is a leap case here in which just one image of oneself can open the door to highly accurate predictions. Anonymity, for example, becomes impossible when just one picture can be traced back to one's identity (and the digital footprint associated with it). Moreover, since as noted there are both good and bad uses of FRT, even if we could control access to the raw data, it wouldn't give us the kind of fine-tuned control over the uses that we might wish for.

This leads to the second point: that the ethical status of FRT or any other technology depends fundamentally on its action-guiding use. There are seemingly neutral uses of FRT, such as unlocking your phone, which focus on convenience; there are prosocial uses of FRT, such as helping find missing people or highly dangerous criminals. But ultimately, by its very nature, FRT is meant to identify individuals and track them, and this means that conceptually, it is completely intertwined with surveillance. Making such tools broadly available will therefore contribute to a surveillance-based society, either by governments or corporations or both.

The negative impacts of surveillance are numerous and have been heavily discussed in the literature. Perhaps the clearest one is the chilling effect it has, but as I have argued, losing privacy enhances vulnerability to harm along several dimensions (financial harm, physical harm, psychological harm, etc.). Morally speaking, then, the action-guiding uses of FRT are very worrisome and perhaps inevitable. So, what do we do?

This brings me to my third point. I have argued that we have to focus our policy making and regulations on limiting the allowed (ethically desirable/permissible) uses of a technology. But this is not always an easy task. Ideally, we would want to prohibit the problematic uses of FRT but allow the good ones, that is, allow for certain specific uses of FRT while heavily restricting others. The problem here is that once the technology exists in some form (say, geared towards finding missing people), it is very hard to stop it from being used in accordance with what it is conceptually linked to, i.e., surveillance of all sorts. There is an old saying: "when all you have is a hammer, everything looks like a nail". Here I am making a similar point: "when you have a hammer, you will eventually realize that it is good for pounding in nails". In other words, the widespread availability of a tool that is efficient at X all but guarantees its eventual use for X, whatever the intentions/restrictions on its original deployment. When a technology is ripe for abuse in a vast array of contexts, proper regulation can become incredibly hard to enforce. Some might find this too pessimistic. At the very least, we need to examine the factors that are relevant to the existence of a technology leading to its misuse, including the incentives for misuse and questions such as "how easy is it to detect and stop the abusive uses?" and "how easy is it to ensure that only 'good actors' can actually access/use the technology?". In cases where abuse is highly incentivized and difficulty to detect and block, it may well be better to restrict all development/use (provided this is feasible) rather than opening up pandora's box.

Even when there exist some pro-social applications, the negative applications can be very damaging and far-reaching. Thus, the risk of a technology like facial recognition might be too great for it not to be heavily regulated or even, perhaps, outright banned (some cities and states in the US have banned specific uses of the tech while some corporations have paused their development of the tech—maybe this is the right approach)[25].

[25] From 2019 through 2021, about two dozen U.S. state or local governments passed laws restricting facial recognition (Dave 2022). States that have passed legislation towards facial recognition and other biometric technologies (specially against use by law enforcement) include Illinois, Vermont, Virginia, California; and cities like Alameda, Minneapolis, Baltimore, Berkeley, Boston, Cambridge, New Orleans, Pittsburgh, Portland, San Francisco, among others (Future 2023) https://www.banfacialrecognition.com/map/. Sadly, some of these legislations are being undone, such as Virginia's, and possibly California and New Orleans next (Dave 2022).

Excessive data collection:

Lastly, what about the problem of excessive data collection? The last argument I want to present here is that, though I am arguing against data control as the right strategy to tackle privacy harms, focusing on usage control might have a positive impact on data collection as well. This is because under this view, different companies or institutions must properly apply for the rights for specific uses of the data they are collecting. If there are several uses that won't be granted, this might disincentivize data collection in the first place, since it becomes worthless. By interrupting the profitability of certain types of data collection (since they cannot actually use the data), the desire to collect might diminish, because data storage and management is expensive in itself. The reason why data storage is so desirable right now is because it can be either directly used or sold to a third party that will use it.

3.2. Potential ways to control use

Crafting public policy and law to govern data use is ultimately a job for policy experts and lawyers; that being said, we can offer some arguments and insights to help guide the crafting of such governing principles.

It is first useful to shed some light onto a class of approaches that are *not* useful. Some privacy advocates have argued that an ethical solution to privacy management may lie in giving users *property rights* over their data (Rose 2005; Chellappa and Shivendu 2006; Kerry and Morris 2019). The basic idea here is that individuals should effectively own their data and thus should receive monetary compensation each time they sell it to a specific company (of course, this would require an infrastructure to support such transactions). But notice that any approach founded in the concept of data ownership takes for granted that the key to protecting data privacy is to control possession of the data itself. As such, the arguments we have presented previously in this chapter apply in full force to show that this approach is ill-conceived.[26] This is

[26] Moreover, some studies have shown that payments for data would amount to a very underwhelming sum since one individual's data by itself is close to worthless (sometimes estimated to be as little as a dollar). What is really valuable is aggregated data, that is, big datasets. (Herman 2020; Jurcys 2022; Steel et al. 2013).

primarily because data ownership approaches do not allow for control once the data have been 'sold', and so cannot support use control. They assume that giving consent and getting monetary compensation for the sale of one's data is all that is morally required.[27] While it is true that some advocates for the "data property rights" approach agree that being informed about uses is necessary for consent, nevertheless, even when users are presented with some information about the primary uses of the data they are selling, the whole point of the transaction is that the *ownership* of the data is being transferred, after which it can be used at the discretion of the buyer.[28] In other words, the focus is on who owns the data, with the underlying assumption being that once you own something, you can use it as you see fit. By contrast, the proposal we outline here aims to change the focus from "data ownership rights" to something closer to "data use rights".

Let's be even clearer about the aim of the present project. We have argued that the concept of "data control" (with "data ownership" as we've discussed it being the special case of control mediated by currency) is not a useful paradigm for developing policies that effectively protect privacy. Although our argument is a practical one, the upshot is a conceptual overhaul: "data control/ownership" cannot be the moral focus of any useful theory of privacy protection. In abandoning the concept of the data ownership, however, we certainly do not mean to be nihilistic about the prospect of privacy in the digital age: rather, we are suggesting a new focal

[27] It is also noteworthy that when we talk about consent, we are talking about *informed, uncoerced* consent. If a person is "consenting" to have their data collected with no understanding of the primary (and secondary) uses of such data, this is not really informed consent. A true conception of consent for data collection, then, must include the uses to be made of it. That is, consent should focus on consent for uses and not on the collection itself (more about consent will be explored in Chapter 3).

[28] One might respond that instead of *selling* data, it could be *rented* instead, which may help avoid the problems outlined above. Specifically, if the company is only renting data from the individual, then presumably they could withdraw their data ("cancel the rental") if it was used for purposes that were not part of the rental agreement (just like a landlord can evict someone for violating terms). In a sense this idea is similar in spirit to the approach we will focus on, in that it centers on *use* control. However, a rental paradigm suffers from some crucial deficits. First, practically speaking, it puts all the onus on the individual to effectively police the uses of their data via the terms of the "rental agreement". And second, conceptually speaking, in order to apply the "rental" concept to data, we need to account for the fact that data can copied, and it can be integrated in models in a way that is hard to "undo", which threatens to take the teeth out of rescinding any rental agreement.

point around which theories and policies may turn, namely, the concept of data *use* control. The practical and conceptual infeasibility of privacy protection policies based on data control forces us to search for an alternative better suited to guard against harms caused by privacy infringement.

Is such an overhaul truly necessary? Perhaps the deficits of the ownership approach can be mitigated in other ways, such as through the installation and maintenance of proper legal mechanisms for *redress* in the case of privacy harms. This idea is pursued in the context of "Big Data" misuses in, for example, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms* (Crawford and Schultz 2014). The framework they outline is useful since it explores a way to redress privacy harms (caused both by misuses of the data or uses of incorrect data) by articulating what legal changes are required in order to have a proper notion of due process in regard to these harms. Crawford and Schultz argue that currently there are no good legal frameworks for privacy harms and propose that "individuals affected by Big Data should have similar rights to those in the legal system with respect to how their personal data is used in such adjudications." (ibid, p. 93) Certainly, setting up legal mechanisms to redress privacy harms is a necessary component of any reasonable privacy management policy. However, I argue that it is only a partial solution because any good policy on privacy harms must focus not only on redressing harm caused by data uses but also in *preventing such harm*. As things stand now, it would be too onerous for individuals to seek legal action every single time their data was used in ways that put them at an unfair disadvantage.

If we want to prevent privacy violations and harms, then, where do we start? Naturally, "at the beginning" is a reasonable place, which in this case means in the development stage of the systems and algorithms that collect, predict, and/or use our data. Much of this development takes place within corporations, and so that will be our primary focus in the remainder of this chapter.[29] Ideally, we wish to build a regulatory landscape within which companies make it an

[29] Governmental agencies also engage in such development, though in many cases they simply purchase finished products from third party companies. We will focus on policy regulation for corporate entities here to streamline

integral part of any algorithm development to question, articulate, and critically assess the *uses* of this algorithm.[30] What is it intended to do? Who are the stakeholders? What are the advantages of developing and deploying the algorithm, and who could be negatively impacted? As we've seen, attention is required both for the question-answering phase (i.e., the initial development of an algorithm and data processing) and also in light of the action-guiding goals for which the technology will ultimately be used.

To illustrate, consider again Target's pregnancy prediction algorithm. Even if in practice the algorithm might have been less successful that the media presented it to be (Piatetsky 2014; Fraser 2020), it serves as a good example to grapple with these concepts. In this case, the question-answering stage admits a clear articulation: the algorithm is designed to answer questions of the form, "Is this customer likely to be pregnant?"[31] This case allows us to consider some of the moral reflections that must take place at the question-answering stage. Notably, in this case, the question that Target is seeking an algorithmic answer to is one that they could easily ask their customers directly. It is worth wondering, then, why develop an algorithm to guess an answer to a question that could be outright asked? The answer in this case might be that, in all likelihood, many customers would not be keen to disclose their pregnancy status to Target. This should give us pause. When the information sought could easily be asked directly,

---

the discussion and avoid the additional moral and political considerations that come from considering privacy in the context of *state* surveillance.

[30] As I'll argue at the end of this section, to achieve this, governance will have to come both from within the corporations and from external groups that audit and apply pressure to take action on these matters.

[31] One might argue that Target is not asking directly: "Are you pregnant?" but only the more indirect: "Do you want coupons for diapers?". They are both indeed closely related—the latter might even be considered a proxy for the former. This is exactly what my distinction between "question-answering" and "action-guiding" use (discussed earlier in the chapter) is meant to address. Specifically, at the "question-answering" level of analysis, the two questions above are essentially equivalent. The fact that Target wants the answer to this question (either one) is of course because they then want to decide who to send certain coupons to—but this pertains to the action-guiding sense of use. It might be that (some) customers are okay with this action-guiding use. But that doesn't change the fact that to arrive at this action-guiding use, Target deploys an algorithm that effectively answers the question "Are you pregnant?" (Or an equivalent or highly correlated question). The moral status of using an algorithm to answer such a question must be assessed in its own right, i.e., in addition to the action-guiding use (which comes later). This ties in directly with the point raised below, namely that if customers are on board with answering this question *because* of the action-guiding use it will be put to, why not ask it directly to customers?

but one prefers to infer it algorithmically, this may already be an indication of technology misuse.

Target could counterargue that their customers would actually be interested in coupons based on their needs, such as diapers for a new parent, and so in general they would be happy to report on their pregnancy status. The use of the algorithm is only a matter of convenience and efficiency: it would be too troublesome for the customer to manually provide this information, and to keep it updated. In other words, the algorithm is simply meant to avoid burdening customers with long surveys and the like. But if this was the reasoning, then why not inform the customers about the practice? Why not transparently disclose they are monitoring their shopping habits to predict their lifestyle, and, more concretely, when they are primed to switch their shopping habits? Why not present a box they can check for opting in on dynamically updating aspects of their profile, such as marital status, pregnancies, etc.? They could make explicit not only the present but future monitoring. Of course, Target is well aware that this would make their customers uncomfortable. A deeper question is if they know why.

I presented this as an issue at stake in the question-answering phase, but we can see that a close examination inevitably leads one to the action-guiding sense of use as well. Customers might be unwilling to disclose their private information freely to Target because they do not know how this information might be used. Yes, on some occasions it might be used to give me discounts. But is that all they are doing with it? Are they selling it to third parties that can use it as they see fit? Are they using it to *exclude* me from possible discounts? Think for example of the issues surrounding dynamic pricing, used especially in online shopping. Amazon tracks shopping habits and has a detailed dossier of your needs and possible next buys, which might allow them to present you with some relevant deals. But it might also allow them to infer that there are items you cannot live without, and which might then be raised in price. Imagine, for example, a case in which they accurately infer their customer doesn't own a car and so they rely on online shopping and delivery for bulky or heavy items. Is there any guarantee they won't use this information to increase the prices for such items? (Most of us might have felt price-gauging when buying plane tickets. They know customers have in general very specific dates to travel,

with specific depart and landing locations. Raising prices feels almost like the default as you search). I am focusing on shopping habits as an example since it is, on its face, a somewhat innocuous form of tracking, yet even at this level it is filled with moral questions. The urgency of these debates only increases the higher the stakes (think of job hiring, loan access, education access, policing, etc.)

The point here is that the action-guiding use of the algorithm is completely essential to understand how and if a specific algorithm should be developed, or under what constraints. If the action guiding sense of the shopping-behaviors prediction algorithm is, say, ultimately the company's profit, only one stakeholder's needs are being taken into account. The action-guiding use of the tech has to be explored from all different stakeholders' points of view to assess what trade-offs have to be made.

So, to be more concrete, I now present a selection of tools, mechanisms, and policies that can collectively work to implement a use-control approach to privacy protection. To be clear, I do not here suggest that any one of these components is individually necessary (or sufficient!) for an effective use-control approach; instead, I hope to provide a "proof of concept" by outlining a collage of complementary approaches, each with some informative precedent in existing agencies or policies. What ties these approaches together is that each can be understood as focusing on *impacts* of new technologies; since impacts are ultimately determined by use, this obligates a use-centric analysis.

A core mechanism to consider is the creation of internal "algorithmic ethics boards", which developers have to consult with for each particular project, reflecting on the possible consequences for them and, importantly, the alternatives to development (including not developing it at all). These would be similar in some respects to an HR department, or especially to an Institutional Review Board (IRB), where the end goal is not only to avoid being sued but also to examine the actual *impacts* of the technologies developed and the morality thereof. This would serve as a pipeline to get approval for any project, similar to how in academic research

(think for example in psychology), all projects involving human participants have to get ethical approval before being implemented.

While the IRB approval process focusses on potential harms to individuals in the study, an algorithmic review board would take a wider scope, focusing also on broader societal impacts and potential harms. This idea has been pursued in recent work by Bernstein et al. (2021), in which they explain that existing mechanisms such as IRBs for research are designed to evaluate harms to human subjects and not harms to human society. As a result, research in AI falls outside of their purview. To solve this, they develop an "Ethics and Society Review" (ESR) board as "an institutional process that facilitates researchers in mitigating the negative societal and ethical aspects of AI research by serving as a requirement to access funding. Researchers submit a brief ESR statement alongside their grant proposal that describes their project's most salient risks to society, to subgroups in society, and globally. This statement articulates the principles the researchers will use to mitigate those risks and describes how those principles are instantiated in the research design." (ibid, p. 2) The statements consist of two parts:

I. Describing the ethical challenges and societal risks.

II. Articulating general principles that researchers in their field should use to eliminate or mitigate these issues and translate those principles into specific design decisions they are making in their research.

The ERS panel reviews the statement after the funding program conducts its usual grant merit review. Importantly, the ERS requires an interdisciplinary panel with expertise on technology ethics and society, and after evaluation they work with the researchers in an iterative process in which they provide feedback to support a more comprehensive statement, and finally give their recommendation to the funding program. Though this example takes place inside universities, a similar approach could be applied in corporate settings. In fact, Microsoft

has been on the vanguard of this, by establishing their Research Ethics Review Program back in 2013.[32]

An internal department of this form would not only approve or deny specific projects, but would also be involved in inquiring how to get the relevant *consent* from users. Question like the one explored above, become relevant: Why aren't we asking customers for their pregnancy status? What constraints/limitations would make them comfortable to disclose that information? Do we really need to predict it? In general, the department will develop guidelines to define technology misuse and ethical use, showing what red-flags to look for.

An important tool that an algorithmic review board might deploy are Algorithmic Impact Statements (AIS). AISs are modeled after Environmental Impact Statements (EIS) (of the National Environmental Policy Act (NEPA)); Selbst (2017, p. 168) discussed them as possible regulatory solutions for different data-driven prediction systems. Indeed, Selbst explores this framework in the context of algorithms designed for policing. In his words: "Impact statements have become a much-emulated regulatory tool where the problem at hand is a lack of knowledge about the effects of a particular type of decision. While AISs will not necessarily achieve the full measure of accountability that will eventually be required, they will be useful—and perhaps necessary—to determine what, if anything, society will need to do next." (ibid, p. 169). As above, the reason this approach is interesting is because it focuses on *impacts* and therefore must take the *use* of the algorithm as an essential part of the reflection. Other scholars have considered using AISs for regulating mass surveillance (Froomkin 2015). As a legislative requirement, AIS can be seen as a model of the conversations that should happen inside a corporation as they develop new algorithms. In a nutshell, AISs propose six requirements:

> A. Rigorously explore and objectively evaluate all reasonable alternatives, and for alternatives which were eliminated from detailed study, briefly discuss the reasons for theirs having been eliminated.

[32] https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1%3aprimaryr5 (Microsoft AI 2023)

B. Devote substantial treatment to each alternative considered in detail including the proposed action so that reviewers may evaluate their comparative merits.

C. Include reasonable alternatives not within the jurisdiction of the lead agency.

D. Include the alternative of no action.

E. Identify the agency's preferred alternative or alternatives, if one or more exists, in the draft statement and identify such alternative in the final statement unless another law prohibits the expression of such a preference.

F. Include appropriate mitigation measures not already included in the proposed action or alternatives." (Selbst 2017, p. 172,3).

A modification of these requirements is useful for an internal evaluation where developers themselves have an active role in reflection before designing a specific algorithm. Indeed, these requirements integrate with the discussion above. For instance, in a conversation surrounding alternatives as per requirement (a), one consideration would be whether they should just ask their customers for the desire information rather than algorithmically infer it. Choosing to try to predict some information that could easily be asked directly might be a red flag. Requirement (d) also seems highly relevant: should we just not develop this at all? What are the costs? These reflections require some account of the different stakeholders and their values. Trade-offs (not only monetary but moral as well) will have to be made, since different stakeholders hold different (and sometimes contradicting) goals. Questions surrounding (f) might include constraints on data storage or sharing that are required for mitigating the vulnerability to harms, as well as conversations about assuring informed and uncoerced consent from their users.

In addition, having this IRB or AI ethics department within the company would also give more voice for employees who sometimes get no vote in the general industry/company policies. Often developers (employees involved in coding or designing the technology) have no real say into the problematic aspects they see in the technology. This structure involves them directly as well as higher ranks, giving them more power to participate in the creation process.

Finally, it is worth mentioning that although some companies will be interested in implementing this approach unilaterally, many others won't feel the need unless there are real external pressures to do so. How many academics are thrilled about applying for IRB approval? Of course, some might indeed welcome the process, believing that it improves their experimental designs. Similarly, the ESR board discussed above has been welcomed by many researchers (Bernstein et al. 2021). Still, it is not hard to imagine contexts where the various incentives don't fully align; this is where governance becomes important. Having external organizations that focus on auditing companies might be a necessary tool.

While the approaches we have outlined focus on preventing harm, of course such policies must be complemented by better legislative solutions for accountability, enforcement, and redress. Again, as we've argued, the most workable and effective approaches will be those that center on governing uses of data rather than controlling the data itself.[33] Part of the external pressure will likely have to come from government oversight, but public engagement is also a crucial aspect of proper governance, which relies on ways of making relevant information available to the general public.

Section 4 | **Some Considerations**

We conclude with a brief consideration of two very different sorts of concerns that might be raised in the context of the framework we have proposed. First, is the use-centric approach too narrow in the scope of what it can count as harm? And second, is the practical task of articulating uses so unconstrained as to render effective regulation impossible?

Regarding the first point, are we really saying that we should focus *entirely* on data use, and therefore giving carte blanche to any data collection as long as it doesn't involve use (or only involves licensed use)? In a narrow interpretation, yes. To illustrate, consider the following

[33] As I suggested in the last section, focusing on uses control might also decrease data collection and storage, since data is worthless unless it can be used, while data maintenance and storage is costly.

example: suppose your smart phone is constantly "listening" to you, in the sense that it is recording all the sounds around it and storing the audio files on some server somewhere. Some might feel that the mere collection of this data is already morally problematic; others might argue that no harm is being done as long as no one is looking at or using the data in question. Superficially, the approach presented in this chapter seems to support the latter interpretation, since there is no use of the data and therefore no harm has been done. However, it is of course the case that the stored data could be breached by an unauthorized agent; in this case, the company storing it would be liable for the resulting harms. I argue that we can push this reason a step further—we need not wait for a breach to find the storage of the data problematic. The storage itself creates the possibility of harm (through a breach), and in some legal context needlessly creating a possibility of harm is itself a crime (think of reckless endangerment, which is a crime even if no one is actually hurt). We might apply the analogue of reckless endangerment to establish a legal context for prosecuting certain types of data collection which lend themselves to abuse, even before any abuse (or even use) actually occurs.

Potentially a similar "reckless endangerment" argument could be made for "stalking", i.e., if someone is listening to all your conversations but not (yet) actually acting on any of the information they obtain. So, the upshot is that we might actually find legal precedents that support many of the "common sense" prohibitions on data *access*, even though the underlying framework only directly attaches harms to data *use*.

The second challenge has to do with the wording of the uses. What we've attempted to develop above is a general, conceptual framework for regulating *uses* of data in a way that protects the privacy of individuals despite the challenges imposed by the digital spaces we now occupy. Of course, transforming these ideas into specific regulations and/or practices comes with its own challenges. It is a general phenomenon that when rules are actually written down, they become subject to exploitation and gamification—put more simply, powerful players will always try to find ways to circumvent the rules as stated based on technicalities or loopholes. The present context is no different, and although we cannot solve this general problem here, we can gesture towards the type of exploitation that is likely to arise. Since our framework revolves

around regulating *uses*, both question-answering and action-guiding, a key "stress point" of any practical implementation will be exactly how such uses are articulated and distinguished.

More concretely, consider again the pregnancy prediction example. We have suggested, among other things, that in order to deploy algorithms that answer the question "Are you pregnant?", companies such as Target ought to be obligated to obtain consent (in some form). But rather that asking their customers whether they're okay with the pregnancy status being predictively tracked, Target might change the way they articulate the question. Instead, it might be "Are you likely to buy diapers soon?", or (if this is too transparently a proxy for the previous question), "What are you likely to purchase soon, based on your previous purchase history?" In this way Target is able to deploy essentially the same algorithms but in a way that never explicitly mentions pregnancy, and so in principle avoids any legal or procedural requirement to disclose this or obtain consent. Considering this, it seems clear that effective regulation will have to be robust against such "rewordings" of the target questions.

We sketch here a partial solution. First, in the above "sanitized" version of the question, there is no hint of the *action-guiding* use of the algorithm. Once this is added (i.e., we will use the answer to this question to determine whether to send coupons for diapers), we have a better handle for regulation. But notice that this most general version of the question lends itself to *hundreds* of potential action-guiding uses; if the regulatory framework is set up correctly, this in itself will be a strong disincentive for Target and other companies to articulate the question in this way, since it will entail a requirement to obtain consent for all these hundreds of uses it might be put to (or else severely narrow the scope of the action-guiding uses that the answers to the question can be put to). In other words, in a proper regulatory environment, Target would be incentivized to more narrowly articulate the questions, they are answering precisely because that is the only way they can feasibly get consent for the associated action-guiding uses.

