Measuring Automated Influence: Between Empirical Evidence and Ethical Values

Daniel Susser
College of Information Sciences and Technology
& Rock Ethics Institute
Penn State University
University Park, PA, USA
daniel.susser@psu.edu

Vincent Grimaldi
College of Information Sciences and Technology
Penn State University
University Park, PA, USA
vincentgrimaldi@psu.edu

ABSTRACT

Automated influence, delivered by digital targeting technologies such as targeted advertising, digital nudges, and recommender systems, has attracted significant interest from both empirical researchers, on one hand, and critical scholars and policymakers on the other. In this paper, we argue for closer integration of these efforts. Critical scholars and policymakers, who focus primarily on the social, ethical, and political effects of these technologies, need empirical evidence to substantiate and motivate their concerns. However, existing empirical research investigating the effectiveness of these technologies (or lack thereof), neglects other morally relevant effects—which can be felt regardless of whether or not the technologies "work" in the sense of fulfilling the promises of their designers. Drawing from the ethics and policy literature, we enumerate a range of questions begging for empirical analysis—the outline of a research agenda bridging these fields—and issue a call to action for more empirical research that takes these urgent ethics and policy questions as their starting point.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Human-centered computing** → *HCI design and evaluation methods.*

KEYWORDS

influence; targeted advertising; dark patterns; nudges; ethics; law and policy; privacy; autonomy

ACM Reference Format:
Daniel Susser and Vincent Grimaldi. 2021. Measuring Automated Influence: Between Empirical Evidence and Ethical Values. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3461702.3462532

AIES '21, May 19–21, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462532

1 INTRODUCTION

Researchers, policymakers, and activists have raised the alarm about the harms of automated influence. Framed by their developers as tools for delivering personalized digital experiences tailored to the preferences and desires of individual users, targeted advertising, digital nudges, recommender systems, and other influence technologies, threaten—at the same time—to extract higher rents [14], unjustly discriminate [71], mislead and polarize [64], and interfere in both individual and collective decision-making [112]. Frequently driven by artificial intelligence and machine learning algorithms (AI/ML), such technologies are especially concerning because automation enables influence that is simultaneously individually tailored and massive in scale [87].

Shoshana Zuboff has made what is perhaps the most urgent case against these forms of influence, arguing that the firms building them have fundamentally re-shaped the economy into what she terms "surveillance capitalism," and have enacted a new form of power—"instrumentarian power"—which, she claims, utilizes digital influence tools to engage in large-scale "behavior modification" [112]. Whether or not one accepts Zuboff's whole story, the suspicion at its center—that digital influence threatens cherished individual and social values—is widely shared. Others warn about the potential harms of "digital market manipulation" [14], "online manipulation" [87], an "emergent limbic media system" [20], and a "weaponized digital influence machine" [64]. Which is to say, for many observers the influences digital technologies enable raise deep social, ethical, and political questions.

Not everyone is worried, though. Predictably, industry actors contend that these technologies simply respond to consumer demand. Regarding targeted advertising, for example, Facebook CEO Mark Zuckerberg has publicly argued that "People consistently tell us that if they're going to see ads, they want them to be relevant. That means we need to understand their interests" [113]. Putting aside the fact that survey research by prominent privacy scholars casts doubt on these claims about public opinion, Zuckerberg's argument implies that if there is any downside to these tools it is worth it, on balance, to consumers who consent (at least tacitly) to data collection, targeting, and related digital practices when they accept Facebook and other platforms' privacy policies and terms of service [96]. But the nature and significance of these costs are not yet well understood, and when consumers are informed about how targeted advertising technologies work they have tended to register significant "discomfort" with it, especially in certain domains, such as political advertising [95].

Interestingly, some critical scholars and industry watchers are equally unconcerned, arguing that the above debates are beside the point, because digital influence tools don't even work. On this view, concerns about digital influence reflect an uncritical acceptance—by researchers and policymakers—of the technology industry's own claims regarding the effectiveness of its products. These criticisms are important, if only because they encourage more careful, reflexive scholarship. However, they tend to be more responsive to unhelpful straw men derived from popular media (such as the idea, perpetuated by the Netflix film *The Social Dilemma*, that the threat of targeted advertising is a kind of "mind control"), than to the actual arguments advanced in ethics research. More often than not, one encounters these arguments informally, during discussions at conferences and on social media, rather than in peer-reviewed publications and other traditional venues for disseminating research. But some published work advancing this type of argument is beginning to emerge [23, 38]. Usually, the digital influence technology these critics question is targeted advertising—whether they believe other influence technologies, such as recommender systems and "digital nudges," are also unconcerning is unclear.