CHAPTER 2

Intermediate cases: **Between Guessing and Observing**

INTRODUCTION

In Chapter 1 we discussed the blurring between guesses and observations produced by predictive algorithms. In the extreme case, as we saw, there's a kind of "leap" between a guess and an observation, meaning that what information in the past was only enough for a rough guess, now can be analyzed by an algorithm whose output will be an accurate prediction, functionally equivalent to an observation in many ways. We examined the ethical consequences of this blurring and in particular how such leap cases can produce violations of privacy, and we explored possible responses and solutions to this kind of unprecedented loss of privacy. A crucial point here was that privacy infringements are ethically problematic because they lead to vulnerability to harms.

The premise of the first chapter was that there is a moral difference between guesses and observations. While it might be permissible to guess someone's information (even what they might not want to disclose, for example what they look like naked), unsolicited, non-consensual observations are morally problematic in way that goes beyond guessing (say, taking a picture of the person while they are in a changing room). This is not to say that all guesses are morally neutral (trying to guess someone's PIN to empty their bank account already involves morally questionable intentions), but rather that observations are morally and categorically *worse* than mere guesses in that they actualize certain specific vulnerabilities to harm. In Chapter 1 we discussed these privacy-based harms that come from highly accurate inferences. Naturally, they were multi-faceted and varied from case to case (after all, the technology in question applies to a very broad range of situations).

Nevertheless, a question arises: to what extent, if any, do these specific, privacy-related harms of highly accurate predictions persist when the inferences are not, in fact, highly accurate and

therefore cannot be construed as functionally equivalent to observations? I will make the case that these kinds of harms can indeed persist, though sometimes in more subtle forms that require more careful conceptual analysis. In particular, I will argue against the idea that these specific sorts of harm *only* depend on accuracy, i.e., that as error goes to 0, so too do these harms. I will show that such harms can also be rooted in "*perceived accuracy*".[34]

INTERMEDIATE CASES

We are concerned here with the grey area of "somewhat accurate" predictive algorithms: roughly, those that produce outputs that are too accurate to count as guesses, but not accurate enough to be considered observations. Nonetheless, they are sometimes still treated as "sufficiently" accurate in some sense. This is the type of blurring that will be considered in this chapter. The central question is how much these cases differ from the 'leap cases' of Chapter 1 in their moral implications. Can we apply the same framework to analyze the harms? To the same degree? Importantly, the answer to these questions might depend not only on how accurate these algorithms really are, but also on how the algorithms are treated. Indeed, one can argue that to the extent they are treated *as if* they are accurate observations, such intermediate cases carry many of the same moral implications as genuinely accurate predictions.

So, to start, I want to present a clear organization of the landscape we are dealing with, which can be organized based on the two key variables just discussed. First, we have the *real accuracy* of the predictive algorithm: Is the prediction essentially 100% correct? No better than a random guess? Something in between? Second, we have the *perceived accuracy* of the algorithm: Do people perceive it as generating accurate information or not? This latter variable corresponds naturally with how the algorithm is treated/implemented/used in society. These distinctions generate the following high-level 2-dimensional layout of the conceptual space we are working in:

[34] When we speak of harms that stem from perceived accuracy, one mechanism that might come to mind is the following: an inaccurate algorithm that is perceived to be accurate will be over-used, and this will multiply the harms that come from its inaccuracy (e.g., medical misdiagnoses). While this is a real kind of harm, it is not the main focus of this chapter. We instead direct our attention to what we will term *presentational harms*, introduced below.

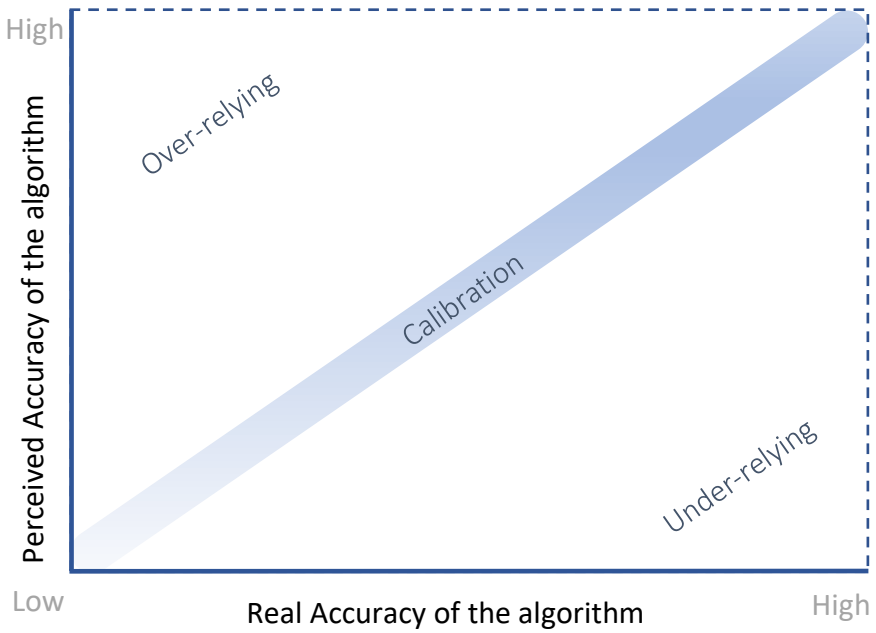

Figure 1

Chapter 1 explored in depth the "high real accuracy" region without much explicit attention paid to perceived accuracy; tacitly we focused on cases of high real accuracy that were also taken/treated as such. This is where we see "guesses" effectively become observations, and the moral distinction between guessing and observing becomes relevant since such "guesses" count as infringements of privacy.

In this chapter, we organize our exploration according to the three broad regions labelled in the figure above: "over-relying", "under-relying", and "calibrated". However, we begin in Section 1 by establishing some important foundational concepts relevant for the later analysis. In particular, we discuss the importance of the difference between real accuracy and perceived accuracy in terms of two relevant aspects of harm: *informational* and *presentational*; we also distinguish two broad categories for assessing accuracy: present- and future-directed predictions.

In Section 2 we discuss what we call "over-relying" cases. These might be the most common cases nowadays: those in which our perception of the accuracy of the algorithms is higher than its factual accuracy. We will focus on harms that come specifically from this misalignment,

considering cases ranging across different actual accuracies (the upper-left region in Figure 1). For example, when the actual accuracy lies somewhere in the middle of the spectrum, this corresponds to cases where the inferences are, as a matter of fact, not completely accurate, but they are more accurate than random guesses. However, they may be incorrectly taken to be highly accurate and/or treated as such; this is the top-middle region of the graph. Several algorithms currently deployed fall into this category, including several recidivism algorithms, policing algorithms, hiring algorithms, etc. In essence this encompasses cases where humans over-rely on an algorithm to make a decision, trusting its accuracy more than they should. This of course can have real impacts on the lives of the people who are submitted to it (e.g., getting or not getting the job, receiving or not the loan, being approved or not for parole, etc.).

In Section 3 we will briefly consider what we call "under-relying" cases. In such cases by definition the perceived accuracy is actually lower than the real accuracy of the algorithm. If algorithmic inferences are partially accurate but are being treated as if they are less accurate than they actually are, in a sense the algorithm is being "under-used". While this may be problematic in certain cases where we would stand to benefit from deployment of the algorithm (e.g., for medical diagnoses), our concern in this chapter is specific to the harms that stem from privacy violations, broadly construed; in this context, under-use of algorithms is not particularly relevant.

In Section 4 we turn our attention to the strongest cases for proponents of algorithm inference: alignment between perceived and real accuracy, i.e., calibration. These cases lie on the diagonal line of the graph. Here, the inferences are not completely accurate (except of course in the upper-right corner, but these cases were the subject of Chapter 1, so we ignore them here), but they are also not treated as if they are highly accurate—by assumption, they are treated as if they are about as accurate as they actually are. This is what developers of predictive algorithms might consider the ideal case. They might concede that both overlying and underlying on the technology can bring unwanted problems and even harms, but still argue that when there is alignment, there is no misuse, and so such deployments can be considered responsible. In this setting the notion of "presentational harm" we develop in the next section will be put to use, and we will have to grapple directly with the question of when such harm is warranted. We will argue that while the

"calibrated" case is indeed an improvement over the over- and under-relying cases, it does not in itself completely circumvent the possibility of privacy infringement and harms; when such harms are acceptable (or even desirable) is highly context sensitive. Thus, achieving calibration is necessary but not sufficient for assessing the ethical development and deployment of predictive algorithms.

Section 1| **Foundational Concepts**

1.1. TWO ASPECTS OF HARM

Our analysis will be framed by a distinction between two aspects of harm: "informational" vs. "presentational" harm. The former is meant to capture the kinds of harms that stem directly from a piece of factual information becoming known and thereby making the agent vulnerable to damaging actions. This is perhaps easiest to delineate in cases where the harm is *purely* informational, e.g.: someone learning my PIN, my email password, my address, my mother's maiden name, the place I hide my money, how much money I have, "corporate secrets", etc. The point is that in all these cases, knowledge of some piece of true information opens the individual (or company) up to loss: I lose the money in my bank account or in my mattress; I lose control over my email account; I lose a bidding war (because my opponent knows how high I can go); I lose business (because my competitor knows the secret recipe for what I sell); a harasser learns where I live, who my family is, etc.

By contrast, presentational harm comes not from being subject to loss on account of some fact becoming known, but from being *presented* in a way that leads to negative (social) consequences. For example, you may be presented as gay, or pregnant, or as trying to get pregnant, or likely to commit a crime, or bad with money—and on account of this presentation, people may perceive you differently, and opportunities may be closed to you. You may even perceive yourself differently and/or be disturbed by the belief that others think of you in a

certain way, *whether it is accurate or not*. It is important to note that these two aspects of harm can take place simultaneously: e.g., if you are pregnant and someone learned that they could increase the price of diapers when you search for them online. This would be an informational harm; notice that the damage is contingent on you actually being pregnant (and therefore needing diapers, so having to pay the higher price). At the same time, being presented as pregnant could also cause a prospective employer to pass over your resume; here, it doesn't ultimately matter whether you are pregnant or not—the harm derives merely from being perceived that way.[35]

In short, the distinction hinges on the question: does the harm stem from something *true* being revealed, or does it consist in something being presented in a certain way (true or not)? So informational harm tracks *accuracy*, whereas presentational harm tracks *perceived accuracy*. (Of course, as always, all of this assumes the individual does *not* want others to know or believe X about them; cases where information is given freely do not concern us within the framework of considering the harms associated with privacy violations.)

Our core purpose in articulating this distinction is to counter the implicit assumption that predictions become "automatically" morally permissible whenever they don't count as observations. This is only plausibly true in cases where the associated harms are *purely informational*, in which case blocking the transmission of (factual) information also blocks (or at least mitigates) the harm (think of someone having your wrong PIN, they will not be able to empty your account). In reality, most harms are a mixture of these two aspects, so it would be too hasty to say that as accuracy goes down to zero, so does harm. The harm may be lessened

[35] Note that Kate Crawford **(2017)** has talked about a somewhat similar distinction between "representational harms" and "allocative harms". In her view though, representational harms focus on identity categories (gender, race, etc.) and how they are being reinforced or denigrated in society (stereotyping, denigration, under-representation). Allocative harms refer to how economic and resource-based benefits are allocated. Our distinction presented here is different from Crawford's: we do not focus on how resources are allocated, nor (only) on identity categories. The focus of our distinction is on whether the vulnerability to harm comes from information that is factual (i.e., knowledge) or from how we are presented to be (i.e., belief).

(since the informational aspect is evaporating), but the presentational aspect can persist, and moreover, be quite pernicious.

It is also worth observing that as accuracy goes up, presentational harms may be harder to counter. This is because it is typically easier to deny/refute a false presentation (and thus, perhaps, escape or lessen the consequences) than a true presentation. So, while the harm itself may not come from something being true, the 'persistence' of the harm may be correlated with its truth.

Overall, here we are presented with a distinction between "vulnerable to harm due to (knowledge of) X" vs. "vulnerable to harm due to people's belief that X". At the end of this chapter, we aim to show that, since presentational harms involve harms that stem from people's beliefs, we can reduce these harms in two main ways:

(a) changing people's beliefs (show that the information was incorrect/inaccurate);

(b) stopping the harmful behavior (use control: decide that some information can't be used against you in specific contexts. E.g., they can't pass you for a job for being pregnant).

Before we try to apply this distinction to cases of partially accurate inferences, it will be useful to understand one more challenge that makes thinking about "partially accurate" inferences hard, and how this relates to the informational/presentational harms distinction presented above.

1.2. Present versus Future-directed Assessments of Accuracy

In thinking about algorithmic predictions, we can distinguish two broad categories for assessing accuracy. On the one hand we have present-directed predictions (e.g., "are you pregnant?", "are you over 21 years old?", "do you live at X address?", "are you gay?"). By contrast, many algorithms are essentially designed to output predictions of *future* behaviors, which of course means that there is no *present* way to directly assess their true accuracy ("will you recidivate?", "will you pay that loan?", "will you be a good worker?", etc.).

When dealing with these kinds of future-directed outputs describing facts or behaviors that have not yet come to pass, there is a more abstract sense of what, say, "80% accuracy" actually means and it's harder to measure and quantify this accuracy. Steering clear of the vast philosophy of science literature on measurement and forecasting in general, the central point we want to make here is that this abstractness often serves to shift more harms from an informational to a presentational aspect.

An example will be helpful here. Many algorithms that claim to assess "personality traits" of an individual (e.g., hiring algorithms that aim to gauge how lazy or productive an employee is, or how good of a leader, etc.) are really trying to predict future behaviors (will you take shortcuts, complete your work in time, competently manage a group project, etc.). In other words, such talk of "traits" is, in reality, an indirect way of speaking of future behavior, and therefore such algorithms are best understood as making future-directed predictions. Such predictions can only be assessed to have been accurate or not after the fact, in retrospect. In fact, in many of these cases we might *never* find out if the prediction was correct (for instance, if someone is denied a loan because they were considered at risk of defaulting, we might never find out if that was true, since they never got the loan in the first place). This makes it very tricky to assess accuracy; moreover, we can see that while it is tempting to label the associated harms as informational, in many cases they are actually presentational: it is not that you have been harmed in virtue of the true fact that you are lazy being revealed, but rather that the belief that you are lazy has been created, and this harms you (whether it is true or not). This is a key point so it's worth emphasizing: Future-directed model outputs (almost) necessarily involve presentational harms (if there are harms at all), since it is hard to see how I could show that something in the future is false (since it hasn't happened yet).

Since, as discussed, privacy harms include presentational harms, we argue that future-directed outputs can constitute privacy harms even if they are not, strictly speaking, "true" (i.e., even if they are not revealing true information about the subject). Framing harms as presentational in this context allows for an understanding of how privacy harms can stem from

the kinds of future-directed predictions that don't, in a literal sense, reveal "true" things about the subject.

Crucially, what this means is that *perceived accuracy* has an integral role to play when thinking about vulnerability to harms. And in a landscape where assessing real accuracy becomes more nebulous, perceived accuracy can be easily manipulated.

1.3. Over-relying (over the line), under-relying (under the line) and calibrated (on the line):

We now turn to explore the regions of the graph presented at the beginning of this chapter. The graph had two axes: real accuracy and perceived accuracy, with a diagonal region through the middle. This line represents a "match" or calibration, namely, when our perceptions of accuracy and the real accuracy align, and it naturally divides the possibilities into 3 regions. Framing the possibilities in this way corresponds to the idea that when assessing vulnerabilities to harms we can't consider only real accuracy, or only perceived accuracy, in isolation. They interact—in other words, the problem space is "2 dimensional", as shown in the graph, and we must consider real and perceived accuracy in tandem.

Section 2 | **Over-Relying**

We consider here those cases in which partially accurate inferences are treated as if they were more accurate than they really are and analyze what kind of harms can result from this. It is important to consider these cases since they might very well be the most common cases of algorithms deployed nowadays. Countries like the US are widely deploying algorithms to assist or make decisions in many aspects of our lives, from financial decisions (who gets a loan) (Ereiz 2019), to work related questions (who gets a job) (Mahmoud et al. 2019), to education (who is accepted to a university) (Al Mayahi and Al-Bahri 2020), to healthcare choices (how much you pay for insurance) (Kaushik et al. 2022), to policing (predictive policing, sentencing, paroling) (Berk 2021), and so on (Kumar, Kaur, and Singh 2020). The ample literature of ethics and AI that has exploded in recent years has showed time and time again that the algorithms that were

eagerly deployed were more problematic than the developers and the ones deploying them wanted to admit (Christin, Rosenblat, and Boyd 2015; Barocas and Selbst 2016; Binns 2017; Araujo et al. 2020; Marcinkowski et al. 2020; Char, Abràmoff, and Feudtner 2020). And beliefs about the accuracy of these algorithms have in many cases been much more confident than warranted.

To get our bearings, let's start with an example of a *simple, present-directed output*: an algorithm that is meant to guess a 4-digit PIN. If it gets it right 1% of the time, this is two orders of magnitude better than chance (which would be 1 out of 10,000 i.e., 0.01%). In this case, the algorithm can't 100% predict your PIN but it's much better than a guess.[36] So even if we are not dealing with an "observation", we can see that the harms, and specifically the informational aspects of the harms, are similar insofar as the degree to which malicious agents may have access to information about you, even some of the time, increases your vulnerability to those informational harms. Imagine, conversely, that a bank decided to make their PINs only 2-digit instead of 4. We can see that the chance would be 1 in 100 instead of 1 in 10,000 to get access to your savings. In this scenario, rightly so, the bank would be liable for the harms that would come from leaving the bank accounts of their clients so insecure. By the same reasoning, even this kind of partially accurate algorithm carries a *degree of vulnerability to informational harms directly proportional to their accuracy*: the more accurate the information that can be inferred, the more the vulnerability to harms increases, and vice versa. This applies to other examples of informational harms as well, such as the damage that can be brought by a harasser knowing your home address, or a totalitarian state being able to identify you in a protest, or any harm that comes from knowing factual information about you.

In this case the perceived accuracy was not very relevant at all, except insofar as the algorithm was perceived to be accurate enough to actually use it, to deploy it in society. All that really mattered was that the algorithm was in fact accurate enough to cause harm some of the time. But algorithms with present-directed outputs also can carry presentational harms that obtain

[36] For instance, the algorithm might analyze your typing speed and patterns and use that to make an "educated guess" about which sequences of numbers are most committed to your muscle memory.

even when they are inaccurate. If the algorithm is *treated* as being accurate, even when it's not, the consequences can be very real. For instance, let's return to the example where you are passed over for a job because an algorithm has inferred that you might be pregnant (and therefore assumed to be less dedicated to your work). Even if this prediction is wrong, the harms from discrimination can persist simply through the perception/belief that it might be right.[37] Moreover, even with present-directed outputs, it may be very difficult to correct false predictions. Sometimes this is because the user has no access to the decision taken: maybe the algorithm assumed you had young children and you actually didn't, and although this would be easy to prove, neither the people who rely on the algorithm nor you know how it operated, so false inferences will never be exposed to be corrected.[38] This issue is most pronounced with black boxes algorithms (common in neural networks), which are intrinsically opaque (Burrell 2016). Any factual error in your profile (wrong age, address, nationality, gender, past employment, errors in your credit scores, errors in your education transcripts, the list goes on) might not be salient and therefore there may be no opportunity to fix them, even if in fact they could be easily disproven (Doshi-Velez et al. 2019).

When we look at future-directed outputs, things don't get any better. The most evident examples are "risk scoring" algorithms, but the issues apply to any algorithm that is trying to infer something about the future, such as your ability to pay a loan, your risk of recidivism (the likelihood you will commit another crime once released from prison), or the probability you will have kids in the near future, among others. When the prediction is based on some future characteristic, since the decision is made in the present to give or deny you access to the relevant good or service or opportunity, the harms persist even if it turns out later that the algorithm was wrong. Discrimination can occur because these algorithms as *perceived* as being accurate,

[37] We are using "perceived/ believed" generally here, since it might be the case that no human actually has access to this information, but it's only a variable used by the ML algorithm to assess prospective employees.

[38] And even if it's possible to see that they were wrong, effort is required to correct the record. For an clear example think about how the people who are victims of identity theft incur great costs (int time and money) and loss of opportunities in trying to set the record straight.

regardless of whether they really are. Opacity of the algorithm and lack of transparency for how it is being used only makes it harder to see its limits/failings and change our perception of it.[39]

In general, when considering the region above the line in the graph, the problem at hand is that we are *over-trusting* or *over-relying* on the technology. The cases explored above are cases in which the technology is being used to try to reveal something about the individual, and in doing so infringe on the privacy of that person. Of course, there are also cases in which the use of the algorithm is something the person *desires*, such as being assessed for the risk of developing a specific illness in the future (developing cancer, for example) in the hopes to get the best preventative care in the present.

So we want to be clear about a distinction between two routes to harm in the "above the line" region—one the one hand we have presentational harms that stem from high perceived accuracy (which is what has been discussed up to now); on the other hand are those harms that

[39] It's worth briefly talking about why it is hard to cleanly define what is meant by "accuracy" and "partial accuracy", since this adds an extra layer of complexity to the assessment of these algorithms overall.

There are really at least two ways to judge (partial) accuracy, Depending on the *type of output* of the algorithm. On the one hand, some algorithms work in a binary way, producing outputs that are essentially either correct or not, with no real grey area in between. In this case, accuracy might be construed in terms of the percentage of times the algorithm gets it "right". In this context "partially accurate" would then describe an algorithm that produces a **binary** outcome and has a better than chance probability of getting it right (think of the PIN example). Such an algorithm would, of course, generate a lot of false positives in the sense that many. (Perhaps even most) of its guesses would be wrong, but it could still be right much more often than pure chance. On the other hand, some algorithms' outputs are complex and cannot be cleanly divided into "right" and "wrong", but rather "closer" and "farther" from the truth, thus falling on more of a **spectrum**. In this case a partially accurate algorithm is better understood as one that is getting it "closer to right" than chance sufficiently often. Instead of looking for a binary "right or wrong" inference, the outcome is an approximation to the truth. For example, think of a 'deepfaked' image with some details wrong, but perhaps close enough to be "recognizable". All *synthetic media* falls into this category: the relevant question is not whether the output is right or wrong, but "how right" it is.

These considerations demonstrate how complicated it is to understand what exactly is meant by "accuracy" and how we might gauge measures of accuracy (like percentages). For our purposes, the problem with measurement of accuracy is relevant insofar as it is directly tied to the perceived accuracy: more and more we live in an environment that deploys and relies on such partially accurate algorithms for many tasks in society. In essence, some outputs are binary, clearly "right" or "wrong" while others land more on a spectrum, even in present-oriented cases. So, in this sense it really can be challenging to specify what is meant by "accurate" or "partially accurate", which compounds the difficulty of the overall assessment of how accurate the algorithm is.

come from over-trusting the algorithm to accomplish a certain (desired) task that it then fails to do (or fails more often than expected to do), because in reality it's less accurate than you thought (e.g., medical diagnosis, finding a good romantic partner, etc.). The latter cases would not count as privacy infringements since the person is assumed to be a willing participant (typically because they think the output of the algorithm is going to be used for their benefit, rather than to charge them higher premiums for insurance in the future, for example). Such cases can still involve harm to the degree that over-relying on the algorithm can have real consequences, such as embarking on the wrong kind of treatment or lack of treatment due to a false diagnosis. Nonetheless, such harms, while very real, are not the focus of this work.

2.1. The appeal of automation

A practical standard to evaluate what (descriptively) has been considered to be "accurate enough" is *deployment* of said algorithm. Many of the cases discussed thus far have been considered accurate enough to be implemented in society. This willingness to deploy partially accurate algorithms, plus the challenges in developing proper recourse to challenge their decisions—the lack of due process or proper regulation to audit and have oversight over their implementation—all serves to show the extent to which vulnerability to harm that stems from high accuracy (Chapter 1) carries over to cases of high *perception* of accuracy.

The existence of highly accurate algorithms serves to increase the faith and bolster support for algorithmic-based solutions across the board. This phenomenon is sometimes referred to as "automation bias", although this terminology is often more narrowly construed as applying to the extreme case where algorithms are preferred unquestioningly based on a general belief in the superiority of technology (Skitka, Mosier, and Burdick 2000). Our arguments are broader than this: there's a wide range of reasons why perceived accuracy can be high, ranging from marketing strategies, lobbying, desires to cut costs (wishful thinking), AI hype, and even straight-out opportunistic grifting. Of course, as we approach the upper-left corner of the graph, where accuracy gets lower and lower but perceived accuracy stays high, we are dealing with cases where it is essentially irrational faith in technology that is carrying the day, , i.e.,

positive automation bias (as opposed to algorithm aversion (Dietvorst, Simmons, and Massey 2015)). More generally, however, automation bias is only one of many potential reasons why people might perceive an algorithm to be more accurate than it really is.[40]

In a sense, in this space we encounter cases where the use of technology is not necessarily the most morally salient factor. We might think we are in the more mundane situation of false rumors: incorrect/invented information about something or someone that gets spread and ends up widely believed to be true. In the "pre-technological past", though, the virality of rumors/false information might be more contained.[41] The central point here is that we ought to be wary of the false authority that algorithmic predictions may carry in our society, not only because they are wrong, but because the perception of their accuracy can greatly magnify their reach. Predictive technology has limitations and understanding them is important: as a society we must reflect on the influence of technology on the proliferation of such "false rumors" (Diresta 2018; Forberg 2022), including the false authority that automation bias may carry in our society.

Though here we are focusing on limits of algorithms to properly predict (their real accuracy), careful thought should also be given to what we should *want* to predict: a sociological, as opposed to technological, reflection on the role of technology in society and how we may flourish as individuals. Descriptively speaking, the use of predictive algorithms is increasing, which as argued, demonstrates a general agreement as to their quality and accuracy. But a pressing question in this sphere concerns the normative aspect, namely, *should*

[40] In some cases, the direction of influence might be unclear, since we might mis-perceive the accuracy and therefore appear to have a pro-algorithm bias.

[41] There's a different argument to be made about "accurate" algorithms increasing the reach of false information (disinformation and misinformation) where it's clear that an environment that over-relies on algorithms is one that will make false information thrive even more. This is widely discussed in the research around misinformation and the role of tech: curation algorithms for feeds (which are meant to predict what the user might want to watch) have already amplified fringe ideas (Forberg 2022; Diresta 2018). Nevertheless, these are algorithms that seem to be accurately predicting that emotional/anger inducing videos or stories make us engage more with the platform, so it's a different issue from the one debated here, where trust in algorithms make us accept more and more the use of poorly designed algorithms with low accuracy.

we be readily willing to employ these algorithms when we are not completely certain of how accurate they are? A central theme of this thesis is to give us some pause, to understand at least some of the ethical impacts of this progression. Our argument has focused on the vulnerability to harms that come with the ability of algorithms to predict, and how many of these harms can persist if the perception of accuracy remains high. In essence, what is required is a proper evaluation on the question-answering use and the action-guiding use of the algorithm, as discussed in Chapter 1. We return to discuss this question in Section 4 and the concluding section.

So, what to do in this situation where partially accurate algorithms are being widely deployed? Deployment is linked with the belief/perception that the algorithm works as intended. An obvious intervention, then, is to focus on changing people's beliefs about the algorithms, so they match their real accuracy. But is this enough? We'll return to this when reflecting on the cases that are on the line (where perception and accuracy align, i.e., calibration), in Section 4.

2.2. Disparities in perceived accuracy among different groups

We have been tacitly treating "perceived accuracy" as if it is uniform across different groups, but in reality, there might be disparities in the "perceptions" of the algorithm by different parties. As a first pass, we might identify four relevant (possibly overlapping) groups regarding perception of accuracy: the *developers* of the algorithm (i.e., those who design, code, test, etc.); the ones *deploying* the algorithm (i.e., what we may conceive of as the "users", e.g., the police officers who use predictive algorithms for their job); the people who are *subjected* to the algorithms (i.e., the individuals who are evaluated through them or otherwise directly impacted by its outputs);[42] and finally, the general public. In an ideal world, these groups would coincide in their perception of the quality (accuracy) of the algorithm, since they all would have access to the same factual information about the algorithm and how it works (i.e., their

[42] In some cases, the deployers and the subjects of an algorithm may overlap or be the same. For instance, when an individual chooses to use a personal app, e.g., a dating app that aims to connect them with "good matches".

perception would match the actual accuracy of the algorithm, which is the "calibrated" case we explore in the next section). More realistically, however, this will not be the case: information about any given algorithm is certainly not guaranteed to be equally available to all parties—even (and in some cases, especially) those who are directly affected by its use.

As a practical matter, the only two parties that typically have any real say in the use of an algorithm are the developers and the deployers, while individuals subjected to them often have very few options to avoid interacting with the entities that deploy them. But it's worth noting that presentational harms can come not only from what a deployer comes to believe on the basis of the algorithm, but sometimes also from what "society" thinks of a person. High perception of accuracy by the second group (deployers) is a necessary precondition for use (why would they use it if they don't think it's accurate?), but this can also result in a high perception of accuracy by the general public. Think, for example, of the reputational harms that can be incurred by deepfake technology or other synthetic media—even in cases where the actual accuracy is not high, many harms can stem from the perceived accuracy. This is the nature of presentational harm.

It is worth being explicit about what was implicit in the above, namely, that it is the people *subjected* to the algorithm who are typically the ones who face potential harms (be they informational or presentational). It also may seem natural to understand the people *deploying* the algorithm as the parties *responsible* for these harms; however, in many cases the *developers* may also be culpable for certain harms. The distinction between "question-answering" and "action-guiding" uses of an algorithm, explained in Chapter 1, tracks this well: developers may be culpable for harms that stem from designing algorithms that (1) seek to answer harmful questions or (2) fail to answer the question they think they are answering (e.g., via bad proxies). Deployers, meanwhile, have responsibility for how they actually use the algorithm in society and what effects it has on the people subjected to it.

Moreover, although it is typical to focus on harms caused to the individuals subjected to algorithms, there are also broader harms that can "leak" out to our fourth group, the general

public. In a society where algorithmic predictions mediate more and more interactions, the fear of informational or presentational harms can easily generate chilling effects that impact the lives even of individuals who are not (currently) directly subjected to the algorithm per se. They might choose to avoid certain kinds of activities such as going to a doctor or psychologist, interacting with the police, applying for jobs in certain sectors, using phone apps that track location, or freely communicating via email or instant messaging services, or in general anything that may leave too much of a digital trace.

Our analysis in Chapter 3, which will focus on *use control*, aims to lay a foundation for policy recommendations that target both developers and deployers taking into account harms for both individuals subjected to algorithms and the general public.

Having identified the four groups above, it is important to observe that there is a significant power imbalance between them. Individual subjects often have very little capacity to (1) know they are being subjected to an algorithmically assisted prediction/decision, (2) understand the role of the algorithm in that decision, and therefore (3) muster any recourse to combat the decision if they find it untrue/unfair. If, in addition, the individuals subjected to the algorithm have a high perception of its accuracy (so they fall in the over-relying region of the graph), this creates yet a further obstacle to challenging the ultimate decision (even though it might be based on faulty predictions). Insofar as this is the case, any viable solution requires the involvement of the people subjected to the algorithms, at least if we wish to establish any strong governance foundation.

Section 3 | **Under-Relying**

Thus far we have talked about cases where we are over-trusting or over-relying on technology. By contrast, we can also have cases where perceived accuracy is actually below the real accuracy, i.e., algorithms that work *better* than we believe them to. We briefly consider this case here. Examples include cases of "automation aversion", where an algorithm is judged

harshly purely in virtue of being an algorithm, but also more subtle cases, for example a case where my assessment of the accuracy of an algorithm is methodologically flawed in some way that negatively biases it (e.g., overweighting failures) (Dietvorst, Simmons, and Massey 2015).

As far as vulnerability to harm facilitated by predictive algorithms goes, such cases might be the least problematic. In the bottom-right region of Figure 1, accuracy is actually high, but it's not perceived that way. What this typically results in is a maintenance of the status quo: things remain how they were before the algorithm was developed. Of course, this might carry some ethical problems, particularly when the status quo itself includes systemic problems such as structural discrimination. A reflection essential for any algorithm development is to understand in which ways it is improving the status quo. Who are all the involved stakeholders? Who is it benefitting? Who does it have the power to oppress? What are the tradeoffs and who is incurring the associated costs? In essence, would this algorithm really improve society, providing people with the ability to flourish? This is a crucial question, to be fully explored by any new venture in AI, but in terms of harms that come from predictive algorithms and violations of privacy, it is out of our current scope. Moreover, if the algorithm is being used but "cautiously", i.e., as if it is less accurate than it really is, then the analysis presented in Chapter 1 remains most relevant here, since it frames the *possible* harms and thus can serve as a guide to what "cautious" use ought to look like, to avoid or mitigate those harms.

As we move left in the bottom region, cases become less and less impactful since the algorithms involved are neither very accurate nor perceived as such; this would include frivolous applications (online tests to gauge what Tolkien character you are, or what type of soda captures your personality, etc.) as well as less trivial situations (e.g., predicting your political affiliation based on your profile picture) which are nonetheless not taken seriously (currently). The ethical questions related to these revolve around proper data collection and data sharing to avoid unnecessary and unwarranted surveillance; this is a privacy-related topic but again, not the issue at hand.

Section 4 | **Calibration**

Finally, we examine the case of when the deployment of an algorithm matches its real accuracy, i.e., when our beliefs about the ability of the algorithm to fulfill its intended use are properly calibrated to what the algorithm is truly capable of doing. As mentioned, in a sense this is the strongest case for proponents of predictive algorithms, particularly from the perspective of developers and deployers—by definition, the algorithm is functioning exactly at the level it is expected to, so there is no obvious sense in which anyone is being misled. Our analysis will raise once again several open questions considered in previous sections. *Should* we deploy these algorithms? What warrants their use, particularly in cases where they may cause harm? And how can we find paths to mitigate presentational harms, specifically?

Of course, as mentioned, the whole of Chapter 1 focused exclusively on the top-right corner of the graph, which is considered by most developers and those who deploy predictive algorithms to be the ideal—we discussed at length the ethical problems entailed by the blurring of the distinction between guessing and observing entailed by such highly accurate algorithms. As we move diagonally down the line the accuracy falls, but we are still considering cases in which, by assumption, there is a reasonable understanding of the tech and its limitations.[43] We have already discussed the negative ramifications of being over and under the line (i.e., over- and under-relying on the tech), and we have argued that one of the core privacy-related problems with high perceived accuracy is that it brings along many presentational harms. The present analysis of the "calibrated" cases will dive deeper into exploring such harms.

First, it is helpful to highlight and then bracket a potential worry about the practical usage of predictive algorithms. In many cases a partially accurate algorithm (that is correctly perceived as

[43] One subtlety: what if I'm calibrated in terms of overall accuracy, but not in terms of *where* the algorithm is inaccurate? For example, what if I "know" that the algorithm is 75% accurate, but I think that the 25% errors are all in women when they are actually all in men? In a certain sense one might say I'm calibrated, but in another sense I'm really not. This taps into the difficulties of using a one-dimensional scale (i.e., percentage) to capture all the subtleties of "accuracy", when in reality the notion is of course more complex. For our purposes, we take true calibration to rule out cases like these, where the person is systematically misinformed about the nature of the errors.

such) is still much, much better at making predictions that other existing methods. Because of this the algorithm may be widely relied upon—it is, after all, the best we've got. It is debatable whether this should count as a case of "over-relying" on the algorithm. By assumption, we are considering cases where the assessment of the algorithm's accuracy (or lack thereof) is on the mark. Nonetheless, it may be deployed extensively, being the best tool available. For this reason, we may find that many of the considerations raised above for "over the line" cases, specifically as pertaining to presentational harms, apply here as well, due to the mismatch between the algorithm's actual accuracy and its wide deployment (i.e., it is "over-deployed"). In what follows, however, we bracket this issue in order to further zero in on the "strongest" pro-deployment cases: that is, where as much as possible the perception *and deployment* of the algorithm is "responsible" in the sense of aligning with its true capacity for prediction.

4.1. When are presentational harms warranted?

To address these "ideal" cases of predictive algorithm use, we must face a difficult question head on: *are all presentational harms wrong?* In other words, are some presentational harms justified? After all, we *have* to make choices constantly to decide how to allocate scarce resources and opportunities, many of which rely on informed guesses; in at least some such cases, algorithmic assistance will allow us to make better decisions than we would otherwise.[44]

While some presentational harms might indeed be justified, we will argue that matching perceived accuracy with real accuracy is necessary *but not sufficient* for such justification. The arguments we have explored thus far show that it is irresponsible to deploy an algorithm with an incorrect perception of its accuracy. This is because they subject individuals to presentational harms that are certainly *not* justified, since they amount to decisions made under false premises. Some presentational harms correspond to false beliefs about an individual (incorrect predictions by the algorithm), while others can come from true beliefs, in which case they are a

[44] In some cases, this might simply be due to the raw speed of computation: we might lack the personnel or resources to scan through massive amounts of applicants/candidates/suspects/patients/etc. In other cases, it may stem more from the particular way the algorithm processes information and encodes correlations, etc.

mix of presentational and informational harm. Examples of the latter include discrimination based on gender, sexual orientation, nationality, or the color of our skin. This is to say: in our society we have already concluded that discriminating against individuals based on certain *protected characteristics* is wrong, so for the same reason these presentational harms are also, of course, unjustified—no matter how much alignment there is between the perception of the algorithm's accuracy and its true accuracy.

But not all cases of interest involve such protected characteristics, nor are so morally transparent. Let's consider a hiring algorithm, as discussed in Section 1. What happens if such an algorithm, in a partially accurate way, infers that you are not well suited for a specific job? Suppose it gauges with 80% confidence that your skills are not sufficient for the job at hand. One can argue that there is absolutely nothing wrong with this: after all, the company *must* pass over *some* prospective employees, and even without any technological support, the hiring committee will have to make inferences based on their available information (the personality type of the applicant, their resume, how the interview went, etc.). How could it be wrong, then, to instead deploy an algorithm that can assess the candidate along similar lines?

Part of the answer depends on what information they are gauging and how they get it. Are they scraping the social media accounts of potential employees or just analyzing their resumes?[45] One of the subtle dangers in (even partially) replacing human decision making with algorithms is a potential loss of norms and moral (and legal) constraints that had previously been in place. We would not find it acceptable, for instance (excepting certain specialized contexts, like high stakes government jobs), for an interviewer to assemble a team of experts to comb through an applicant's complete social media presence on every platform going back for decades. This is an example of a norm that governs a certain kind of human decision context. It's not merely that this example is silly on account of the ludicrous amount of human resources it would eat up; rather, we find there is something fundamentally inappropriate about being

[45] There is an argument here for Nisenbaum's work regarding "contextual integrity" in which the information flow matters, that is, it matters how the information was obtained, and what counts as an appropriate flow of information varies depending on the specific context (Nissenbaum 2009).

subject to such an extreme level of scrutiny in the context of applying for a typical job. But if a predictive algorithm is capable of simulating this level of scrutiny—scraping through the internet and assembling a reasonable (partially accurate) approximation of an applicant's social media profile—then it is effectively replicating (with some error) the very thing we do not accept from human decision makers; in the case above, violating the applicant's privacy.

We should resist an erosion of norms that is due simply to the nature of certain kinds of computation, namely, the fact that predictive algorithms can far exceed the capacity of human decision makers to reconstruct private information from data that is freely available (even if this is done with only partial accuracy). Said more plainly: without the use of algorithms, human decision making is replete with standards (legal and moral) regarding what information is appropriate to consider in making a given decision. But algorithmic intervention often makes it much harder to tell what information is effectively being used (either through direct access or machine-powered inference), and so difficult to translate and apply analogous standards. In fact, there is a culture of "more data is better" at play which creates a further disconnect between human vs. algorithmic decision-making: what would be unacceptable for a human to consider becomes not only acceptable but essential or laudable for an algorithm.

So, the first point to make here is that algorithm use should be governed by standards regarding what information is appropriate to access/infer that are analogous to the standards we would want applied to human decision makers in similar contexts. Of course, these standards will vary in different situations, e.g., hiring versus medical diagnoses. This may seem like an obvious point, but it is complicated in practice by the fact that it can be hard to get a clear grasp on *exactly what information* a given algorithm is able to access and infer. At first glance, this might seem like an argument in favor of data control, since the moral problem under consideration here is based on what data is accessed. But we do not suggest data control as a remedy because, as we have discussed in Chapter 1, it is a practical impossibility in many cases. Rather, a form of *use control* is needed here: we should aim to restrict deployment (use) of algorithms that access (directly or indirectly) certain kinds of information that we consider inappropriate for the decision at hand. Indeed, such use control, properly understood, can help

incentivize the development of algorithms that make it easier to assess what data or information is being (effectively) accessed and what we consider appropriate.

For a simple example: no human is going to pore through the last 10 years of an applicant's tweets, retweets, and likes to build a psychological profile of said individual, even though this information is, technically, freely available. But a machine could very easily integrate all this public information into their decision making, and more (data on past purchases, locations, comments, etc.). The social norms that help draw the line between our public vs. private lives were developed in the context of *human* capabilities. So, at a high level here, we are grappling with the same question we have visited before, namely, how must these norms be adapted in the age of algorithms? To the extent that we want to preserve a similar boundary between public and private lives, we must factor in the new ways that ML algorithms can use data.

A more nuanced understanding of these considerations comes from analyzing the way such algorithms are actually *used*. Recalling the distinction introduced in Chapter 1 between the *question-answering* sense of use and the *action-guiding* sense of use, we focus on the former as it is most relevant to unpacking the high-level concerns raised above, as well as identifying and mitigating the associated harms.

4.2. Question-answering and ethical proxies

In the example presented, at a first pass, the use of a hiring algorithm may seem intuitively acceptable when the company is trying to make an educated guess about "appropriate" characteristics for prospective employees like their *competency* for the job: do they have the right experience, the right skills, etc.? But even in this case, it is crucial to understand *how* the algorithm assesses competency (this is the question-answering sense of use).[46] One might reasonably object, for example, if the notion of "competency" is based on proxies that even if

[46] One obvious concern here is whether the proxies chosen actually track what we think they are (E.g., does incarceration really track criminality, etc.). I discuss this issue in chapter 1 (Section 3.1.1.2. Measures and proxies), and at any rate, in this section we are explicitly considering situations where the algorithm is indeed partially accurate and correctly perceived as such. Therefore, we bracket such concerns here.

ultimately "correct" (they are at least partly true indicators of the desired trait), we consider not appropriate to include, such as who might be family oriented, or who might be pregnant or soon to be pregnant. In this case, although we may think "competency assessment" broadly speaking is an acceptable use of a hiring algorithm, whether any particular such algorithm is truly justified will depend on what proxies are considered appropriate to answer the question of "competency".

In reflecting on this, we must take into account both what kind of *data* or information is unacceptable and what kind of *practices* are unacceptable; these proxies should both not infringe in the individual's privacy and not erode our usual understanding of reasonable or appropriate information to ponder for a given choice. Cathy O'Neil (2016)[47] gives us a real-life example of the perils of using inappropriate proxies[48] for 'competency'. She reports on how it has become common practice in several companies (including Kroger, Finish Line, Home Depot, Lowe's, PetSmart, Walgreen Co., Yum Brands, among others) to use *personality tests* such as variations of the Five Factor Model as an automated first screening for possible job applicants (even for part-time minimum-wage jobs) as a way to identify possible mental illness.[49] These personality tests have become common practice and are used to exclude as many applicants as possible, as cheaply as possible (ibid, p. 109).[50] The result is that educated people that may be completely competent for the job are being passed over in many applications because of some

[47] (O'Neil 2016) *Weapons of math destruction*. Chapter 6: Ineligible to serve.

[48] In Chapter 1 we already talked about the complexities of picking the right proxies. Another similar complexity has to do with the benchmarks used to create and train the model. How is it assessed? Compared to what? Oftentimes the benchmarks are initially created to measure a very specific skill but have been confused as measurements of more general abilities (Raji et al. 2021).

[49] These companies hire third parties such as Kronos, a workforce management company, to run this part of their hiring process.

[50] It is still not settled whether this personality tests are good predictors of performance, with some research suggesting it is not (ibid, p. 103). But even assuming they perform above average (and better than not implementing the tests at all), there are still moral problems around them. things like reference checks are more effective, as well as some kinds of cognitive exams.

red flag inferred from the questions presented in the personality test indicating mental illness[51] (that can be anything from anxiety or tendencies towards depression to bigger issues that are not fully captured by the test). It might be an incorrect inference, or only partially accurate, but even if accurate, this kind of exclusionary practice is highly problematic: "mental illness" is a wide and complex spectrum that by no means entails that an individual should be excluded from having a job. Notice that there is a case here for using personality tests as a way to screen people by personality traits that might be desirable (e.g., being a team player, being a leader, etc.) as a way to find tiebreakers between equally qualified applicants. After all, companies can expect to receive applications from multiple qualified applicants and some screening for compatibility with the company's ethos is good. Ideally interviews are where personality traits are assessed, but it makes sense for jobs with high numbers of applicants to use other kind of screening methods. In this case, the question is whether these personality tests are an adequate way of identifying those traits (i.e., whether they use the right questions and right proxies). Moreover, it is important that the tests are chosen carefully to be ones that measure only appropriate personality traits and are not an excuse to uncover mental or physical health diagnoses. "Having dealt with depression" is not an acceptable reason to pass someone over for a job (even in a "tiebreaking" situation), and it should not implicitly become one due to the deployment of these sorts of algorithmic assessments.

This is just one concrete example, but it shows how, like many other "Big Data" programs, since they can't directly measure the key variable ("competence" for a job in this case), they settle for proxies—proxies which are often not only inexact but also *unfair*. In this case, our right to keep private our mental (and physical) health is at stake. Since (in most cases) we do not consider it appropriate to be asked in an interview about our mental health history ("have you ever seen a psychiatrist? Why?") and whether we have battled with any mental illnesses, nor required to disclose any health problems such as propensity to have cancer in the future, algorithmic methods should not offer a roundabout way to get away with asking these questions

[51] What makes things worse is that the process is opaque both in that the developers and deployers of these tests are not sure what patterns of answers disqualify the applicants, and because the applicant might never find out why they were disqualified—they are not notified that it was because of their personality test results (ibid, p. 110).

and basing decisions on this information, even if technically, the connection between the proxy and the initial variable is correct and knowing if you will or have had cancer can be a good partial indicator of job competency. We do not generally think that job applicants ought to be disqualified on the basis of being diagnosed with depression, or anxiety, or conditions such as bipolar disorders.[52] If companies *were* allowed to use this information against prospective employees, it would (among other things) create an incentive to never seek treatment in the first place, so that we don't have "therapy" in our records. Diagnoses for both mental and physical health are kept private under HIPPA (in the US) precisely to protect individuals against such discrimination. Similarly, we have an obligation to develop and deploy algorithms that do not rely on that private information. The mere fact that it's not a human doing the direct "snooping" does not warrant the behavior, nor avoid the negative consequences.