What both these worries and responses demonstrate is the need for robust empirical research investigating digital influence. Theoretical work has been valuable for identifying, defining, and motivating these problems, but to marshal a meaningful response policymakers and the public need to know more about the actual capabilities of these technologies, and about the nature, range, and scope of their effects. Here, there is good news and bad news. The good news is a large body of related empirical research, spanning multiple disciplines, already exists. The bad news is the findings tell a mixed and context-specific story. Measuring influence is extremely difficult—people make choices for any number of reasons, and isolating the effects of particular influences, especially outside experimental contexts, is a challenge. Rarely able to directly detect "effectiveness," researchers instead explore a variety of proxies. They find that some influence technologies work, to some extent, in some contexts. Others, in other contexts, don't.

More significantly, many of the ethical concerns critical scholars raise are independent of questions about effectiveness. Which is to say, some highly effective automated influence technologies might be perfectly acceptable, while others could be harmful even if they don't work as promised. Influence, on its face, is neither good nor bad—we are all unavoidably influenced in myriad ways. What critical scholars worry about is *how* technologies influence people and *whose interests* such influence serves. For empirical research to advance our understanding of the ethics of automated influence, it needs to investigate the specific issues ethics and policy scholars identify.

In this paper, we further these discussions in two ways. First, we explore the empirical literature on a range of influence technologies to provide a broad overview of the landscape of automated influence, some of the ways empirical scholars have tried to measure the effectiveness of these technologies, and a sense of their findings. Given the breadth and depth of this literature, our analysis is not comprehensive; still, we are able to extract a number of useful insights.[1] Second, we discuss ethical concerns critical scholars and policymakers have raised about automated influence, identify empirical questions they generate, and highlight research that has begun to investigate these questions. In doing so, we hope to inspire a kind of call to action, motivating new empirical research that speaks more directly to the urgent ethics and policy discussions currently underway.

The paper proceeds as follows. In Section 2, we discuss three forms of automated influence: targeted advertising, digital nudging, and recommender systems. We situate these technologies in illustrative application domains—e.g., commercial and political advertising, health and sustainability nudges, and dark patterns. And we explore empirical research that investigates these technologies, with an eye toward the concrete questions researchers are asking. In Section 3, we turn to the ethics and policy literatures. We discuss worries about privacy, autonomy, economic and epistemic harms. Finally, in Section 4, we begin to operationalize these worries in terms of measurable empirical questions, sketching a research agenda that could deepen our understanding of these problems and, perhaps, guide our responses.

2 THE LANDSCAPE OF AUTOMATED INFLUENCE

We are influenced by technology in uncountably many ways, many unintentional. Our focus in this paper is not on the totality of digital influence; rather, our aim is to map technologies designed *intentionally* to influence, and to ask how the effects of such influence are and might be measured. More specifically, we draw attention to the question of *automated influence*—the use of technologies driven by machine learning (ML) or artificial intelligence (AI) to individually tailor influence strategies at scale [76].

"Artificial intelligence" (AI) is a famously slippery concept, often deployed more for the sake of marketing than taxonomic precision. For our purposes, what is significant about the various computing strategies usually collected under the "AI" umbrella (including machine learning) is that they facilitate the real-time automation of influence. By "influence," we mean intervening to change a person's beliefs, desires, emotions, and/or behaviors [87]. Thus, automated influence technologies are technologies that automate the complicated work of changing what people think, want, feel, and do. They are, as Latour might say, a means of "delegating" this work of influencing to machines [48].

The term "automated influence" will conjure, for many readers, images of targeted advertising. And not without reason—advertisers have always been in the influence business, and the promise of applying new technologies to the problem of selling things has been a significant driver of the technology industry's interest in this field [73, 105]. But targeted advertising is not the only technology designed to influence our behavior, and inducing people to buy things is not the only end automated influence serves. Automated influence takes on a variety of forms, is found in a range of application domains, and seeks to further a number of different ends.

[1] This section draws on a more comprehensive, systematic literature review, currently underway.

In this section, we begin to map this terrain. We identify three types of automated influence: targeted advertising, digital nudges, and recommender systems. We describe various contexts they are applied in and ends they are designed to serve—e.g., commerce, health, environmental sustainability, and politics. We examine the kinds of questions empirical scholars tend to ask about these technologies, and how they conceptualize and measure their effectiveness.

2.1 Targeted Advertising

"Targeted advertising" (sometimes referred to as "behavioral advertising," "online behavioral advertising," or "personalized online advertising") designates a range of technologies that tailor advertisements to individuals, either by matching them with particular ads ("segmenting" audiences), or by shaping ads to suit them. While the language of "advertising" suggests the familiar display ad—text, images, or videos appearing alongside search results and other online content—targeted advertising comes in many varieties, including so-called "native ads" (advertisements designed to blend in with non-advertising content), targeted marketing emails, and other digital appeals.

No different, in principle, from television, magazine or billboard advertising, the special promise of targeted advertising is its precision [17]. By collecting and analyzing vast amounts of information about each of us—our needs, desires, preferences, and interests, our spending power, our previous purchasing behaviors and the purchasing behaviors of people like us—targeted advertising attempts to deliver just the right message to just the right person at just the right time to maximize its influence [41]. As Joseph Turow writes, "Advertisers in the digital space expect all media firms to deliver to them particular types of individuals—and, increasingly, *particular* individuals—by leveraging a detailed knowledge about them and their behaviors that was unheard of even a few years ago" [94, p. 12, emphasis in original].