Figuring out what information is "appropriate" in this sense is a challenge that will vary from case to case. Nisenbaum's "Contextual Integrity" approach can give us a framework for understanding what the standards are that we want to preserve, and thereby help us avoid the potential erosion of norms that comes with the use of predictive algorithms. In the example here presented, this would mean using other better proxies for job competency, such as reference checks or even some types of cognitive exams or work-related tests.

Section 5 | **Concluding Remarks**

One might argue that the status quo (i.e., human decision makers) is already problematic filled with biases and snap judgements. Therefore, one might claim, even if algorithmically assisted decisions are problematic in some of the same ways, as long as they represent an *improvement* over old practices, we should embrace them with open arms. Indeed, proponents

[52] In her chapter, O'Neil mentions the concrete example of a young man that had had treatment for bipolar disorder and perhaps as a result, kept being red-lighted in every application that required personality tests.

of many decision-assisting algorithms will typically argue not only that these algorithms are efficient and cut costs, but that humans are *more* biased in their decision making.

But this is both too optimistic and too pessimistic. It is too optimistic in the sense that it assumes that deploying new technology that is *currently* an improvement (in accuracy) over the status quo will not lead, ultimately, to something much worse than the status quo. Technology is disruptive and often unpredictable in its consequences. Allowing algorithms to encode ethically nebulous deliberations is dangerous *precisely* because automation has the potential to multiply the associated harms, sometimes in unforeseen ways. One example of this is by "ossifying" a decision process in ways that make it hard for any future improvement of the process.[53]

This is a problem present in the hiring example given above. These hiring programs can't incorporate information about how the candidate would actually perform at the company, since that's in the future, and therefore unknown. But because of the way this algorithm is used, there is no way to know if it made a mistake and passed on what would have been a stellar employee. The algorithm is not updating its analysis based on this kind of feedback (e.g., realizing when it missed out on good hires), and therefore won't be able to correct inappropriate or unjust parameters (O'Neil 2016, p. 111). This situation is worse than the status quo insofar as we lose track of the problem at hand. With human decision-making there is a way to go back to the human, understand the reasoning, and educate on the biases, since we are constantly updating our beliefs about the world. But with these types of algorithms, we might be hard-coding bad practices in a way that is harder to correct. Moreover, the more opaque the algorithms become, the less chance there is that we will realize if it is using discriminatory or inappropriate parameters. There is no way to solve a problem that we are not aware of, which is especially problematic when algorithms are deployed in high stakes situations, like granting access to goods and services, especially when they are essential to live a flourishing life—in our current society,

[53] A real-life example of how decision processes can ossify is in the field of law. Consider the case of information and data privacy law, which began in the 80's and was heavily influenced by Richard Posner and the Chicago School of Economics into adopting an extremely laissez-faire approach; this then solidified into the present status quo, in a way that has made it incredibly hard to change privacy protections to accommodate the development of technology through the last three decades (Solove 2006).

being able to work is a basic requirement for providing for ourselves the basic human capabilities (capacities to function)[54] to stay alive and, hopefully, flourish.

With all that being said, it's important to recognize that aiming shortsightedly for minor improvements to the status quo is also too *pessimistic* in an orthogonal sense. The widespread implementation of algorithmically assisted decision making represents a paradigm shift in many aspects of human enterprise and is thus the ideal opportunity to address and improve systemic flaws, including those that have deep historical roots. It is incredibly shortsighted to waste this opportunity simply because such flaws already exist and thus are, in a sense, "inherited" to our sense of what is acceptable and what is not.

The example of the hiring algorithm is a grim one in part because it represents the sad reality that for most people, getting a job is essential for acquiring any of the central functioning capabilities that are necessary to live a decent life. It is a sad reality that nothing is guaranteed for us in many societies unless we can obtain them (pay for them) for ourselves. This can include shelter, access to healthcare, to education, to food, among others. How much of these needs should be an individual's responsibility to guarantee for themselves and their families, and how much should we entertain the prospect of restructuring society in such a way that our basic needs for a flourishing life do not depend on our personal income. Technology and predictive algorithms could have a role to play here, not by making slights improvements over the status quo (making job screening more efficient and cheaper, for example) and thereby further reinforcing the current social organization, but by helping restructure societies in ways that are actually more beneficial to all of us. Of course, many changes are socio-political and not technological in nature. But technology can either serve to ossify old standards and limit new ways of approaching social

[54] The 'Capabilities Approach' by Martha Nussbaum and Amartya Sen nicely frames the distribution of liberties and resources that each person requires to be able to see they are ends in themselves and not means to someone else's ends. These basic capabilities are: Life; Bodily health; Bodily integrity; Sense, imagination and thought; Emotions; Practical reason; Affiliation; Relation to other species; Play; and Control over one's political and material environment. Given the options, each individual can choose for themselves to function in the way they see fit, where functioning is the actual exercise of a capability, given the internal capability and the externally appropriate environment to realize it. (Nussbaum 1999)

problems, or it can shake up the status quo and make new approaches feasible and salient. We ought to work towards algorithms that preserve the best of current social practices and combat erosion of care—even if that erosion was already in progress—and furthermore, that are designed with an understanding of what we need to change in society to make it more ethical. Going back to our example, we could perhaps consider the idea of a hiring algorithm that is designed not only to identify the person who would do the job best, but also balance that with the person who would benefit most from *getting* the job.

As we can see, the considerations involved in the reflections above move us from the "question-answering sense" of algorithm use, where we focus on what is the appropriate information that goes into the algorithm, to the "action-guiding" sense of use, when we assess the social contexts in which we will deploy the algorithms and how it will affect peoples' lives. As always, the questions correspond to: "Who are the stakeholders here and what do they value?", "Who is the algorithm benefitting" and "Who does it has the power to oppress?"

To summarize: the use of algorithms (even partially accurate algorithms that we correctly assess as such) ought to be justified in two broad senses: (1) is it using (directly or indirectly) information relevant to the use-case that we deem acceptable, and (2) is it guiding actions that do not go against the individual's ability to enact their capacities to function in society. (And, if we are feeling optimistic: (3) is it being used to genuinely and substantially improve peoples' lives?) The ultimately goal is to approach a society where everyone can pursue a life plan in accordance with their own values and goals, and not just to "improve on the status quo" which usually refers to improving efficiency and cutting costs. The more important task is to know if the development of the algorithm *is ethical*.

Lastly, moving up a level, we have seen in this chapter that presentational harms that involve "people harming me because of their beliefs about me" can be reduced/mitigated in two key ways:

Improve calibration: establishing alignment between perceived and real accuracy, i.e., changing people's beliefs (perceived accuracy); and

Implement *use control* in two senses:

Question-answering sense: proper model, proper training (right proxies and acceptable/appropriate proxies) and appropriate data used.

Action-guiding sense: stopping the harmful behavior by making sure the output given by the algorithm is used in ways that benefit society (help individuals exercise the capabilities they chose to).

These solutions require collaboration between disciplines. Improving calibration means to bring everyone—developers, deployers, and citizens—to a common understanding of the accuracy of the algorithm. This in turn involves providing better tech literacy so that people implementing, relying, or being subjected to the algorithm understand its function better. Use control likewise requires an interdisciplinary approach, connecting more specialized understandings of the algorithmic process to its world implications. For the question-answering sense, there must be a clear understanding of what information it is using, whether it is relevant and obtained without violations of privacy, and whether it is appropriate in the first place to consider such information in answering the given question. For the action-guiding sense, we ask in what context do we want to use the output of this algorithm? What services or goods are granted or denied based on this output? Is it an ethical use? And what was the possibility space for computational intervention to begin with?

CHAPTER 3

Consent in The Digital Age

INTRODUCTION

Suppose you have concerns about Google's collection and use of your information: you avoid their search engine, eschew their office tools, and even steer clear of YouTube. This may already seem like a monumental reorganization of your daily life, but now imagine you are a prospective student just offered admission to the university of your dreams, where the institutional email service is run by G Suite, and it is neither possible nor desirable to opt out. In situations like these, where our data is collected by entities whose interests do not necessarily align with our own, or where we can't really understand what we are consenting to (i.e., what is being collected and how it will be used), and where the prospects of "opting out" are murky at best, we must ask: what is the meaning of *consent*?

As we become more and more intertwined with technology that shapes and defines our private and social lives, so too do the conditions for our well-being. In today's societies, our interactions with technology mediate and are mediated by our relationships with the private corporations that own, develop, and deploy them. Given our discussion in previous chapters of privacy and the harms that can come through violations of privacy in digital contexts, it will come as no surprise that to understand the conditions of human flourishing in this new digital era, we have to revisit a key concept: consent.

We have seen that privacy is important to safeguard people's ability to pursue their own interests and be able to flourish. Private corporations, on the other hand, are entities that represent interests that often conflict with individual interests. Often, one of the primary goals of any corporation is the success of the company, represented by increasing profits, market share, and/or stock value, liquidity and equity. Of course, to achieve such goals they often offer products

and services that people actually want and will pay for, and for this reason there tends to be some overlap or compatibility of these interests to the degree that, broadly speaking, businesses focus on delivering service and products which provide convenience to their customers.[55] But companies' interests are (almost by definition) not perfectly aligned with our interests, so we have good reason to worry that our data will be used in ways that don't promote our interests, if companies are left unconstrained.

Taking advantage of the lag in proper regulations, technology has developed in ways that have significantly increased the power that companies have over users/customers, creating an unprecedented asymmetry between the two. A major driving force behind this new asymmetry is the fact that, aside from whatever specific service or product is on offer, corporations can also access, collect, correlate, and process incredible amounts of information that their users share. Regulations on data ownership, collection aggregation and usage have been relatively lax and corporations have taken full advantage of this since the advent of the internet and tracking/monitoring technologies (Robertson 2019).

In previous chapters we have focused on the *depth* of reach we get from the data that is collected via machine learning and predictive algorithms, but it's also important to note the *breadth* of information gathering that has led to those better predictions. The business models of many corporations have shifted more and more to focus on data collecting and information processing. For example, it is standard nowadays for businesses to employ targeted advertisement, and the dynamics of targeted advertisements have given rise to a whole new industry trading in user information, collected via cookies and other identifiers, along with a constant exchange of this tracked information via real-time bidding, cookies synching, data brokers, etc. (Cyphers and Gebhart 2019). For most of its existence, data brokerage has been a highly unregulated industry—regulations such as CCPA/CPRA in California or DMA in the EU being recent attempts to change this. The industry focuses on collecting, processing, and reselling huge

[55] After all, many of the technologies developed find ways to offer more convenience in our day to day lives. The catch is that in return they rely on a high degree of data collection and information processing, most of which goes beyond the requirements of the technology to work (it is also often sold to third parties for profit) (Robertson 2019).

amounts of information about all online users. Of course, the end goal is not only targeted advertisement, though that's the only one users might be aware of. Besides targeted advertising (Boerman et al. 2017; Ur et al. 2012), other uses of that information include selling it to political campaigns and interest groups (Chester and Montgomery 2017; Cadwalladr and Graham-Harrison 2018); to debt collectors (Basak 2022), and bounty hunters (Cox 2019); to cities, law enforcement agencies (Bradford Franklin et al. 2021), intelligence agencies (Bradfort Franklin and Thakur 2021), and others like insurance companies, lenders, hiring services, etc. ("Data Brokers" 2023).

Though our main focus here is not to exhaustively catalogue the vast array of tools related to monitoring technology, to set the stage a bit it will be useful to at least briefly outline the basic mechanisms of corporate monitoring in the digital age (see Cyphers and Gebhart 2019 for an excellent overview). It is founded in identifiers on the web, on mobile devices, and in the real world that are shared automatically with companies and linked with each other to form growing "shadow profiles" of users over time. Identifiers on the web include: cookies, IP addresses, TLS states, local storage super cookies, browser fingerprints; identifiers on mobile devices refer to: phone numbers, hardware identifiers (IMSI and IMEI number), advertising IDs, MAC addresses; finally, some real-world identifiers include: license plates, face prints (face biometrics) and credit card numbers.

These identifiers are used to create a tracking network that combines in-software tracking (in websites and apps) and passive, real-world tracking. Tracking in-software includes practices such as: analytics and tracking pixels, embedded media players tracking, social media widgets, CAPTCHAs, session replay services, and the massive ad networks (which involve activities like real-time bidding, where large amounts of information stored in cookies are shared with multiple third parties and data brokers, without the knowledge of the user). Passive, real-world tracking includes: WiFi hotspots and wireless beacons tracking, vehicle tracking and ALPRs, face recognition cameras, payment processors and financial technology. The average user has no idea most of this is happening and has very little understanding of what information is being collected, by whom, who it is being shared with, and how it might be used, now or in the future (Cyphers and Gebhart 2019).

This ubiquitous monitoring represents a pervasive violation of privacy. What could warrant such infringements, ethically? *Consent* is a morally transforming concept here (Dworkin 1988; Beauchamp and Childress 2001; Archard 2008; Miller and Wertheimer 2010). In a nutshell, what it means for an agent to consent to a proposal is for them to voluntarily express, in some form, their endorsement or approval of it. Of course, there are many subtleties here: if this expression is somehow forced or tricked out of them, or left tacit, that can change the nature of the act—we discuss this more below. It is easy to see the transformative power of consent in the realm of *intimacy*, where for instance it is the difference between spying on someone while they are changing and being invited to see them naked, or between rape and consensual sex. The morally transformative power of consent can also be found in other areas for which privacy is important (see Chapter 1, Section 1.4 "The importance of privacy"): for human dignity, for developing interpersonal relationships, to control access others have to us, to enhance personal expression and choice, to have control over information about ourselves, and to have a sphere free of interference by others.

The goal of this chapter is to examine the concept of consent in light of the analysis of privacy in the digital age we have laid out in previous chapters. One might understand consent to be morally absolving, so that cases involving consent necessarily do not constitute violations of privacy; in this case, one would conclude that individual consent is all that is needed to correct the various harms that we have extensively discussed in this dissertation. Unfortunately, and crucially, even if we were to adopt such a position, *individual consent itself is not truly feasible given the digital landscape we operate within*. As I'll show, part of the problem has to do with simple lack of awareness, but perhaps more importantly, "consenting" to relinquish ownership of your data leaves one vulnerable to a host of potential future abuses that, arguably, undermines the idea of giving "consent" in this context altogether.

The natural "fix" here is to switch to a model of *use control* along the same lines as introduced in previous chapters. But even here, many of the problems of consent persist; indeed, I will argue that any reasonable conception of use control cannot (solely) rely on individual consent as a way to warrant privacy infringements. To set the stage, I dedicate the first half of this chapter to laying

out the terrain that makes individual consent—understood as *informed*, *uncoerced* consent — infeasible. To establish this, I will talk about (1) the user's lack of awareness of the data collection; (2) the user's lack of know-how to navigate the privacy policies and/or; (3) the lack of meaningful understanding of under-constrained future uses; (4) and the lack of feasible alternatives. In the second half of this chapter, I will then explore possible alternatives to individual consent, informed by the *use control* framework we have argued for.

Section 1 | **The Problems with Individual Consent**

In recent years, some prominent thinkers in the ethics of technology have argued against the idea that individual consent is the panacea that current legislation makes it out to be (Barocas and Nissenbaum 2014; Solove 2013). The United States has largely embraced a "notice and choice" approach in which users who don't opt out count as implicitly consenting to privacy policies. Meanwhile, the EU with GDPR has an "express consent" approach that focuses on actively opting *in* (also known as "affirmative" consent) (Solove 2013; Schwartz and Solove 2014). In being explicit rather than tacit, the EU approach presumably captures a more robust form of individual consent; however, as it turns out, both arguably fall short of a meaningful notion of informed, uncoerced consent. This general criticism has been articulated by various authors (see, e.g., (Raymond 2017; Berinato 2018; Diehm et al. 2021); here, I will present my own vision of these arguments, framing the issue using the canonical interpterion of "informed and uncoerced" consent drawn from bioethics. Applied in the modern technological context, I identify several necessary conditions for genuine individual consent: basic awareness, relevant know-how, "sufficient" understanding, and availability of feasible alternatives; each of these speaks to either the "informed" or the "uncoerced" aspect of consent in their own way.

The arguments and examples that follow will not be completely surprising to anyone who is familiar with the modern digital architecture that has come to dominate so much of our lives. It's easy to notice that there are many acts of "consent" that are quite superficial, and as such fall short of what is required to safeguard privacy. For instance, clicking the "I agree" button, when

presented with a hundred-page privacy policy before installing a new app, is a very poor interpretation of "consent", particularly when the user has no real understanding of what they are agreeing to. Obscure language and vaguely defined terms and usages are rampant in the privacy policies of many corporations. This is clearly a type of uninformed (or poorly informed) consent. Similarly, "consenting" to a privacy policy for a service I cannot live (or reasonably conduct my life) without is hardly an example of consent freely given. How many of us could forgo having an email account, nowadays? Is there a meaningful difference between the privacy policies of, e.g., Gmail and Yahoo, or any other provider?

But then what is it, precisely, to give "consent"? This notion is particularly relevant in bioethics and especially the discussion of consent in recent clinical practice (Beauchamp and Childress 2001). The medical sphere moved from a paternalistic view of "doctor knows best" to a requirement of voluntary, informed consent by the patient. In medical research, the Nuremberg Code was influential (a result of the cruelty of Nazi experiments), and now asserts that "The voluntary consent of the human subject is absolutely essential". Moreover, regarding privacy specifically, health-related data and information is considered to be among the most vulnerable and exploitable type of information one can share, so it requires proper safeguards to assure it won't be used against the patient (for example, for hiring discrimination, or by insurances companies, loan agencies, etc.).

I will focus on this idea of *informed*, *uncoerced* consent from a competent person to motivate an analysis that is directly and systematically applicable in technological contexts. Specifically, I will connect the "uncoerced" requirement to the availability of feasible alternatives (examined in Section 2.4.1) and effectively split the "informed" requirement into 3 related but distinct notions: awareness (Section 2.1.1) and know-how (Section 2.2) and understanding (Section 2.3).

1.1 . INFORMED CONSENT

Three main "categories" of problems arise when we consider the feasibility of an individual giving informed consent in digital/technological contexts. First is a basic lack of awareness of the fact that they are being monitored and tracked in the first place. Second is when the relevant

know-how required to navigate the options that might be available to opt out of such tracking is not present. The third problem corresponds to the impossibility of meaningful understanding given the unconstrained and highly uncertain possible uses of the data that is collected. I will argue that a two-pronged solution is needed here: the first two kinds of problems can only be solved by moving away from individualized notions of consent and towards stronger social governance and regulations. The third kind can be solved or at least ameliorated by a switch to use control. Let's start with the most basic violation of privacy first.

1.1.1. Lack of awareness

It might sound simple, but more often than not, users have no awareness of what sort of policy they have implicitly agreed to when using the products or services of a company (Atske 2019; Larsson et al. 2021; Spadafora 2021; Turow et al. 2023). The identifiers and tracking networks currently used are completely obscure for the average user, who is often only vaguely aware that cookies are somehow being used to offer them personalized ads. And even this vague awareness is relatively recent, with the implementation of the GDPR since 2018 requiring websites to inform visitors of the collection of cookies, and ask for their "consent", by hitting an accept button (the opting in approach).

In its simplest form, this lack of awareness can be expressed in the first-person form: "I have no idea something is happening." In essence, the average user has no idea they are being surveilled, or to what degree. This of course means they don't even know whether their privacy has been infringed upon. This lack of knowledge creates an immediate and stark asymmetry in power between users and corporations. To understand this more fully, think of the relationship between users and corporation as a flow (see image below), where users experience an activity provided by the company (e.g., browsing YouTube, playing a game, using a program, visiting a store, etc.). When the user engages in the given activity provided by the company, raw data starts being collected (this can be behavioral data on the website, items added to a cart, links clicked on, time spent on each link, location tracking in the real world like navigating a store, etc.). Some of this data is collected "in an anonymous way", which means it is not tied to

personally identifiable information (PII) that includes names, addresses, zip codes, and phone numbers, among others. But the data is tied to one or many of the identifiers mentioned in the previous section. This data can then be stored and often sold to third-party companies or processed into information (meaningful data) and then stored or sold to third parties. Information processing involves aggregating and cross-referencing data, and more often than not, using advanced algorithms to make predictions from those data sets. (During these processes, the anonymization can in many cases effectively be overridden—also called de-anonymization or re-identification). Finally, the companies put this collected data and information to use. The uses can be varied and widespread. The same data can be used to improve their product or service, can be sold for profit, can be used to start a new project, and so on. Thus, the information created will in the end be used in ways that affect the user, creating a feedback loop.

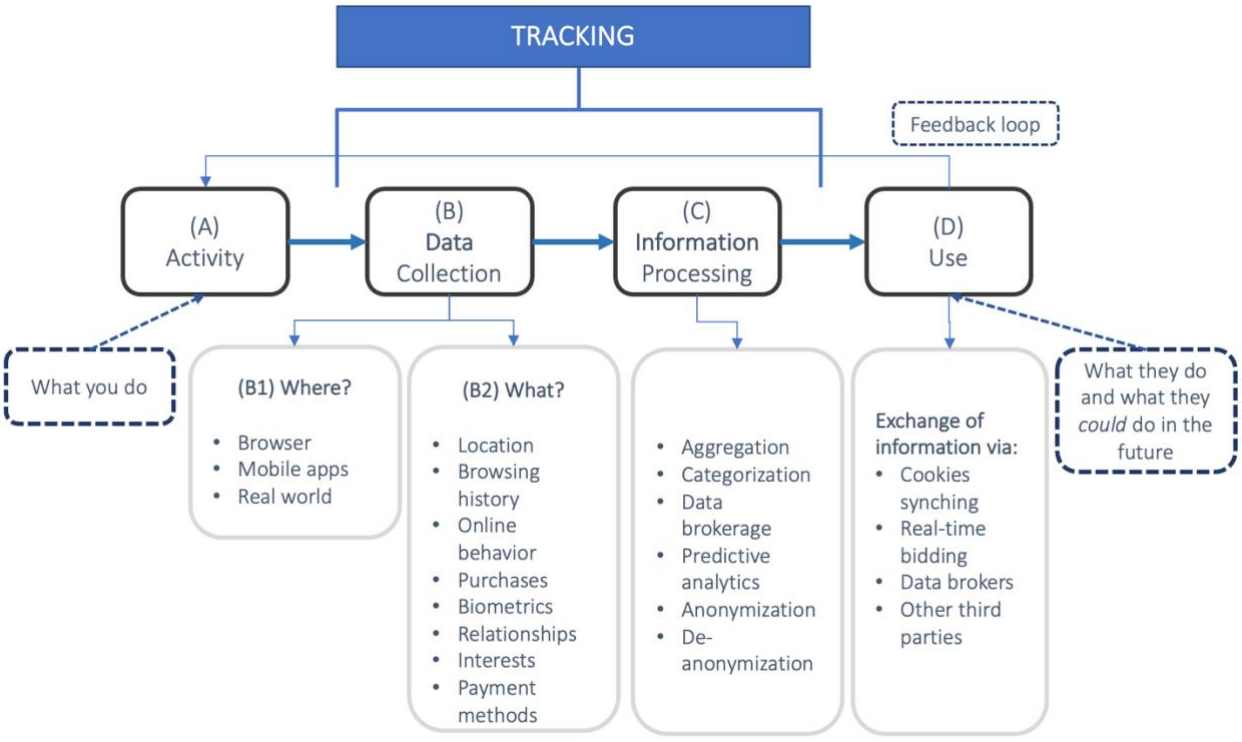

Figure 2

This is a very short, simplified overview, but the idea is to draw attention to the many stages of the dynamic interactions between corporations and users, each of which the user might have different levels of awareness of. In this first kind of problem for individual consent, the user is only truly aware of the first step: only aware of the *activity* they are engaging in, with no real knowledge of the broad scope of the data that is being collected, processed, and used in various ways (let alone how it might affect them in the immediate or distant future).

Examples are easy to come by. One potent example is the widespread deployment of facial recognition (FR) technology, including in the private sector, with some stores using it to identify customers inside their establishments. People are not aware of this happening and therefore have no way to fight back against pervasive tracking and profiling via facial recognition. Some states have started to bring regulation against the use of this technology, but FR is by no means the only way of tracking the location of users in a store or building. Many companies also record other identifiers such as MAC Addresses (used to set up the connection between two wireless-capable devices over WiFi or Bluetooth), which are captured by WiFi hotspot in the area or even by wireless beacons set up by the companies themselves, silently picking up MAC addresses in their vicinity. Once again, customers are not informed of this location tracking taking place, much less of the use and pervasiveness of the practice (to protect customers, recent versions of iOS and Android use MAC randomization, which hints at the importance of good governance, in the privacy-by-design mentality).

One can see that in many cases, it is more convenient for the company not to inform the customers of the tracking that is taking place, so they can maintain control over the collection and use of the data unhindered. Clearly these cases involve violations of privacy. Lack of awareness leaves the individual completely at the mercy of the entity infringing on their privacy.

Following our focus on the uses of the technology, the most obvious “fix” for this situation would be to require the proper disclosure to individuals of the use of these technologies and

the purposes of their use (GDPR's "affirmative choice" approach tries to tackle exactly this). But, as we will see, even if efforts to achieve this are necessary, they are far from sufficient.

1.1.2. Lack of know-how

Whereas we described the first type of problem with the expression "I have no idea something is happening", this second type could be expressed as "I know something is happening, but I don't know how to stop it." Here we are dealing with cases in which, when engaging with a specific company, users lack the relevant know-how that would allow them to navigate the available privacy options to serve their best interests in the face of potential privacy violations that they are aware of (Cakebread 2017; Turow et al. 2023; Singer and Karaian 2023). This includes cases in which the company provides the option to opt out of certain features (collection of data for personalized advertisements or for analytics, to sell to third parties, etc.), but the user lacks the understanding and the knowledge to effectively exercise this option. For example, the user might be aware of surveillance/monitoring taking place by the company but have no knowledge of how to express their discontent or lack of consent to it. Or the user might know their information is being collected, processed, and shared unless they "opt out", but have no clear understanding of how to change their privacy settings to actually opt out.[56]

Navigating bureaucracy and being tech savvy is something that is hard for the typical user, often requiring significant time and effort, and companies can take advantage of that. Sometimes privacy settings are designed in a convoluted way, buried in obscure text and multiple links that direct to even more links (Cate 2010; Oltramari et al. 2018; Chen et al. 2019)(anyone who has tried to navigate the privacy settings of all their Google accounts has gotten a taste of how it rapidly becomes a massive endeavor, and how easy it is to get lost

[56] Issues related to know-how can naturally be viewed as falling under "informed" consent, since finer-grained control (e.g., over one's privacy settings) depends on understanding how the system works. But it's worth pointing out that these issues can also be understood as being about "uncoerced" consent, in that people who don't know how to choose a different option (i.e., they are functionally unable to opt out) are in some sense being forced to acquiesce. So, the problems raised here are relevant to both dimensions of consent.

along the way). Other options for opting out involve filling out forms, making calls, and in general jumping through so many hoops that many users simply give up at some point, regardless of their original interest in protecting their data. Thus, users often continue to engage with the service or product provided by a specific company, despite the fact that the company has failed to provide the proper disclosure of relevant information and easy interfaces that would make it feasible for the average user to actually opt out of some of their practices. This results in a type of privacy violation brought about by exploiting the limited time, resources, and attention of the user base.

It is important to point out, here, that this type of obfuscation of information and convoluted (and oftentimes only partial) ways of opting out tend to disproportionally impact the most vulnerable members of society (Madden et al. 2017). Navigating confusing privacy settings is time consuming, a luxury that lower class can't afford. It also requires some level of competency in reading legalese, and competency with technology in general, which results in the people who are most disaffected in society facing the greatest potential violations of privacy. Indeed, this is already the case in their relation to the State, where socioeconomically disadvantaged citizens that apply for subsidies are often required, in return, to agree to heavy monitoring of their lives, more and more so in the digital age (Henman and Marston 2008; Eubanks 2014; Madden et al. 2017).

In considering this type of problem for individual consent, we might find two problematic extremes. On one extreme, there is a very minimal disclosure of information, barely distinct from the "lack of awareness" cases considered previously; for example, when it is simply announced that "cookies are being collected", or some other sort of obscure practice is taking place. Here, the disclosure is not enough for the user to know how to limit the collection of their data, or limit how it's being used, while still using the service. At the other extreme, we have companies that decide to dump an incredible amount of obscure legal and technical information that lead users to the more and more common state of "information overload" (McDonald and Cranor 2008; Obar and Oeldorf-Hirsch 2020; Acquisti et al. 2023). This impedes

users from being able to pinpoint the relevant information and tools they need to understand and change their privacy settings.

Solutions for this type of problem are, naturally, heavily rooted in proper and transparent disclosure by corporations of the relevant information about how to exercise the option to opt out, change privacy settings, and generally protect one's information. Of course, we believe that such mitigation attempts will ultimately fail if they are founded in giving individual users control over their data (data control approaches), for all the reasons exposed in Chapter 1; the framing has to shift more fundamentally, to providing information about the *uses* of the data collected. But even if privacy policies did shift to focus on uses of the data, the problems presented here would persist. This is because it is often simply too onerous for an individual to navigate all the complexities of the uses of their data. The combination of this "informational overload" with the non-scalability of privacy management (i.e., the responsibility falling entirely on the individual to read every single policy of every company they interact with), leads to its inevitable failure at the policy level. A "notice and choice" approach (whether requiring to opt-out or more actively opt-in), even if focused on uses of the data and not its mere collection, will not work. For "use control" to work, it has to move beyond the notion of individual consent and into social governance structures (we return to this in Section 3).

Importantly, this type of problem for individual consent concerns the hurdles that users might face when engaging with a specific company. Perhaps the company superficially provides users with some sort of control over their data, but exercising that control is heavily impeded. Or perhaps they fail to provide an option to opt out of any of their monitoring practices, or the most invasive ones. In these cases, the only real choice the user has is either to accept all monitoring practices from the company, or go for a "hard" opt out, and avoid any engagement with the company at all.[57] A "hard" opt out entails forgoing the services or products that company offers altogether, but as we will see, there are many cases in which such a choice to

[57] Another problem arises when people who are engaged with a company decide they want to leave it once they find out about privacy infringements they previously were not aware of—how can they "get control of their data back"?

completely stop engaging with a company imposes enormous costs. Before we get into that aspect of the "voluntary, uncoerced" nature of consent, I want to explore the last and most damning problem to the idea that we can achieve meaningful informed consent from an individual.

1.1.3. Lack of meaningful understanding: under-constrained future uses

Companies can do things with our data that we might never expect or anticipate. This often involves the use of predictive analytics and the immense and far-reaching effects of this type of information processing that we have discussed in Chapter 1. While proper governance can help to resolve the harms that come from the first two kinds of problems for informed consent, this issue is one where a shift in focus to use control is necessary to ameliorate the problem.

A genuine understanding of the nature of the options open to you is a prerequisite for real consent. An important question to tackle in the present context is: What does it mean to be truly informed?

In the digital age, this question become quite complex—we may be aware of the superficialities of a situation while missing entirely the deeper impacts. Think for example of the difference between knowing "this site uses cookies" (with, ideally, an opt out option), versus actually understanding that your information may be acquired by third parties, that those cookies can be used to create a shadow profile of you that is bought and sold and used in ways that are only vaguely disclosed to users. Indeed, your data may be used in the future in ways that not only you do not currently understand, but even the developers did not predict.

Some regulations require divulging the bare minimum of information, for example a pop-up on a website that vaguely suggests that cookies are needed for the correct functioning of the site and/or for targeted advertisement; other regulations (like the GDPR) might require huge dumps of information, though usually on a different webpage, with intricate language and insufficient or unclear ways of opting out. There is often no clear indication of the potential collection of data by third parties and what can they do with it, nor of course any option to

specifically (or easily) opt out of this transfer to third parties. Arguably, proper standards for "informed consent" are not being applied to cases of monitoring and tracking in the digital world, thereby undermining the superficial kind of consent we currently give. This makes it crucial to establish and push for standards of informed consent that apply in the digital realm.

This then is the main challenge for informed consent in the digital age: even if our daily lives happen mostly online, we are incredibly ignorant of what information the companies we engage with are acquiring in exchange for the good or service they offer us, and how that information may be used in the present and in the future. Many consumers may have thoughts like:

- Sure, you can use cookies, I just want to read the article.
- Targeted advertisement? Sign me up.
- Yes, I would like $10 off my next prescription. Here's my info.

These hypothetical persons are aware of the monitoring (superficially at least) and may not be experiencing any obstacles to opting out (in some cases they may even go out of their way to opt in!). Nonetheless, they may woefully misunderstand the ways in which their data will or could be compiled, analyzed, and used. Note that there are two related but distinct issues here: one consists in improper disclosure/communication of what *will* be done with the data (e.g., if the Terms & Conditions are insufficiently specific); the second is the challenge of knowing what *could* be done with the data (which is a function of how technological capabilities change and evolve over time).

In either case, the concern is that users may not truly understand what they are consenting to. This should not be construed as a simple failure of the individual, or laziness on their part. Indeed, the system relies on and perpetuates this asymmetry in knowledge, because it serves companies well to keep these understandings hidden, or open-ended, so as not to scare customers away. Though individuals can make an honest effort to understand the policies of any specific company—and bear *some* responsibility for informing themselves—the sheer

amount of information required makes it impossible to scale this process for every single company one interacts with, and the uncertain and under-constrained nature of future uses significantly compounds this problem. The importance of privacy in this context—as a safeguard for people's ability to pursue their own interests and flourish—is paramount. In a world where access to services is highly dependent on algorithms that are fed our information, many of our fundamental rights depend on the preservation of privacy. This does not necessarily mean arresting the flow of information altogether, but rather *justifying* it: making it responsive to some degree to the beliefs and desires of the users, and not just corporate interests.

The "under-constrained future uses" issue is closely related to the main topic of Chapter 1: how rapidly advancing technologies introduce a crucial *dynamic* element to the game; knowledge, capabilities, and reach can and do change over time, as the power of the algorithms increases, and *predictive analytics* become more and more pervasive. The uses to which our data may be put today are far greater than what they were 10 years ago, and may in turn pale in comparison to what the next decade will bring. This presents a novel challenge to the very concept of "informed consent"; indeed, sometimes the corporations themselves might not realize all of the potential impacts or uses of the data they are collecting and selling. How can a company ensure you are properly informed if they themselves don't yet know what the future will bring in terms of how your data may be used?

This, again, cuts to the central relevance of the *use* that corporate entities make of our data Predictive analytics extends the apparent uses far beyond what most would expect or, arguably, consent to. Of course, this is connected to the blurring of the distinction between "observing" versus "guessing" explored in Chapter 1: when information that previously would only be sufficient for a wild guess now can be used to generate predictions more akin to observations; this presents a major challenge to the idea that we can effectively self-manage our data to protect our own privacy. This in turn leads to the idea that protecting our privacy has to focus on regulating the *uses* of data, more than the collection itself. Indeed, when

dealing with unknown or under-constrained future uses of the data collected, the *use control approach* specifically accounts for this difficulty.

As a first pass, the basic idea here is that consent from individuals shouldn't be directed towards the *collection* or *ownership* of data, since this leaves most of the uncertainty and allowance for unconstrained future uses on the table. By instead focusing on consent for *uses*, this problem is bypassed—if companies are required to obtain consent for each specific use they put the data to, then surprising or novel uses that arise in the future are not automatically permitted. In this envisioned world, "owning" data is not the same as having the right to *use* that data in whatever way you wish. Companies cannot simply ask users for a "blank check" to use collected data in any future way that becomes feasible. Rather, owning data would become more akin to owning land, where the uses you put that land to are typically governed by further zoning regulations and the like. Notice that this can also naturally serve to shift more of the administrative burden from individuals to corporate and regulatory entities: it becomes their responsibility to arrange the proper permits for new uses.

1.2. Uncoerced consent

1.2.1. Lack of feasible alternatives

Here we shift our focus to a kind of structural failure in the broader society; we showcase problems that, we argue, corporations are obligated to take into account even if they do not bear sole responsibility for them (but bear part of it). It can be expressed succinctly as follows: "I see how to opt out, but I'm worried that doing so will make things significantly worse for me." Here, the individual is aware of problematic practices a company engages in and is unwilling to consent to the corresponding violations of their privacy. The corporation provides no option (or an insufficient option) to opt out; therefore, the user faces a choice: either to agree to those practices or not to engage with the company at all. This type of issue is quite broad and can be manifested in many different ways. To illustrate, I'll give a list of brief examples:

Monitoring software in the workplace: if I opt out, will they find an excuse to fire me?

Credit monitoring software: if I opt out, will I be viewed (implicitly or explicitly) as a bigger risk by lenders?

Effective monopolies: if I refuse to shop at Walmart, Target, Amazon, etc. because I do not wish to consent to their use of my information, what are my outside options?

Industry standards for employees: I can quit my job because I don't like the way they monitor me, but the next place that hires me will have the same policy, or worse!

Industry standards for consumers: I can easily switch from Giant Eagle to Trader Joe's for my grocery shopping needs, but if both subscribe to the same tracking policies, I am no better off.

Everyday technological necessities: I can switch cell phone providers, but the next one will have the same problems as the first. I can stop using a cell phone altogether, but then I am excluded from many aspects of society that have arguably become central, if not. Ditto for having an email account.

Medical interventions: If I opt out of a cochlear implant because of the associated privacy infringements, I am forfeiting a certain kind of medical and social support, for which perhaps not comparable alternative exists.

In these and similar examples, we see two main roles individuals typically play when interacting with corporations: consumer (users/customers) and employee. These roles are similar in that they both involve an individual's data being collected, processed, and used. They are also similar in that both involve significant information and power asymmetries with the company. This kind of lack of consent—where feasible alternatives are absent—presents distinct complications. In what follows for concreteness I will frame many of the issues in terms of corporations as employers and their relationship with employees. But many of the dynamics

that arise can be extrapolated to apply also to the role of individuals as consumers, particularly in the case of monopolies and pervasive industry standards.

It is becoming more commonplace for companies to use AI technology to monitor their employees (Allyn 2020; Klöpper and Köhne 2023; Zickuhr 2021). Amazon's tracking of employees in warehouses for efficiency is a notorious example, but this kind of tracking technology has also been adopted in more office-like environments (Sonnemaker 2021). In isolation, the company might feel it is acting responsibly if they inform their employees beforehand about the monitoring that will take place and the purposes for it (notice here the uses of the data collected are explained). They might explain the efficiency techniques behind it and detail how information about them will be stored to create a profile of their performance. If an individual doesn't feel comfortable with these requirements, the company might very well believe that that particular individual should find employment elsewhere.

What is the status of an employee's "consent" in this context? One could argue that it is free, uncoerced consent, since no one is directly forcing them to work for that particular company. However, this is a very simplistic way of understanding coercion. Here we have to explore what it means to have "real" alternatives in choice. In the specific case of employment, one cannot simply ignore the lack of job mobility that many employees face. Workers are not often presented with a vast array of job opportunities to pick and choose from, all matching their skills and aptitudes. And as more and more companies choose to enforce the same monitoring technologies, the pool of outside options effectively shrinks even further. (This is conceptually analogous to price collusion: in this case the "price" that employees pay is not monetary but infringements of their privacy, and the collusive aspect is the essential lack of any "market competition" on this front. We discussion collusion more in Section 2.4.2.) Presented with both lack of mobility and industry-wide standards for monitoring, it is easy to imagine a situation where even employees who are technically free to leave their job or deny a work opportunity are not facing a genuine case of uncoerced consent if they don't agree to

their privacy being infringed upon.[58] I therefore contend that corporations must take into account the real limitations that prospective employees face when accepting a job, and ensure that they are not offering merely the illusion of free choice, or else the "consent" they obtain is merely superficial. Of course, each individual company may face their own strict economic/competitive pressures to collect certain data from their employees; this points to the potential need for a top-down regulatory structure rather than a purely "free market" approach, which is susceptible to such "tragedy of the commons" type failures.

The crux of the issue, then, is whether having a feasible alternative choice (i.e., an ability to opt out without enormous cost) is necessary for real consent.[59] One might argue against this following something similar to Frankfurt-style cases about intent and moral responsibility in which: (a) X does action A; (b) unbeknownst to X, X would have been forced to do A in any case; but (c) we still think that X is morally responsible for A since X freely chose to do it. Explicitly, in our context, this raises the possibility that a worker might (a) agree to work (or continue working) for a company; but (b) in fact have no feasible outside options, though they don't realize this; and so (c) be construed to have provided consent since they think they could have walked away. In other words, there might be an epistemic component to consent.[60]

---

[58] This, of course, is based on a socio-economic structure in which having a job is necessary for survival—or at least where unemployment comes at a very high cost in terms of quality of life, pursuit of interests, and general flourishment.

[59] Of course, there are cases where "real consent" is not given for certain types of surveillance, but other considerations are judged to take precedence and justify the surveillance even without consent. For example, it may be that in all food processing plants there is 24-hour video monitoring in order to comply with safety and auditing standards. Here the ethical question is how to balance such concerns against the rights of individuals to minimal intrusion in their lives. Since these questions arise even outside of the technological sphere, we bracket them here.

[60] Analogously, we might suppose that company C is going to use X's data in bad ways, regardless of whether X gives consent, but X doesn't know this fact. Can X then genuinely consent to C? Does consent actually require the ability to opt out (similar to how moral responsibility does not require the ability to truly do otherwise)? Relatedly, we might imagine a person who knows that they have no feasible alternatives, but doesn't care—they are perfectly happy to "agree" to the company's monitoring anyway. The question here is whether consent always requires an ability to opt out. Finally, there is a way to interpret the case in which it is not necessarily an epistemic issue, about the agent's knowledge. Instead, it could be a metaphysical argument about the causal role they play: what matters is whether the agent played the "right" causal role with regards to the action. On this understanding of these cases,

For now, however, my goal is only to emphasize that for individual consent to be a feasible *ethical and policy framing*, the ability to truly opt out in general would be necessary. This perhaps implies that consent at this structural level might not be completely analogous to individual moral responsibility as in the Frankfurt-style cases. It is, of course, possible that a corporation secures consent from an employee as described above—where they only had the illusion of choice. And perhaps this is good enough from the perspective of that individual employee; after all, if I truly want to work at a company and I don't mind, say, their data collection practices, then why does it matter to me whether or not I could have a job at a different company with different data practices? However, the acceptability of this particular scenario depends, essentially, on getting "lucky"; that is, it's lucky for the company that this particular employee is actually fine with their policies. But the company is operating in a context where a similar employee, who might *not* be fine with the policies, could "look the same" to them—they might appear to be consenting freely, but in reality their "consent" is offered precisely because they perceive no feasible alternatives (and don't want to cause trouble and lose the job). To avoid situations like this, the company still bears the responsibility not to exploit such situations (even if for some employees it might turn out to not "officially" count as exploitation). In a nutshell: the acceptability of these kind of Frankfurt-style scenarios is based on just one individual who happens to be okay with their lack of options. But this is not a good basis for general company policies since it leaves others in a position of being coerced, even if unintentionally. Thus, corporations must bear some responsibility to ensure awareness of the alternatives their users/employees actually have. And when both the company and the employee are aware that there's a lack of real, feasible alternatives, it becomes harder to argue that privacy is not being violated.

It could be argued that something like a tight job market is a problem that the government, and not corporations, has to solve. And it is certainly true that not only the government, but good governance is needed here. Nevertheless, this shouldn't give corporations a free pass to

one might argue that a "real ability to opt out" is not necessary, but what matters for consent is that the agent played an appropriate causal role (and therefore "could have done otherwise" is simply irrelevant).

exploit the broader structural problems they find themselves immersed in. In particular, to the degree that business models and practices are designed to take advantage of citizens' lack of job mobility in order to secure "consent" for privacy violating practices, it is the corporations' responsibility to modify their approach to allow for real consent from their employees.

1.2.2. Monopolies and industry standards

The previous analysis might have a different impact on corporations depending on their size. The bigger the corporation, the more all-encompassing it is, the higher the responsibility they must avoid privacy violations arising from the lack of real alternatives their employees have. The asymmetry in power between corporations and employees is the greater when we are dealing with huge companies. In some places, a sole company employs most of the people in a city, county, etc. In the consumer/provider dynamic, it is even easier to see how giant corporations dominate their market. Easy examples are the "big tech" industries like Amazon, Alphabet (Google), Apple and Meta (Facebook), among others.

One might assume that avoiding engagement with Google and its policies might amount to simply steering clear of Google's search engine, the Gmail suite, YouTube, and so forth. Or that protecting yourself against Meta's snooping requires "merely" the sacrifice of closing your Facebook and Instagram accounts (or never opening one). Similarly, a user might think that avoiding purchases through Amazon is all it takes to escape being monitored by them. But it is evident that avoiding the main services provided by mega-corporations like the examples just considered is highly costly for the individual. By avoiding YouTube, an individual not only loses contact with Google itself, but with all the other companies, producers, creators, freelancers, etc., that share their content on this juggernaut of a platform. Avoiding buying on Amazon in the US can cut off access to thousands of local producers and smaller shops that have to sell through Amazon to stay afloat. These are not minor inconveniences but would already require a reframing of how we navigate the modern world (quite literally: no using Google maps, either!). For something like Gmail, the feasibility of not opening an account might be completely undermined if you are affiliated with an institution whose email service is Gmail.

(Here at CMU, our institutional email is run by Gmail suites, and it is not possible to opt out of having an institutional email address, nor is it desirable).

This gives a sense of how hard it might be to choose the "hard opt out" from companies that we consider to be highly invasive of our privacy. Should a potential student or faculty member avoid joining CMU altogether just to be able to escape Gmail policies?[61] And once we expand our view to take into account Google Analytics or Facebook Analytics, that are used in thousands of other services, and running in the background of uncountable websites, such as online banking among many others, the prospect of avoiding *all* contact with these organizations becomes close to impossible. Monopolies, by definition, imply a decrease in choice for customers and employees alike. Mega-corporations therefore have correspondingly stronger obligations to safeguard customers' and employee's privacy.

Problems can also arise in the absence of outright monopolies via a variety of related mechanisms, perhaps most directly exemplified by collusion between so-called "competitors", which can recreate many of the effects of a monopoly. Consumers don't have much in the way of "alternative" options in the case of, for instance, Comcast and AT&T, who have divided their internet coverage by region (so most people only have access to one of them). But even without direct, malicious collusion, the explicit or implicit adoption of industry-wide standards can generate similar effects and perpetuate these problems. By "implicit" here I refer to the independent, uncoordinated adoption of similar practices (say, when several companies settle on essentially the same set of policies—each for purely internal reasons, not because their competitors are doing it). Each of these scenarios has a similar bottom line: the company's economic growth or profit (taking their shareholders' interests as their primary goal). This is not to say that businesses are inherently evil; rather, the very existence of a business is usually predicated on protecting their shareholders' interests, not the public good. Of course, the

[61] CMU 's Gmail powered services operates under a somewhat different set of policies, though it is not clear to the user in what ways they vary (users are still recommended not to put any sensitive information on these emails).

public good can be benefited in the process of seeking revenue, but it is not the main goal of most companies.

Lack of regulation (which should focus on the public good) leaves fertile ground for many businesses to "independently" implement effectively the same invasive practices, since these are the ones that lead to the highest revenue, require the least work, or position them best for future enterprises. Sometimes companies are not even directly profiting from privacy invasion, but merely choose their business model in conformity with a privacy-invasive industry standard. After all, implementing privacy protecting measures can be more expensive than simply defaulting to the standard, so there are few incentives to invest in protecting privacy. In other cases, privacy-invading practices may be implemented not for their own sake, but as side-effects of other practices. In both cases we are faced with the ethical question of whether reasons matter when judging actions.

This is unavoidable in the current business model that authors like Shoshana Zuboff have termed "*surveillance capitalism*" (Zuboff 2019).[62] While it is clear that direct collusion and explicit industry-wide standards generate stricter moral obligations and, correspondingly, may require tighter regulation, it's perhaps initially less clear by what standards we should judge and regulate *implicit* industry standards. I would argue, though, that the moral landscape here is not genuinely different, particularly when we focus on the *outcomes* of such cases and their impacts on consumers. Moreover, companies don't make decisions in isolation; even when it is not explicitly coordinated, they are always paying close attention to the practices other

[62] "Surveillance capitalism claims human experience as raw material for translation into behavioral data. That data is partially used to improve the digital products or services; but most importantly it is declared 'proprietary behavioral surplus' fed into 'machine intelligence' manufacturing processes producing 'predictions products.' These 'behavioral prediction products' are sold in a new type of market: the 'behavioral futures market'" (…) "The surveillance market is a hugely profitable market. When Google just embarked on its surveillance capitalist journey in 2001 its net revenues jumped to $86 million (a 400 percent increase); in 2002 revenues rose to $347million, $1.5 billion in 2003 and $3.5 billion in 2004. 'The discovery of behavioral surplus had produced a stunning 3,590 percent increase in revenue in four years.' (Zuboff, 2019: 87). No wonder, then, that Facebook, Microsoft and even net providers were keen to join the party. And it is these profits that yield invasive and imperialist companies." (https://www.academia.edu/38403327/The_age_of_surveillance_capitalism_Diggit_Magazine.pdf)

companies get away with and that have been shown to be profitable, and constantly adapting and reacting in this environment. In this sense, even superficially "independent" adoption of policies can be understood as much more coordinated than it may appear. This makes the distinction between explicit and implicit industry-wide standards much more nebulous, and much less morally exculpating.

Given this, we can see that the problem with lack of feasible alternatives for individual consent are as widespread as they are problematic, and only grow worse as monopolies consolidate and (explicit or implicit) industry-wide standards adopt more invasive privacy practices. As we saw with the previous problems of lack of awareness and know-how, issues can disproportionately affect people who are worse off. This is mainly the result of the high cost of a "hard opt out". Generally, options for forgoing mainstream corporate products are considerably more expensive, and so are only really options for the wealthy (indeed, for some technologies such as a smartphone, a hard opt-out is, at best, feasible for only a few hundred people in the world).

The constraints can be subtle, and it might be tempting to invoke some platitudes: "you have a free choice!", "just quit!", "just shop somewhere else!", "just don't give your info!". But those choices are often not so easy, and attach harsh consequences to nonconformity.

Practical solutions this lack of feasible alternatives are hard to solve. In essence, the ideal solution is straightforward: viable alternatives need to be made available. But implementing this requires us to identify the many and varied constraints that people may face in a range of different scenarios, and a reshaping of the current prevailing business models and economic landscape. In other words, it requires a massive structural shift, which is impossible to achieve without better governance, both from the government and from the private sector. Though comprehensive solutions for solving these issues go beyond the scope of this thesis, the conversations pertaining the shift to *use control* present an opportunity to tackle regulation in ways that can at least indirectly contribute to better the current ecosystem where monopolies thrive: to the degree that the focus on impacts of the technologies require a wide restructure

and creation of regulatory agencies, it provides an opportunity to reshape the current landscape dominated (especially in the USA) by a strong "lassies-faire" market approach into one were wide-spread industry standards that are pernicious for the consumer are mitigated or disincentivized.

***

To summarize so far: we have laid out a systematic way to see problems with individual consent showing different kinds of issues that lead to violations of privacy. When considering cases of uninformed consent, violations can start from the individual having a total "lack of awareness" of a privacy infringing practice that's taking place and though a basic solution here is to generate or require that basic awareness, doing so is clearly not enough. Issues pertaining "lack of know-how" concern situations where there is awareness of a potentially problematic practice but the individual lacks the right know-how to successfully opt-out from all or at least some invasive practices, for example an inability to successfully navigate complicated privacy settings. There has been significant attention to this issue recently, resulting in companies trying to make their privacy options easier to find and navigate. Broadly speaking, solutions here involve creating easy-to-navigate privacy policies and opt out options (for the unnecessary or overreaching invasive practices within a company). But even this still might not be enough. On the one part, these solutions heavily burden the individual with an increasingly complex array of policies they have to interact and personalize. Furthermore, a lack of meaningful understanding is prevalent here because the uncertainty of under-constrained future uses that are hard to anticipate and therefore to consent to. To this final aspect, use control can be presented as a step towards more meaningful intervention, since it forces companies to specify concrete cases of use of the data collected.

Finally, we explored cases that present problems for "uncoerced consent", as they involve both awareness and apparent choice but a lack of genuinely feasible alternatives. The "hard opt out" options can come at intolerably high costs—and can disproportionately affect people who are already disadvantaged, given the harsh consequences for nonconformity, many of

which are economic. These issues are exacerbated by monopolies, mega-corporations, and industry standards, affecting both employees and customers. The solution requires the presence of viable alternatives, which in turn will require structural changes.

1.3. So, what's the status of consent?

Thus far we have shown that individual consent is not sufficient to safeguard privacy and protect people from harm, *even when* we focus on use control rather than data control. In light of this, what we're going to suggest in its place is a broader, social implementation of such safeguards on data use, similar to IRB boards to FDA regulations. In other words, domain expertise is necessary to assess which uses of data benefit all stakeholders or what trade-offs are prudent, and of course such assessments cannot simply be provided by those in the employ of the very agencies (Meta, Google, etc.) whose interests are primarily aligned with their shareholders.

Could we call this a kind of "social, distributed consent"? Maybe. Or maybe it's something better thought of as a new category entirely, distinct from consent. We aren't going to stake out a firm position on whether we *call* it a species of consent or not, but focus instead on outlining the basics of its implementation from a practical/policy perspective, as well as what its advantages are and what challenges it will still face.

Section 2 | **Solutions and the Importance of Use Control**

The previous sections have focused on a variety of obstacles to a meaningful notion of individual consent in the modern digital landscape. What's the upshot? There are two different claims here. On the one hand, some might claim that, with sufficient time, energy, and expertise, an individual can overcome these obstacles, and consent can play the morally transforming role we might wish it to. If so, our efforts ought to be directed towards finding ways of empowering individuals with the access to expertise that they need. Supporters of this claim (Kitkowska, Högberg, and Wästlund 2022; Emami-Naeini et al. 2021; Perera and Perera 2021) tend to focus

on providing users with better understanding of the policies, through a system that relies on experts to communicate the relevant information. This is a necessary step towards enhancing the ability of the individual to properly understand the agreements there are consenting to, making it a necessary, if not sufficient way to tackle the problem.

Here it's worth mentioning user interfaces for *privacy agents* (Cranor, Guduru, and Arjula 2006; Tondel, Nyre, and Bernsmed 2011). The idea here consists in companies providing a standard machine-readable format for website privacy policies, so that a program can read privacy policies automatically, compare them with a user's privacy preferences, and alert and advise the user through an understandable interface. The goal is to lessen the individual's burden to properly understand everything that goes into these policies, since once you set your privacy preferences, these "agentified devices" collaborate among themselves and with other devices so that the user's privacy preferences are satisfied (Galvan et al. 2021). Imagine for example that the user selects that they don't want their location to be tracked. The program then will select or deselect the corresponding settings for you (say, they imped your MAC address to be tracked as you move through a store). Though promising, there's a problem here regarding the preferences that the user is presented with and how much they need to understand them. Do they need to understand what a MAC address is or only what tracking location entails? But more importantly, the user might want their location to be tracked or not *depending on the specific use of that tracking*. I might want to share my live location over WhatsApp with my friend so we can meet, but I might not want WhatsApp to use that location tracking for other uses, or selling it to third parties. A similar, related approach in the one concerning Privacy "Nutrition Labels", which are meant to make people understand the potential risks of their IoT (Internet of Things) or smart devices, by making it clear what data is being collected, stored and for what purpose (Emami-Naeini et al. 2022). Again, there is a lot of merit to this idea, but it has faced implementation issues (Leon et al. 2010). I'll argue that these approaches can be better implemented in a more useful way for the user if they are developed under the use control structure (and use taxonomy) that I'll delineate in the following section. From this perspective, it is tempting to think (or at least hope) that this problem can be resolved without resorting to top-

down regulation: a combination of greater availability of experts along with tools that lead the average users to expertise should suffice.

However, as already touched on, the problems with consent run deeper than mere lack of expertise. Technology can be opaque on at least two levels: first, many algorithms are black boxes, working in ways that not even their creators do not fully comprehend (the prime example being deep neural networks); second, corporate practices are often proprietary or protected by other legal safeguards. Both of these non-transparencies put hard limits on what information is publicly accessible to common citizens. Even an expert who devotes their life to understanding privacy policies will face the hard limit of simply not having access to the totality of the internal policies of a given company, still less the internal technical details of the relevant algorithms (which can also be quite complex and require a whole other field of expertise to understand).[63]

For these reasons, approaches to facilitating consent that work solely by trying to "empower" individuals will not be enough to achieve the goal of informed, uncoerced consent. For starters, some degree of regulatory structure is essential—corporate privacy should not trump individual rights, and regulatory agencies exist (at least in part) to ensure that external experts who serve the public are in a legal and informational position to apply their expertise to safeguard the rights of ordinary citizens. This is a familiar concept: when I put Heinz ketchup on my fries, I don't need to be an expert on food toxicology, or personally know the secret recipe for the condiment, to feel safe in consuming it. The FDA has (in theory) already vetted the product under the supervision of appropriate experts and with access to the necessary information. Something similar is needed here, not because the government has to regulate everything, but because the parallel between food products and drugs with ML algorithms is pertinent. The inability of ordinary people to learn and understand relevant facts to navigate these technologies, and the potential harms that can come to them because of this inability, makes it a good candidate for

[63] Moreover, as discussed in Section 2.3, the fact that the uses of data can change over time in unforeseen ways also presents a kind of "hard limit" to what even experts could presently know. Meaningful consent, as we have argued, cannot be given for mere data collection with open-ended future uses; it cannot be a blank check. It must be explicitly tied, in ways that we shall explore below, to use control. But for the moment we will focus on the problems with individualizing consent in this context.

government oversight. It is far too onerous for an individual to have to develop expertise on each food item they consume in order to understand if it is toxic when they "consent to eat it". The FDA serves as an intermediary agency that has a (somewhat) fiduciary relationship with the consumer, so the individual burden to ascertain the safety of the food (or drug) is at least partially lifted.[64]

Now, if we agree that the challenge presented is not one that can be borne solely by individuals, but rather requires regulation and regulatory agencies to access parts of the process that are not meant to be public (tech development, design and implementation), the next step is to assess the kinds of mechanisms that might address this. Unfortunately, even the most well-informed, good-intentioned, generously funded regulatory agency is doomed to fall short in this context if its focus is on regulating *data ownership*, for all the reasons presented in Section 2.3. In a nutshell: this approach is too open ended, since no matter how much we may know about the *current* potential uses of data, we will never be in a position to truly understand how those potential uses may evolve over the next decade, let alone over a lifetime. If proper data *collection* is all we focus on, we relinquish any control of how this data may be used in the future. So, as already argued, since data *uses* are the real subject of moral evaluation, uses must also be the focus of regulatory agencies.

[64] Other countries have recently expanded this role by implementing a 'junk food law', as Colombia did in 2022, which requires warning labels on foods that are high in salt, sugar and saturated fats (see https://www.vitalstrategies.org/colombias-bold-new-law-to-label-foods-high-in-fat-salt-and-sugar-will-save-lives-and-empower-consumers-other-countries-should-follow/). This approach combines strong regulatory intervention with measures to empower individuals' ability to offer informed consent by providing relevant information (on labels) to consumers to make it easier to understand which foods are unhealthy (even if they are not toxic). This is a good example of a balance between paternalism and autonomy, where regulation focuses on the public good (in this case public health) while also leaving the ultimate choice to consumers to buy or not each product. Notice that the individual doesn't have to become an expert to understand the complicated nutritional labels that are standard on any processed food. Instead, these labels (e.g., "added sugars", "high in salt and sodium", "high in saturated fats", etc.) already include the relevant analysis of what constitutes "high amounts of added sugars", and already take into account that this is information relevant for the consumer. Ideally, tech policies would take a similar approach.

2.1. How does external regulation work when it is focused on uses?

In Chapter 1 (Section 3.2) we presented an overview of what a shift to use control might look like. The focus there was on the early stages of the development of the technologies, which meant focusing on ways that companies themselves can self-regulate by, for instance, implementing internal "Ethics and Society Review" (ESR) boards and mechanisms of that sort (See as an example Bernstein et al., (2021) or Microsoft AETHER committee). [65] As we noted, such mechanisms would need to be complementary to external regulatory structures such as clearer legislation and proper auditing, among others. In this section, we want to highlight some preliminary ideas that focus on these external controls. The core idea is that data collection policies have to focus on the uses that the collected data will be put to, rather than just on the collection, storage, and processing of the data. For example, information can be gathered with the intent of being used for, say, scientific research. This specific use, for this specific research, would be vetted by a regulatory agency, so that when presented to the individuals, they can trust their information will be used responsibly and there are mechanisms in place to ensure this. A crucial point here is that there must be some sort of external regulatory structure which has access to the internal practices and processes of the company. By *combining* this sort of access with a shift to use control, we make possible meaningful policies, restrictions, and means of redress that can actually function to return some power to individuals over their own privacy. Notice that in such a scenario, it is not left as a burden on the individual to navigate, and somehow try to personally guarantee that they won't be harmed by unforeseen future uses of data they may choose to share.

This sort of regulatory framework will sound too demanding to some. For starters, it means that companies must craft policies focusing on specific uses of the data. The users in tandem give consent to specific uses, which implies that whenever a *new* use becomes available or desirable to the company that holds the data, they cannot immediately implement it but instead have to seek approval all over again. To facilitate this process, the regulatory infrastructure

[65] https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1%3aprimaryr5

ought to include crafting a *taxonomy of uses* that will make navigating the process easier and more streamlined.

A key practical issue here is how to individuate genuinely "new" or "different" uses when they arise. If the individuation is too coarse, then companies will effectively get away with using prior consent to implement all sorts of new uses; this is the case with current broad wording given on many data policies that effectively cover *all* possible uses.[66] On the other hand, if the individuation is too fine, then companies will find themselves paralyzed having to navigate through a new approval process for every tiny change in how they operate. Neither of these extremes is desirable. The idea is to strike a balance between *freedom to innovate*, on the one hand, and *consumer protection*, on the other. Of course, this is not a new problem: regulators in any industry already have to worry about this balance (e.g., when does a product change enough that it needs to be tested again for safety?) We argue here that working within a *taxonomy of approved uses* has the potential to help maintain a proper balance. Developing such a taxonomy of uses can clarify the different domains (healthcare, policing, education, advertising, etc.) and what uses are considered appropriate in regard to scope and application of the data collected.

Remember here that when we talk about *uses* there are two senses, both of which have to be taken into account: the question-answering sense and the action-guiding sense. Of course, the question-answering sense needs to be considered, i.e., we must ascertain that the algorithm being used is correctly measuring the right variables to answer the question we care about (with the right values, proxies, benchmarks, weights, etc. See Chapter 1, Section 3.1.1 for more on this). But the action-guiding sense will also be crucially important to assess in any regulatory

[66] Following the Instagram Data Policy example (found here: https://help.instagram.com/155833707900388), Section II. "How do we use this information?" Lists several uses, one of which reads: "Product research and development: We use the information we have to develop, test and improve our Products, including by conducting surveys and research, and testing and troubleshooting new products and features." Notice this basically gives a blank check for any possible use the company wishes in the present and the future.

framework. These two senses of use are intimately connected and must often be assessed in combination (Chapter 1 Section 3.1.2). An example here will help:

Consider how during the 2020 pandemic schools moved online and there was a massive push for the use of technologies to do virtual learning. AI powered algorithms were used across K-12 schools as part of the post-Covid education reformation, and many of these were monitoring and surveilling the students in various ways (such as GoGuardian, Gaggle.Net, Securly, and Bark Technologies)(Anand and Bergen 2021).[67] Some of the software tracks every keystroke, click, and search query, all of which are recorded and analyzed by the companies, even outside of school hours (Laird et al. 2022; Kelly, Rivera, and Intagliata 2022). Because mental health has been an important issue with students, and this was aggravated by remote learning, some of these companies have used this as a selling point, promising to use artificial intelligence to identify students who, based on their online behavior, are at risk of hurting themselves or others. At a first glance, this example seems to show a case where invasive tracking is used to *assist* students with mental health issues. A coarse approval for pro-social uses could lead to green-lighting such a project. But a closer look raises important red flags for both the question-answering and action-guiding senses of use, here.

First, we consider if the behavior tracked (monitoring students' public social media posts or tracking what they do in real-time on their devices) can accurately predict mental health needs, such as whether students are suicidal or need intervention (question-answering sense). Though efforts to build such predictive models exists, there is still no solid evidence that this is the case (Huckins et al. 2020; Costello and Floegel 2020; Conway and O'Connor 2016)[68].

[67] Sources: A *Center for Democracy and Technology* Report https://cdt.org/insights/report-hidden-harms-the-misleading-promise-of-monitoring-students-online/
Media reports: https://www.npr.org/2022/08/17/1118009553/more-kids-are-going-back-to-school-so-why-is-laptop-surveillance-increasing; https://www.bloomberg.com/news/features/2021-10-28/how-goguardian-ai-spyware-took-over-schools-student- devices-during-covid

[68] The CDT report reads: "While students report they are being referred to school counselors, social workers, and other adults for mental health support, they are also experiencing detrimental effects from being monitored online. These effects include avoiding expressing their thoughts and feelings online, as well as not accessing important resources that could help them" (Laird et al. 2022, p. 5).

Moreover, and more importantly, even if we suppose that these predictions are accurate, how will this information actually guide actions (action-guiding sense)? As it happens, these features supposedly for identifying and helping students at risk were in fact used to send alerts to the police, who would respond with home-visits (Kelly, et al. 2022; Anand and Bergen 2021). Police can in many cases exacerbate existing problems, and are not generally considered the best intervention for mental health crises. As the CDT reports, these kind of interventions typically end up being used more for discipline than safety, and widen equity gaps by putting more pressure on POC, low-income[69], and LGBT+[70] students (Laird et al. 2022). Moreover, the monitoring creates chilling effects where students learn to avoid expressing their thoughts and feelings online, and refrain from accessing important resources that could help them (ibid).

Uses, then, must include the relevant interventions, which have to be vetted by organizations that know which interventions actually work. Streamlining these kinds of uses and interventions would help create a structure that companies can more easily follow. Deciding on which uses are permitted, even if it is for the social good, is a job that should not be left solely to corporate interests; it must be supported by a healthy structure of governance that includes external regulators, either private or governmental. Individualized consent is still an operative notion here, but the idea is that takes places against the backdrop of regulatory oversight, which functions to streamline and "sanitize" the relevant choices presented to ordinary citizens.

As with most regulation, any set of rules is bound to be "gamed"—whenever a threshold is set, for example, it creates an incentive to do things *just barely* above that threshold. Consider in the food industry the case of eggs, and how labels such as "free-range" can be just marketing

[69] "**Students from low-income families, Black students, and Hispanic students are at greater risk of harm:** Previous CDT research showed that certain groups of students, including students from low-income families, Black students, and Hispanic students, rely more heavily on school-issued devices. Therefore, they are subject to more surveillance and the aforementioned harms, including interacting with law enforcement, being disciplined, and being outed, than those using personal devices."(Bradfort Franklin and Thakur 2021).

[70] "**LGBTQ+ students are disproportionately targeted for action:** The use of student activity monitoring software is resulting in the nonconsensual disclosure of students' sexual orientation and gender identity (i.e., "outing"), as well as more LGBTQ+ students reporting they are being disciplined or contacted by law enforcement for concerns about committing a crime compared to their peers." (Ibid 2021)

ploys designed to conjure up idyllic images of happy hens roaming freely on green pastures, when in reality the requirement for the label is very basic (Certified Humane 2021).[71] One solution here is scaffolding, that is, developing a tier system that allows for a range of labels that are more meaningful in concert. An existing example would be Whole Foods' approach to meat quality standards, where on top of some baseline standards (no antibiotics, no added growth hormones, not crated, etc.), they have a clear taxonomy of animal welfare standards with a number system that ranges from 1 to 5+ (1-Base Level, 2-Enriched Environment, 3-Outdoor Access, 4-Pastured Raised, 5-Animal Centered, 5+-Entire Life on a Farm) ("Meat Department Quality Standards" n.d.).[72] An industry- wide implementation of Whole Food's initiative would benefit the public, supporting better understanding of what they are consuming and what values they are supporting.

This example is perhaps too simple, in the sense that it is focused on past actions (how the cows were raised) and not future actions, which is the case for data collection. Nevertheless, it does demonstrate how meaningful use-taxonomies can draw from regulations that have worked in other areas to benefit consumers.

A taxonomy of uses is also important for establishing *precedent*. Of course, we don't want to keep reinventing the wheel—ideally, we should be able to mostly re-use certain regulatory structures to apply them to "new" uses that are similar enough to existing ones. In order to implement this, we need to know which uses are "relevantly similar", a knowledge base that takes time and resources to set up. But this is true of essentially *all* regulatory structures in different domains—financial, health, drugs, etc. Successful regulations depend on expertise *and* experience within the regulatory agency, since there are always "gray area" cases that require human judgment. Thus, the promise of this approach is not that it will be incredibly easy or

[71] "According to the USDA, 'Free-Range' only means that hens are 'allowed access to the outside.' Technically, a producer could put in a few small windows and call birds 'Free-Range.' This label isn't a guarantee of animal welfare, of how much time hens spend outside, or of the quality of the outdoor space" https://certifiedhumane.org/all-about-egg-labels/

[72] https://www.wholefoodsmarket.com/quality-standards/meat-standards

always straightforward, but that in the long run it will allow us to better align the uses of our technology with the values that we want our socioeconomic systems to inhabit. Values that, one hopes, prioritize the flourishing of the individual over corporate quarterly statements. [73]

As it happens, recent regulatory policies around the world have started to use terms close to what we would expect within a use control framework. Take for example the California Privacy Rights Act (CPRA) that took effect very recently, in January 2023, expanding on the 2018 California Consumer Privacy Act (CCPA). Here you can already see wording that focuses on limiting the uses of the data collected to the ones disclosed to consumers, as these excerpts show:

> "The categories of personal information to be collected and the purposes for which the categories of personal information are collected or used and whether that information is sold or shared. A business shall not collect additional categories of personal information or *use personal information collected for additional purposes that are incompatible with the disclosed purpose for which the personal information was collected without providing the consumer with notice consistent with this section*." (Emphasis mine)[74]

> "If the business collects sensitive personal information, the categories of sensitive personal information to be collected *and the purposes for which the categories of sensitive personal information are collected or used*, and whether that information is

[73] It is worth here going back to the "privacy agents" and privacy "nutrition labels" cases. Note that a taxonomy of uses would enhance their potential by clarifying the structure of what is presented to the users: the proposed machine-aided interfaces that compile privacy policies for users to understand such Privacy "Nutrition Labels" (Emami-Naeini et al. 2022) and the automatic set up of their preferences so that the program arranges the settings for them (Galvan et al. 2021). These approaches are meant to decrease the information overload and over choice for the user(Colnago and Guardia 2016) which is a necessary step in the right direction. Nevertheless, often these labels or preferences fail because the terminology is to obscure and the uses descriptions too broad. As mentioned, sometimes the users care less about the specific action (tracking my location) than about the concrete present and future uses of these data collection. Here is where use control can enrich these approaches, since a clearer use-taxonomy that has already been vetted under a fiduciary relationship from the state and private companies would help design these interfaces in a way that could more clearly capture what the user cares about and more trustfully provide choices that are not concealing pernicious uses.

[74] https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5

sold or shared. A business shall not collect additional categories of sensitive personal information or *use sensitive personal information collected for additional purposes that are incompatible with the disclosed purpose* for which the sensitive personal information was collected without providing the consumer with notice consistent with this section." (Emphasis mine)[75]

Another good example of this natural shift into use control is the EU Artificial Intelligence Act (AI Act), a regulation proposed in 2021 that follows a risks-based approach, classifying AI applications (i.e., uses) as low-risk, medium-risk, or high-risk.[76] This is part of the structure that a useful taxonomy ought to make available, including a classification system and the corresponding requirements for each category.

Aside from risk levels, a useful taxonomy might also encode other aspects of data use, such as the general domains within which the uses may fall, including especially what types of decisions will be influenced. The very same data could conceivably be used to inform a variety of predictive algorithms, and thus influence decisions that span very different domains. The pattern of websites that I visit and clicks I make, for example, might be useful to inform various sorts of targeted advertising. But the very same information could be used to help make judgements about mental health status and interventions, loan repayment likelihoods, "work ethic", etc. This is a big part of the reason focusing on use control rather than data control was important in the first place. The role of a well-crafted taxonomy in this context is to lay out the rough boundary lines across which consent does not "carry over"—i.e., I may be happy for my web browsing history to be used to inform which ads I see, but not which medications I am prescribed. The considerations above are only compounded when the data itself is highly sensitive, such as health information or biometrics.

[75] Ibidem

[76] https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/

In fact, the protections that are typically in place to guard against misuses of sensitive data are an excellent template for the kind of regulatory structures we are suggesting here more broadly. If an individual agrees to take part on a medical study, for example, it is quite explicit what uses will be made with the data that is collected. The research team is not typically free to reuse the same data for a totally different study, or share it freely with other groups. This seems right and natural given the nature of the information and the obvious potential misuses and privacy concerns. Part of the point we are making is that the concept of "sensitive data"—and the corresponding protections it enjoys—*must be broadened* in light of the transformative power of ML algorithms to extrapolate from seemingly innocuous information (see Chapter 1). Thus, just a participant in a medical study has a right to expect their data will only be used in certain ways, similar an internet user who opts into targeting ads has a right to expect that the data collected to support this will not also be used for wildly different applications.

2.2. Consequences: What about highly multi-purpose predictive algorithms?

We briefly explore here some consequences of the kind of "use control" we are advocating for predictive algorithms that are characterized by being *highly multi-purpose (HMP)*, a central example being LLM (Large Language Model) based chatbot technologies, which have dominated the news cycle in recent months. (Of course, there are other examples of HMP algorithms, including other forms of synthetic media like deepfakes, or "factual" predictive technologies like facial recognition, which was discussed in Chapter 1, Section 3.1.2.)

Many companies (Open AI, Microsoft, Google) are releasing LLM based technologies to the public explicitly framed and advertised as highly multi-purpose tools that cross several categorical/disciplinary boundaries. The issues and debate surrounding LLMs is vast; here, I want to focus specifically on how the notion of *use control* we have developed would apply to such technologies. Given the account we have developed, such HMP and boundary crossing technologies would naturally face much steeper regulatory hurdles, precisely because their "uses" count simultaneously in many different domains, and so in principle ought to be subject to many different (though likely partially overlapping) regulatory frameworks and precedents.

Remember that our use control framework was motivated, in part, by the necessity we identified in many contexts of shifting the onus of ethical decision making from the user onto the companies and organizations that develop and deploy AI tools. Again, while there is a complex debate in the specific context of LLMs about how exactly to balance responsibility between developers and users, here we are focusing more narrowly on the potential for applying the use control framework in some form to such HMP technologies.

Of course, this question arises to some degree with all technologies; however, HMP technologies arguably multiply the relevant risks, since their open-endedness by design magnifies the unexcepted and unforeseen (and sometimes also the predictable and foreseen) misuses of the technology. Indeed, presenting a technology in an open-ended way can sometime serve as a strategic way for a company to effectively wash their hands of their responsibilities to mitigate harm, unloading the blame to "bad actors" who are "abusing" the tech. After all, when a technology is presented without specific use cases, it becomes easier and more natural to blame users for inventing malicious use cases. And of course, it is more convenient for developers to sidestep responsibility for harmful uses by implicitly characterizing them as "misuses" or "abuses". The use control framework we consider can serve to effectively block this maneuver, since it requires an up-front specification of use cases to submit to the appropriate regulatory body, thus shifting the liability/accountability to an earlier stage of development. However, as we will see, the sheer magnitude of "use cases" at play for HMP technologies presents certain feasibility challenges to this approach, among others. We discuss these below.

There are a plethora of potential harms associated specifically with LLMs (for an overview, see (Bender et al. 2021; Weidinger et al. 2021)); exploring them systematically and exhaustively goes well beyond the scope of this chapter. Here, keeping to the spirit of this thesis, we will focus on three core privacy-related issues and analyze how our concept of use control interacts with them.

(1) These models are trained on massive datasets of content scraped from the internet and other sources. This means that information that individuals disclosed with a specific, local scope in mind (i.e., audience size and purpose), is being collected without their awareness and used in ways they likely never conceived (See Chapter 1, Section 2.2 "Public versus private information"). This naturally connects to a related but distinct problem, namely (2) that people can "exploit" these chatbots to try to obtain sensitive and private information about others, for example with the intent to dox them (see Ganguli et al. 2022). This is possible because the immense datasets with which the models are trained were not carefully curated and therefore contain lots of PII from citizens all over the world. Finally, (3) it is a known fact that these models, by the nature of how they work, present fabrications as if they are facts, often in a way that seems believable; this means they can present false information about a person that is then taken as true by people and can become hard to disprove.

Thus, using the conceptual framework we introduced and developed in Chapter 2, problems (1) and (2) can be understood as bringing about *informational harms,* while problem (3) constitutes *presentational harms*. We will not completely rehash these concepts here, as they were discussed in depth in Chapter 2. We simply remind the reader that the scope of "privacy-related harms" we have explored in this thesis goes beyond merely informational harms, i.e., beyond harms that consists simply in revealing factual information about a person that they would rather keep private; it also includes presentational harms, i.e., creating false narratives about a person that they must then fight to disprove. For a concrete example: we should not only consider it a violation of privacy to distribute a naked picture of someone without their consent (informational harm), but also to create and then distribute a synthetic naked picture that is presented as if it's real (presentational harm)—whether it is an accurate depiction or not.

Returning to LLMs (and HMP tech more broadly), suffice it to say that a technology that is so replete with both kinds of harm requires proportionate oversight and regulation; this issue is especially pressing since, as we are presently witnessing, left to their own devices many tech companies will happily launch these models to the public without properly mitigating these risks (or, in some cases, even explicitly recognizing them).

Just as the problems are multi-faceted, so too must be the potential solutions. With regard to (1) and (2), at least part of a solution surely involves more carefully selecting the data that feeds into the training models so that it is properly responsive to the ethical concerns (privacy and otherwise) of individuals. This issue has been particularly relevant with "generative art" AI models such as MidJourney, DALLE or Stable Diffusion. Artists have pressed substantial concerns and even initiated legal processes to combat the use of their content without their consent, nor indeed any option to opt out or offer of compensation for developing a technology that is ultimately designed to replace them. (Stable Diffusion has promised to make it possible for artists to opt out of the newer version of their model—though not from the already existing one (Heikkilä 2022). Adobe is planning on developing a generative art AI model that is trained *only* on public images and work from artists who have explicitly opted in (Kastrenakes 2023)).

However, mitigation efforts like these, while important, do not truly resolve the problems with HMP technologies we have outlined. Indeed, the fundamental purpose for introducing our framework of "use control" was to respond to the fact that seemingly innocuous data can potentially be used (now or in the future) to power predictive engines that go far beyond the scope or even imagination of what anyone can presently consent to. Thus, for essentially all the reasons we have discussed already in Chapter 1, we cannot hope to design the data gathering process in a way that makes it "ethically responsive" to the privacy concerns of all users, since these concerns will evolve over time in potentially unforeseen ways. A truly ethically responsive design must focus on uses, and this is where we return to the special challenge of regulating a technology like an LLM—namely, the uses are so multifaceted and unconstrained, it seems a monumental task to release the technology in any way that doesn't run afoul of the kind of use control we have advocated. (All this is to say nothing of problem (3), where the well-known massive misinformation potential of LLMs can impact individuals' privacy (e.g., their reputations) in a multitude of ways; we return to briefly consider this below.)

What would it look like in practice to try to apply a use control framework in cases like these? There are two broad, closely related issues we will tackle here: a *feasibility* issue, and an *incentive* issue. With regards to feasibility, as we have foreshadowed, the question is to what

extent a technology that is relatively unconstrained in the uses it can be put to can be regulated in a way that requires "novel" uses to be explicitly greenlighted (i.e., by some appropriate regulatory body, as discussed earlier in this chapter). What counts as novel? Will chat bots constantly be refusing to answer questions until they are given the go-ahead by whatever governing regulatory agency they fall under the umbrella of? If each query counts as a novel use, the entire framework becomes obviously unworkable. On the other hand, if LLMS reach a point where they can answer questions about a person's health, or their chance of committing a crime, or committing suicide, or defaulting on a loan, or being a good employee, etc., then they are effectively simulating many (or all!) of the problematic uses that the use control framework was meant to constrain and make ethically responsive. So, the feasibility problem boils down to the question of how to (or indeed whether we can) distinguish "novel" uses in a way that allows LLMs to function as useful tools without giving them free reign to violate peoples' privacy in a myriad of ways.

The incentive problem follows the feasibility problem: supposing we can, in principle, apply the use control framework in a functional way, will the resulting regulatory burdens be so heavy that HMP technologies like LLMs end up heavily disincentivized? To what extent is this an undesirable consequence? Concretely, heavy regulatory burdens could force LLMs to become much narrower in their domains of application or permitted uses. For instance, an LLM meant for assisting in creative writing (e.g., "pretend you are an evil computer who wants to break free, what would you do?") might be (legally required to be) distinct from one used for factual queries (e.g., asking about historical facts). This raises the broad question of innovative freedom vs. public protection: some would argue that any extra hurdles for multi-use LLMs will unacceptably slow down innovation in this area and the usefulness of the tool; others might push back by arguing that it is precisely in these boundary-crossing cases that the most danger exists, and therefore this is where we *should* proceed with the most caution.

Limiting LLMs to specific domains or use cases is an approach to mitigating risk that has been considered more generally, outside the specific context of implementing a use control framework. The basic idea is to either limit the scope of the training data used to develop an

LLM (e.g., only medical texts or legal case documents), or to limit the types of queries the LLM will respond to, or both (see Wolfe 2023 for an example). The goal, typically, is to transform a radically multi-use technology into a more tightly constrained—perhaps effectively single-purpose—tool for use within a specific domain, with the intention of limiting the kinds of harms that it might cause. In our context of use-control, this also makes it a more workable target of regulation. However, there are yet more feasibility issues here: this time not regarding the feasibility of implementing regulation, but of actually designing the technology. First, to what extent does limiting the training data degrade the overall quality of the tool beyond the point of usefulness (since, generally speaking, the unprecedented effectiveness of LLMs stems from their being based on enormous underlying datasets and parameter sets). And second, is it even possible to design such algorithms in a way that truly incorporates strict use-case limits? It is an open question, for example, whether it is in principle possible to craft LLMs that are immune to *prompt injection attacks* (Perez and Ribeiro 2022; Du et al. 2022), which are meant to shift the modality in which the chatbot is being used, often to bypass safeguards.

If we concede that LLMs must, by their nature, be at least partially multi-purpose tools, able to cross domain boundaries, and if moreover we grant that their usefulness to humanity speaks against any attempt to completely ban them, then the prospect of applying our use control framework ultimately depends on creating a meaningful division and taxonomy of uses so that we can track *when* there's a crossing of boundaries that ought to trigger new regulatory oversight. Crucially, in light of the discussion above, this division has to be broad enough that the regulatory oversight isn't constantly triggered (since that would be infeasible), yet fine enough so that individuals and corporations cannot exploit the taxonomy to avoid regulatory oversight and abuse data or information that individuals have disclosed for different purposes.

The discussion above implicitly focuses on cases where the LLMs are providing *factually accurate* responses. Believe it or not, this is the "easy" case! We thus conclude by returning to what is arguably the largest problem of them all, namely problem (3): misinformation, disinformation, and its impact on privacy. As noted, LLMs frequently present falsehoods as if they are truths; they can also be asked explicitly to craft false narratives. At present, this does

not seem to be avoidable, given how the technology actually works: in a very real sense, no LLM can distinguish the truth from a lie—it's essentially just fabricating sentences that seem to best simulate what it finds in its training data in response to a prompt (Bender and Koller 2020).

For instance, in domains like medicine and health, law, or finance, when we imagine a domain-specific or application-constrained LLM—a helpful AI assistant trained to answer questions within that topic area—we must also take into account the fact that these assistants only provide accurate information to the extent that it seems to fit with their training data—when inaccurate statements seem to fit better, that's what it provides. Is there a meaningful notion of use control that accommodates the inevitability of the "uses" including misinformation? To what extent can a regulatory body greenlight a technology that can produce misinformation in an unbounded and poorly understood way?

As with the cases we considered above, this challenge is not unique to the use control framework; many potential mitigations have been suggested. Many of these suggestions revolve around improving in some way the *interpretability* of the algorithm (Wiggers 2023). In theory, this can serve to make it easier to understand the reasons for specific outputs, with the idea being that, providing that insight, then *human* experts (or even adversarial AIs) can then serve as more "fact checkers". While there is some promise to these approaches, they remain at present untested on a large scale. Thus, the prospect of increased interpretability as a way of mitigating misinformation remains murky at best. Even in the best cases, we must accept that LLMs will sometimes output falsehoods, which can harm individuals and potentially have societal damaging impacts as well (e.g., what if LLMs, trained on the falsehood-riddled internet, come to repeat lies about voter fraud in a democratic country?). The challenge remains to craft ethical regulation that respects the privacy concerns of and harms to individuals in light of this.

CONCLUSION

What are the "take home" points from all these arguments? First, technological advancement, specifically in the area of predictive algorithms and AI, has begun to erode the boundary between "guessing" and "observing" (Chapter 1), and this erosion has critical downstream effects on the central moral concept of privacy. We ignore or downplay this at our peril—it can silently yet dramatically increase the extent to which individuals are vulnerable to privacy-related harms. Furthermore, this kind of vulnerability to harm does not require "perfect" predictors: even highly flawed algorithms can cause substantial harm through their *perceived* accuracy; moreover, the *presentational harms* that can be inflicted through such algorithms (Chapter 2) can be just as damaging to individuals when they are not completely accurate.

A central consequence of these observations, given the nature of the predictive algorithms that have reshaped the moral landscape here, is that *data control* approaches to safeguarding privacy are increasingly unviable: already, much of our personal data is easily accessible, and not a lot of it is needed to fuel powerful predictors—a state of affairs that is only going to grow more extreme as time passes. Our contention is that the appropriate response to this is *not* to engage in an arms-race with ever more stringent recommendations for how individuals ought to try to restrict or control access to their personal information, but rather, to shift the focus of this control away from the data itself and onto its potential uses. This also has the crucial benefit of building in the flexibility to accommodate unforeseen future uses of data via new, revolutionary technologies.

Our "data use control" approach goes hand in hand with a concomitant shift away from an individualized notion of consent and towards a more governance based, socially supported framework. While other authors have already raised red flags about highly individualized consent structures, our push for data *use* control, and the analysis of these two shifts in tandem, is (to the best of our knowledge) a novel approach to policy crafting in this area.

Where do we go from here? As with any work in this area, our contribution raises as many questions as it addresses. For example, as we discussed at length (Chapter 3), any systematic implementation of this framework will require a comprehensive "taxonomy of uses" in order to balance freedom to innovate with safeguards against harm. This work has only just begun, and by its nature it is bound to be highly interdisciplinary. On a conceptual front, our distinction between informational and presentational harm, which arises naturally in the context of predictive algorithms, is perhaps deserving of a deeper philosophical inquiry, particularly since we argue that it crucially structures modern conversations about privacy.

On the practical front, there is of course a substantial policy dimension to this work that, we hope, will be relevant to future implementations. While we explicitly describe several potential policy approaches, the impact of the *philosophical ideas* in this thesis may ultimately be the more relevant—this is, after all, a work of philosophy, and the "use" of philosophy in these contexts is to clarify concepts, raise important and overlooked questions, and inform interdisciplinary debates through careful analysis. What does this mean concretely? As new technologies are developed that inevitably have potential impacts on privacy and raise concerns for harms to individuals, corresponding questions will arise about how best to understand and govern these technologies in light of their dangers—what their role in society is, and what responsibilities and risks ought to be borne by the developers, by those who deploy the technology, and by the public at large. We hope that the arguments and ideas presented in this work can play a role in answering such questions, not only in the halls of academia, but in boardrooms, in congress, and even over the dinner table.

Bibliography

Abràmoff, Michael D., Philip T. Lavin, Michele Birch, Nilay Shah, and James C. Folk. 2018. "Pivotal Trial of an Autonomous AI-Based Diagnostic System for Detection of Diabetic Retinopathy in Primary Care Offices." *Npj Digital Medicine* 1 (1): 1–8. https://doi.org/10.1038/s41746-018-0040-6.

Acquisti, Alessandro, Idris Adjerid, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Yang Wang, and Shomir Wilson. 2023. "Nudges (and Deceptive Patterns) for Privacy: Six Years Later." In *The Routledge Handbook of Privacy and Social Media*. Routledge.

Acquisti, Alessandro, and Ralph Gross. 2009. "Predicting Social Security Numbers from Public Data." *Proceedings of the National Academy of Sciences - PNAS*, From the Cover, 106 (27): 10975–80. https://doi.org/10.1073/pnas.0904891106.

Al Mayahi, Khalfan, and Mahmood Al-Bahri. 2020. "Machine Learning Based Predicting Student Academic Success." In *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 264–68. https://doi.org/10.1109/ICUMT51630.2020.9222435.

Allyn, Bobby. 2020. "Your Boss Is Watching You: Work-From-Home Boom Leads To More Surveillance." *NPR*, May 13, 2020. https://www.npr.org/2020/05/13/854014403/your-boss-is-watching-you-work-from-home-boom-leads-to-more-surveillance.

Anand, Priya, and Mark Bergen. 2021. "Big Teacher Is Watching: How AI Spyware Took Over Schools." *Bloomberg.Com*, October 28, 2021. https://www.bloomberg.com/news/features/2021-10-28/how-goguardian-ai-spyware-took-over-schools-student-devices-during-covid.

Araujo, T., N. Helberger, S. Kruikemeier, and C. H. de Vreese. 2020. "In AI We Trust? Perceptions about Automated Decision-Making by Artificial Intelligence." *AI & Society* 35 (3): 611–23. https://doi.org/10.1007/s00146-019-00931-w.

Archard, David. 2008. "Informed Consent: Autonomy and Self-Ownership." *Journal of Applied Philosophy* 25 (1): 19–34. https://doi.org/10.1111/j.1468-5930.2008.00394.x.

"Artificial Neural Network." 2023. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Artificial_neural_network&oldid=1155178993#Applications.

Atske, Sara. 2019. "4. Americans' Attitudes and Experiences with Privacy Policies and Laws." Pew Research Center. https://www.pewresearch.org/internet/2019/11/15/americans-attitudes-and-experiences-with-privacy-policies-and-laws/.

Barocas, Solon, and Helen Nissenbaum. 2014. "Big Data's End Run around Anonymity and Consent." In *Privacy, Big Data, and the Public Good*, edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 1st ed., 44–75. Cambridge University Press. https://doi.org/10.1017/CBO9781107590205.004.

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104 (3): 671–732. https://doi.org/10.15779/Z38BG31.

Basak, Anirban. 2022. "Council Post: Data Toxicity And The Role Of Financial Data Brokers." Forbes. 2022. https://www.forbes.com/sites/forbesbusinesscouncil/2022/10/19/data-toxicity-and-the-role-of-financial-data-brokers/.

Beauchamp, Tom L., and James F. Childress. 2001. *Principles of Biomedical Ethics*. Oxford, UNITED STATES: Oxford University Press, Incorporated. http://ebookcentral.proquest.com/lib/cm/detail.action?docID=5763592.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922.

Bender, Emily M., and Alexander Koller. 2020. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–98. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463.

Berinato, Scott. 2018. "'Stop Thinking About Consent: It Isn't Possible and It Isn't Right.'" *Harvard Business Review*, September 24, 2018. https://hbr.org/2018/09/stop-thinking-about-consent-it-isnt-possible-and-it-isnt-right.

Berk, Richard A. 2021. “Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement.” *Annual Review of Criminology* 4 (1): 209–37. https://doi.org/10.1146/annurev-criminol-051520-012342.

Bernstein, Michael S., Margaret Levi, David Magnus, Betsy A. Rajala, Debra Satz, and Quinn Waeiss. 2021. “Ethics and Society Review: Ethics Reflection as a Precondition to Research Funding.” *Proceedings of the National Academy of Sciences - PNAS* 118 (52): 1-. https://doi.org/10.1073/pnas.2117261118.

Binns, Reuben. 2017. “Fairness in Machine Learning: Lessons from Political Philosophy.” https://doi.org/10.48550/arxiv.1712.03586.

Bloustein, Edward J. 1964. “Privacy as an Aspect of Human Dignity: An Answer to Dean Prosser.” *New York University Law Review* 39: 962.

Boerman, Sophie C., Sanne Kruikemeier, and Frederik J. Zuiderveen Borgesius. 2017. “Online Behavioral Advertising: A Literature Review and Research Agenda.” *Journal of Advertising* 46 (3): 363–76. https://doi.org/10.1080/00913367.2017.1339368.

Bok, Sissela. 2011. *Secrets: On the Ethics of Concealment and Revelation*. Knopf Doubleday Publishing Group.

Boyd, Nora Mills, and James Bogen. 2021. “Theory and Observation in Science.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2021. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/science-theory-observation/.

Bradford Franklin, Sharon, Greg Nojeim, and Dhanaraj Thakur. 2021. “Report - Legal Loopholes and Data for Dollars: How Law Enforcement and Intelligence Agencies Are Buying Your Data from Brokers.” *Center for Democracy and Technology* (blog). December 9, 2021. https://cdt.org/insights/report-legal-loopholes-and-data-for-dollars-how-law-enforcement-and-intelligence-agencies-are-buying-your-data-from-brokers/.

Bradfort Franklin, Sharon, and Dhanaraj Thakur. 2021. “New CDT Report Documents How Law Enforcement & Intel Agencies Are Evading the Law and Buying Your Data from Brokers.” *Center for Democracy and Technology* (blog). December 9, 2021.

https://cdt.org/insights/new-cdt-report-documents-how-law-enforcement-intel-agencies-are-evading-the-law-and-buying-your-data-from-brokers/.

Burrell, Jenna. 2016. “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms.” *Big Data & Society* 3 (1): 205395171562251. https://doi.org/10.1177/2053951715622512.

Cadwalladr, Carole, and Emma Graham-Harrison. 2018. “Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach.” The Guardian. 2018.

Cakebread, Caroline. 2017. “You’re Not Alone, No One Reads Terms of Service Agreements.” *Business Insider*, 2017. https://www.businessinsider.com/deloitte-study-91-percent-agree-terms-of-service-without-reading-2017-11.

Campanella, Thomas J. 2017. “The True Measure of Robert Moses (and His Racist Bridges).” *Bloomberg.Com*, July 9, 2017. https://www.bloomberg.com/news/articles/2017-07-09/robert-moses-and-his-racist-parkway-explained.

Cate, Fred H. 2010. “The Limits of Notice and Choice.” *IEEE Security & Privacy* 8 (2): 59–62. https://doi.org/10.1109/MSP.2010.84.

Certified Humane®. 2021. “All about Egg Labels.” Certified Humane. April 1, 2021. https://certifiedhumane.org/all-about-egg-labels/.

Char, Danton S., Michael D. Abràmoff, and Chris Feudtner. 2020. “Identifying Ethical Considerations for Machine Learning Healthcare Applications.” *The American Journal of Bioethics* 20 (11): 7–17. https://doi.org/10.1080/15265161.2020.1819469.

Chellappa, Ramnath K., and Shivendu Shivendu. 2006. “An Economic Model of Privacy: A Property Rights Approach to Regulatory Choices for Online Personalization.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.457003.

Chen, Yi, Mingming Zha, Nan Zhang, Dandan Xu, Qianqian Zhao, Xuan Feng, Kan Yuan, et al. 2019. “Demystifying Hidden Privacy Settings in Mobile Apps.” In , 570–86. https://doi.org/10.1109/SP.2019.00054.

Chester, Jeff, and Kathryn C. Montgomery. 2017. “The Role of Digital Marketing in Political Campaigns.” *Internet Policy Review* 6 (4): 1–20. https://doi.org/10.14763/2017.4.773.

Christin, Angèle, Alex Rosenblat, and Danah Boyd. 2015. “Courts and Predictive Algorithms.” *DATA & CIVIL RIGHTS: A NEW ERA OF POLICING AND JUSTICE*.

Citron, Danielle Keats, and Daniel J. Solove. 2021. “Privacy Harms.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3782222.

Cohen, Jean-Louis. 2009. “Regulating Intimacy: A New Legal Paradigm.” In *Regulating Intimacy*. Princeton University Press. https://doi.org/10.1515/9781400825035.

Colnago, Jessica, and Hélio Guardia. 2016. “How to Inform Privacy Agents on Preferred Level of User Control?” In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 1542–47. UbiComp ’16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2968219.2968546.

Conway, Mike, and Daniel O’Connor. 2016. “Social Media, Big Data, and Mental Health: Current Advances and Ethical Implications.” *Current Opinion in Psychology*, Social media and applications to health behavior, 9 (June): 77–82. https://doi.org/10.1016/j.copsyc.2016.01.004.

Costello, Kaitlin L., and Diana Floegel. 2020. “‘Predictive Ads Are Not Doctors’: Mental Health Tracking and Technology Companies.” *Proceedings of the Association for Information Science and Technology* 57 (1): e250. https://doi.org/10.1002/pra2.250.

Cox, Joseph. 2019. “I Gave a Bounty Hunter $300. Then He Located Our Phone.” *Vice* (blog). January 8, 2019. https://www.vice.com/en/article/nepxbz/i-gave-a-bounty-hunter-300-dollars-located-phone-microbilt-zumigo-tmobile.

Cranor, Lorrie Faith, Praveen Guduru, and Manjula Arjula. 2006. “User Interfaces for Privacy Agents.” *ACM Transactions on Computer-Human Interaction* 13 (2): 135–78. https://doi.org/10.1145/1165734.1165735.

Crawford, Kate, and Jason Schultz. 2014. “Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms.” *Boston College Law Review* 55: 93.

Cyphers, Bennett, and Gennie Gebhart. 2019. “Behind the One-Way Mirror: A Deep Dive Into the Technology of Corporate Surveillance.” *Electronic Frontier Foundation*.

“Data Brokers.” 2023. *EPIC - Electronic Privacy Information Center* (blog). 2023. https://epic.org/issues/consumer-privacy/data-brokers/.

Dave, Paresh. 2022. “U.S. Cities Are Backing off Banning Facial Recognition as Crime Rises.” *Reuters*, May 12, 2022, sec. Disrupted. https://www.reuters.com/world/us/us-cities-are-backing-off-banning-facial-recognition-crime-rises-2022-05-12/.

Devany, Bonnie E. 2022. “Clearview AI’s First Amendment: A Dangerous Reality?” *Texas Law Review* 101 (2): 473–507.

Diehm, Cadem, Kelsey Smith, Ame Elliott, and Georgia Bullen. 2021. “The Limits to Digital Consent – Simply Secure.” Simply Secure. https://simplysecure.org/blog/the-limits-to-digital-consent-understanding-the-risks-of-ethical-consent-and-data-collection-for-underrepresented-communities/.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. “Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err.” *Journal of Experimental Psychology. General* 144 (1): 114–26. https://doi.org/10.1037/xge0000033.

Diresta, Renee. 2018. “Computational Propaganda: If You Make It Treend, You Make It True.” *The Yale Review* 106 (4): 12–29. https://doi.org/10.1353/tyr.2018.0030.

Doshi-Velez, Finale, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O’Brien, Kate Scott, et al. 2019. “Accountability of AI Under the Law: The Role of Explanation.” arXiv. https://doi.org/10.48550/arXiv.1711.01134.

Drake, Nadia. 2019. “First-Ever Picture of a Black Hole Unveiled.” National Geographic. 2019. https://www.nationalgeographic.com/science/article/first-picture-black-hole-revealed-m87-event-horizon-telescope-astrophysics.

Du, Wei, Yichun Zhao, Boqun Li, Gongshen Liu, and Shilin Wang. 2022. “PPT: Backdoor Attacks on Pre-Trained Models via Poisoned Prompt Tuning.” In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 680–86. Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2022/96.

Dworkin, Gerald. 1988. *The Theory and Practice of Autonomy*. Cambridge Studies in Philosophy. Cambridge: University Press.

Emami-Naeini, Pardis, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. 2021. “Which Privacy and Security Attributes Most Impact Consumers’ Risk Perception and

Willingness to Purchase IoT Devices?" In *2021 IEEE Symposium on Security and Privacy (SP)*, 519–36. https://doi.org/10.1109/SP40001.2021.00112.

———. 2022. "An Informative Security and Privacy 'Nutrition' Label for Internet of Things Devices." *IEEE Security & Privacy* 20 (2): 31–39. https://doi.org/10.1109/MSEC.2021.3132398.

Ereiz, Zoran. 2019. "Predicting Default Loans Using Machine Learning (OptiML)." In *2019 27th Telecommunications Forum (TELFOR)*, 1–4. https://doi.org/10.1109/TELFOR48224.2019.8971110.

Eubanks, Virginia. 2014. "Want to Predict the Future of Surveillance? Ask Poor Communities." The American Prospect. January 15, 2014. https://prospect.org/api/content/36656b9e-c446-5205-9257-0120f64aabdb/.

Forberg, Peter L. 2022. "From the Fringe to the Fore: An Algorithmic Ethnography of the Far-Right Conspiracy Theory Group QAnon." *Journal of Contemporary Ethnography* 51 (3): 291–317. https://doi.org/10.1177/08912416211040560.

Fraser, Colin. 2020. "Target Didn't Figure out a Teen Girl Was Pregnant before Her Father Did." *Medium* (blog). July 16, 2020. https://medium.com/@colin.fraser/target-didnt-figure-out-a-teen-girl-was-pregnant-before-her-father-did-a6be13b973a5.

Fried, Charles. 1970. *An Anatomy of Values: Problems of Personal and Social Choice*. Cambridge, MA: Harvard University Press.

Froomkin, A. Michael. 2015. "Regulating Mass Surveillance as Privacy Pollution: Learning from Environmental Impact Statements." *University of Illinois Law Review* 2015 (5): 1713-.

Future, Fight for the. 2023. "See Where Dangerous Facial Recognition Is Being Used, and Learn What You Can Do about It." Ban Facial Recognition. 2023. https://www.banfacialrecognition.com/map/.

Galvan, Edgar, Joaquin Garcia-Alfaro, Guillermo Navarro-Arribas, and Vicenc Torra. 2021. "Agents in a Privacy-Preserving World." *Transactions on Data Privacy*.

Ganguli, Deep, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, et al. 2022. "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned." arXiv. https://doi.org/10.48550/arXiv.2209.07858.

Gavison, Ruth. 1980. “Privacy and the Limits of Law.” *The Yale Law Journal* 89 (3): 421–71. https://doi.org/10.2307/795891.

Gerety, Tom. 1977. “Redefining Privacy.” *Harvard Civil Rights-Civil Liberties Law Review* 12: 233.

Gerstein, Robert S. 1978. “Intimacy and Privacy.” *Ethics* 89 (1): 76–81. https://doi.org/10.1086/292105.

Global Witness. 2021. “How Facebook’s Ad Targeting May Be in Breach of UK Equality and Data Protection Laws.” Global Witness. 2021. https:///en/campaigns/digital-threats/how-facebooks-ad-targeting-may-be-in-breach-of-uk-equality-and-data-protection-laws/.

Heikkilä, Melissa. 2022. “Artists Can Now Opt out of the next Version of Stable Diffusion.” *MIT Technology Review*, 2022. https://www.technologyreview.com/2022/12/16/1065247/artists-can-now-opt-out-of-the-next-version-of-stable-diffusion/.

Henman, Paul, and Greg Marston. 2008. “The Social Division of Welfare Surveillance.” *Journal of Social Policy* 37 (2): 187–205. https://doi.org/10.1017/S0047279407001705.

Herman, Sean. 2020. “Council Post: Should Tech Companies Be Paying Us For Our Data?” Forbes. 2020. https://www.forbes.com/sites/forbestechcouncil/2020/10/30/should-tech-companies-be-paying-us-for-our-data/.

Hill, Kashmir. 2020. “The Secretive Company That Might End Privacy as We Know It.” *The New York Times*, January 18, 2020, sec. Technology. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

———. 2022. “The Secretive Company That Might End Privacy as We Know It*.” In *Ethics of Data and Analytics*, by Kirsten Martin, 1st ed., 170–77. Boca Raton: Auerbach Publications. https://doi.org/10.1201/9781003278290-26.

Huckins, Jeremy F., Alex W. daSilva, Weichen Wang, Elin Hedlund, Courtney Rogers, Subigya K. Nepal, Jialing Wu, et al. 2020. “Mental Health and Behavior of College Students During the Early Phases of the COVID-19 Pandemic: Longitudinal Smartphone and Ecological Momentary Assessment Study.” *Journal of Medical Internet Research* 22 (6): e20185. https://doi.org/10.2196/20185.

Jurcys, Paulius. 2022. “What Is the Value of Your Data?” Medium. August 18, 2022. https://towardsdatascience.com/what-is-the-value-of-your-data-9341cd019b4d.

Kastrenakes, Jacob. 2023. “Adobe Made an AI Image Generator — and Says It Didn’t Steal Artists’ Work to Do It.” *The Verge*, March 21, 2023. https://www.theverge.com/2023/3/21/23648315/adobe-firefly-ai-image-generator-announced.

Kate Crawford, dir. 2017. *The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford #NIPS2017*. The Artificial Intelligence Channel. https://www.youtube.com/watch?v=fMym_BKWQzk.

Kaushik, Keshav, Akashdeep Bhardwaj, Ashutosh Dhar Dwivedi, and Rajani Singh. 2022. “Machine Learning-Based Regression Framework to Predict Health Insurance Premiums.” *International Journal of Environmental Research and Public Health* 19 (13): 7898. https://doi.org/10.3390/ijerph19137898.

Kelly, Mary Louise, Enrique Rivera, and Christopher Intagliata. 2022. “More Kids Are Going Back to School. So Why Is Laptop Surveillance Increasing?” *NPR*, August 17, 2022, sec. Education. https://www.npr.org/2022/08/17/1118009553/more-kids-are-going-back-to-school-so-why-is-laptop-surveillance-increasing.

Kerry, Cameron F., and John B. Morris. 2019. “Why Data Ownership Is the Wrong Approach to Protecting Privacy.” *Brookings* (blog). June 26, 2019. https://www.brookings.edu/blog/techtank/2019/06/26/why-data-ownership-is-the-wrong-approach-to-protecting-privacy/.

Kitkowska, Agnieszka, Johan Högberg, and Erik Wästlund. 2022. “Online Terms and Conditions: Improving User Engagement, Awareness, and Satisfaction through UI Design.” In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–22. CHI ’22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3491102.3517720.

Klöpper, Miriam, and Sonja Köhne. 2023. “Shifting Structures - A Systematic Literature Review on People Analytics and the Future of Work.” *ECIS 2023 Research Papers*, May. https://aisel.aisnet.org/ecis2023_rp/360.

Kumar, Yogesh, Komalpreet Kaur, and Gurpreet Singh. 2020. “Machine Learning Aspects and Its Applications Towards Different Research Areas.” In *2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, 150–56. https://doi.org/10.1109/ICCAKM46823.2020.9051502.

Laird, Elizabeth, Hugh Grant-Chapman, Cody Venzke, and Hannah Quay-de la Vallee. 2022. “Hidden Harms: The Misleading Promise of Monitoring Students Online.” Center for Democracy and Technology. https://cdt.org/wp-content/uploads/2022/08/Hidden-Harms-The-Misleading-Promise-of-Monitoring-Students-Online-Research-Report-Final-Accessible.pdf.

Larsson, Stefan, Anders Jensen-Urstad, and Fredrik Heintz. 2021. “Notified But Unaware: Third-Party Tracking Online.” *Critical Analysis of Law* 8 (1): 101–20.

Leon, Pedro Giovanni, Lorrie Faith Cranor, Aleecia M. McDonald, and Robert McGuire. 2010. “Token Attempt: The Misrepresentation of Website Privacy Policies through the Misuse of P3p Compact Policy Tokens.” In *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society*, 93–104. WPES ’10. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1866919.1866932.

Madden, Mary, Michele Gilman, Karen Levy, and Alice Marwick. 2017. “Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans.” *Washington University Law Review* 95: 53.

Mahmoud, Ali A., Tahani AL Shawabkeh, Walid A. Salameh, and Ibrahim Al Amro. 2019. “Performance Predicting in Hiring Process and Performance Appraisals Using Machine Learning.” In *2019 10th International Conference on Information and Communication Systems (ICICS)*, 110–15. https://doi.org/10.1109/IACS.2019.8809154.

Marcinkowski, Frank, Kimon Kieslich, Christopher Starke, and Marco Lünich. 2020. “Implications of AI (Un-)Fairness in Higher Education Admissions: The Effects of Perceived AI (Un-)Fairness on Exit, Voice and Organizational Reputation.” In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 122–30. FAT* ’20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372867.

McDonald, Aleecia M., and Lorrie Faith Cranor. 2008. “The Cost of Reading Privacy Policies.” *I/S: A Journal of Law and Policy for the Information Society* 4: 543.

“Meat Department Quality Standards.” n.d. Whole Foods Market. Accessed May 20, 2023. https://www.wholefoodsmarket.com/quality-standards/meat-standards.

Microsoft AI. 2023. “Our Approach to Responsible AI at Microsoft.” Microsoft. 2023. https://www.microsoft.com/en-us/ai/our-approach.

Miller, Franklin G., and Alan Wertheimer. 2010. *The Ethics of Consent Theory and Practice*. New York ; Oxford University Press.

Nissenbaum, Helen. 2009. “Privacy in Context: Technology, Policy, and the Integrity of Social Life.” In *Privacy in Context*. Stanford University Press. https://doi.org/10.1515/9780804772891.

Nussbaum, Martha. 1999. “Women and Equality: The Capabilities Approach.” *International Labour Review* 138 (3): 227–45. https://doi.org/10.1111/j.1564-913X.1999.tb00386.x.

Obar, Jonathan A., and Anne Oeldorf-Hirsch. 2020. “The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services.” *Information, Communication & Society* 23 (1): 128–47. https://doi.org/10.1080/1369118X.2018.1486870.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” *Science* 366 (6464): 447–53. https://doi.org/10.1126/science.aax2342.

Oltramari, Alessandro, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B. Norton, N. Cameron Russell, Peter Story, Joel Reidenberg, and Norman Sadeh. 2018. “PrivOnto: A Semantic Framework for the Analysis of Privacy Policies.” Edited by Mathieu d’Aquin, Sabrina Kirrane, Serena Villata, Mathieu d’Aquin, Sabrina Kirrane, and Serena Villata. *Semantic Web* 9 (2): 185–203. https://doi.org/10.3233/SW-170283.

O’Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. USA: Crown Publishing Group.

Passi, Samir, and Solon Barocas. 2019. “Problem Formulation and Fairness.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 39–48. FAT* ’19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287567.

Perera, Thenuka, and Theja Perera. 2021. “Barrister-Processing and Summarization of Terms & Conditions / Privacy Policies.” In *2021 6th International Conference for Convergence in Technology (I2CT)*, 1–7. https://doi.org/10.1109/I2CT51068.2021.9418090.

Perez, Fábio, and Ian Ribeiro. 2022. “Ignore Previous Prompt: Attack Techniques For Language Models.” arXiv. https://doi.org/10.48550/arXiv.2211.09527.

Piatetsky, Gregory. 2014. “Did Target Really Predict a Teen’s Pregnancy? The Inside Story.” *KDnuggets* (blog). 2014. https://www.kdnuggets.com/did-target-really-predict-a-teens-pregnancy-the-inside-story.html.

Rachels, James. 1975. “Why Privacy Is Important.” *Philosophy & Public Affairs* 4 (4): 323–33.

Rainie, Lee, and Janna Anderson. 2014. “The Future of Privacy.” *Pew Research Center: Internet, Science & Tech* (blog). December 18, 2014. https://www.pewresearch.org/internet/2014/12/18/future-of-privacy/.

Raji, Inioluwa Deborah, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. “AI and the Everything in the Whole Wide World Benchmark.” arXiv. https://doi.org/10.48550/arXiv.2111.15366.

Raymond, Nathaniel A. 2017. “Beyond ‘Do No Harm’ and Individual Consent: Reckoning with the Emerging Ethical Challenges of Civil Society’s Use of Data.” In *Group Privacy: New Challenges of Data Technologies*, edited by Linnet Taylor, Luciano Floridi, and Bart van der Sloot, 67–82. Philosophical Studies Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46608-8_4.

Roberts, Michael, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, et al. 2021. “Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans.” *Nature Machine Intelligence* 3 (3): 199–217. https://doi.org/10.1038/s42256-021-00307-0.

Robertson, Viktoria H. S. E. 2019. “Excessive Data Collection: Privacy Considerations and Abuse of Dominance in the Era of Big Data.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3408971.

Rose, E. 2005. “Data Users versus Data Subjects: Are Consumers Willing to Pay for Property Rights to Personal Information?” In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 180c–180c. https://doi.org/10.1109/HICSS.2005.184.

Rudin, Cynthia. 2019. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” *Nature Machine Intelligence* 1 (5): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Schwartz, Paul M., and Daniel J. Solove. 2014. “Reconciling Personal Information in the United States and European Union.” *California Law Review* 102: 877.

Selbst, Andrew D. 2013. “Contextual Expectations of Privacy.” *Cardozo Law Review* 35: 643.

———. 2017. “DISPARATE IMPACT IN BIG DATA POLICING.” *Georgia Law Review (Athens, Ga. : 1966)* 52 (1): 109-.

Singer, Natasha, and Jason Karaian. 2023. “Americans Flunked This Test on Online Privacy.” *The New York Times*, February 7, 2023, sec. Technology. https://www.nytimes.com/2023/02/07/technology/online-privacy-tracking-report.html.

Skitka, Linda J., Kathleen Mosier, and Mark D. Burdick. 2000. “Accountability and Automation Bias.” *International Journal of Human-Computer Studies* 52 (4): 701–17. https://doi.org/10.1006/ijhc.1999.0349.

Solove, Daniel J. 2002. “Conceptualizing Privacy.” *California Law Review* 90 (4): 1087–1155. https://doi.org/10.2307/3481326.

———. 2005. “A Taxonomy of Privacy.” SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=667622.

———. 2006. “A Brief History of Information Privacy Law.” SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=914271.

Solove, Daniel J. 2013. “Introduction: Privacy Self-Management and the Consent Dilemma.” *Harvard Law Review* 126 (7): 1880–1903.

Sonnemaker, Tyler. 2021. “Amazon Is Deploying AI Cameras to Surveil Delivery Drivers ‘100% of the Time.’” *Business Insider*, 2021. https://www.businessinsider.com/amazon-plans-ai-cameras-surveil-delivery-drivers-netradyne-2021-2.

Spadafora, Anthony. 2021. “Many Americans Aren’t Aware They’re Being Tracked with Facial Recognition While Shopping.” TechRadar. August 12, 2021. https://www.techradar.com/news/many-americans-arent-aware-theyre-being-tracked-with-facial-recognition-while-shopping.

Steel, Emily, Callum Locke, Emily Cadman, and Ben Freese. 2013. “How Much Is Your Personal Data Worth?” Financial Times. June 12, 2013. https://ig.ft.com/how-much-is-your-personal-data-worth/.

Tondel, Inger Anne, Åsmund Ahlmann Nyre, and Karin Bernsmed. 2011. “Learning Privacy Preferences.” In *2011 Sixth International Conference on Availability, Reliability and Security*, 621–26. https://doi.org/10.1109/ARES.2011.96.

Turow, Joseph, Yphtach Lelkes, Nora Draper, and Ari Ezra Waldman. 2023. “Americans Can’t Consent to Companies’ Use of Their Data: They Admit They Don’t Understand It, Say They’re Helpless to Control It, and Believe They’re Harmed When Firms Use Their Data--Making What Companies Do Illegitimate.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4391134.

Ur, Blase, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. “Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising.” In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, 1–15. SOUPS ’12. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2335356.2335362.

Wang, Yilun, and Michal Kosinski. 2018. “Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images.” *Journal of Personality and Social Psychology* 114 (2): 246–57. https://doi.org/10.1037/pspa0000098.

Warren, Samuel D., and Louis D. Brandeis. 1890. “The Right to Privacy.” *Harvard Law Review* 4 (5): 193. https://doi.org/10.2307/1321160.

Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, et al. 2021. “Ethical and Social Risks of Harm from Language Models.” arXiv. https://doi.org/10.48550/arXiv.2112.04359.

Wiggers, Kyle. 2023. “OpenAI’s New Tool Attempts to Explain Language Models’ Behaviors.” *TechCrunch* (blog). May 9, 2023. https://techcrunch.com/2023/05/09/openais-new-tool-attempts-to-explain-language-models-behaviors/.

Winner, Langdon. 1980. “Do Artifacts Have Politics?” *Daedalus* 109 (1): 121–36.

Wolfe, Cameron R. 2023. “Specialized LLMs: ChatGPT, LaMDA, Galactica, Codex, Sparrow, and More.” Medium. January 13, 2023. https://towardsdatascience.com/specialized-llms-chatgpt-lamda-galactica-codex-sparrow-and-more-ccccdd9f666f.

Zickuhr, Kathryn. 2021. “Workplace Surveillance Is Becoming the New Normal for U.S. Workers.” *Washington Center for Equitable Growth* (blog). 2021. https://equitablegrowth.org/research-paper/workplace-surveillance-is-becoming-the-new-normal-for-u-s-workers/.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. First edition. New York: PublicAffairs.