Ad targeting technologies range in sophistication. Contextual advertisements simply target people based on the content they are viewing. Google search results, for example, are often accompanied by ads targeted to the search terms users enter [111]. Behavioral advertising, by contrast, leverages more—and more *specific*—data, aggregated in consumer "profiles," to configure the ads each person sees. Profiles combine demographic information (e.g., age, gender, and income) with information about each person's actual online behavior (e.g., web browsing history, search history, and online shopping history) to deliver advertisements that appeal to specific individuals, rather than to the content they are viewing [7]. Promising even greater precision—and effectiveness—"psychographic profiling" involves advertisements tailored not just to demographic and behavioral information, but to a person's specific *psychological traits*. For example, an advertisement might appeal to "consensus" ("This product is a best seller!"), "authority," ("Recommended by experts!"), or "scarcity" ("Only one left in stock!"), depending on which is predicted to best align with the target's psychological dispositions [41, 44]. These techniques have been used in a range of application domains, most prominent among them e-commerce and politics.

2.1.1 Commercial Advertising. Obviously, a central purpose of targeted advertising is to influence consumer behavior—to sell things—and researchers have long tried to measure how well it works [99]. Measuring real-world purchasing behavior is difficult, because advertising platforms, such as Facebook and Google, carefully guard user data [17, 38]. What research exists suggests that targeted advertisements are relatively more effective than non-targeted ads—one prominent study, for example, found that advertisements targeted to psychographic traits are nearly 50% more effective than generic ads [57]. In absolute terms, the effects were extremely modest, with even targeted advertisements only leading to "conversion"—i.e., purchases—less than 1% of the time. Given the scale and reach automation enables, however, such figures could be misleading: 1% of millions or billions of viewers is a lot. The advertisements in this experiment reached over 3.5 million people.

Lacking the data required to measure real-world purchases, many researchers rely on lab experiments or survey instruments instead, and attempt to measure advertising effectiveness indirectly via proxies, such as "click-through rates" (i.e., how often someone presented with an ad clicks on it) [5, 26, 43, 50], self-reported *intention* to purchase [40, 98], or how much subjects are willing to pay for a given product [42]. Some utilize eye-tracking technology to measure how long a display ad commands a lab subject's attention, suggesting that the more time someone spends looking at an ad the more effective it is [53]. These studies reach a range of conclusions—from little evidence of targeting effectiveness, to evidence of significant effects, to finding that targeting can have *negative* effects, backfiring when users perceive it as too intrusive.

2.1.2 Political Advertising. People are increasingly aware that targeted advertising technologies have moved beyond the commercial sphere and entered the realm of politics. Scandals, such as the Facebook/Cambridge Analytica affair, have brought the stakes of this movement to light.[2] Just as targeted ads in the consumer realm are best understood as an evolution of older commercial advertising practices, so too is digital political advertising an old practice in new form [49]. But more data, increasingly sophisticated statistical modeling techniques, and greater access to individuals via digital platforms mean political campaigns can deliver more precisely tailored digital advertisements than ever before [92].

As in the case of commercial advertising, many empirical scholars have sought to determine just how effective these advertisements are. Since the intention is to influence political behavior, rather than consumer behavior, however, "effectiveness" is conceptualized differently in this context. The objective of targeted political advertisements is generally either (1) to increase voter support for a candidate or position, or (2) to increase voter participation rates amongst existing supporters. To determine how effective political advertisements are at achieving these objectives, political scientists turn to a variety of empirical strategies. Again, the results are mixed.

Some researchers have explored these questions in experimental settings. For example, Zarouali et al. developed a mock social media platform to test whether psychographically targeted political ads were more persuasive than generic ones, and find that to a significant degree they were [109]. Some attempt to measure such effects

[2] See, e.g., https://www.theguardian.com/news/series/cambridge-analytica-files.

more directly: Valenzuela and Michelson use randomized control experiments to measure the impacts of tailored messages on voter turnout (finding that certain kinds of identity appeals do increase turnout), and validate the results using real-world turnout data [97]. Others attempt to measure these phenomena indirectly, relying on survey instruments that inquire about voter intention. Gerber et al. demonstrate that certain kinds of psychographic targeting can increase expressions of intention to vote [31]. And Dixon et al. show that targeted appeals can increase expressions of support for policies to combat climate change [22].

By contrast, however, Krotzek finds that while targeted appeals can improve the way people feel about a candidate, they do not increase the likelihood that they will express an intention to vote for them [45]. Moreover, others find that *mis*targeting—i.e., delivering targeted appeals that are incongruous with a voter's identity—can lead to voters "penalizing" the candidate. Hersh and Schaffner, for example, find that ads targeted at born-again Christians can increase their support for the advertised candidate, but when the same ads are delivered to non-born-again Christians they decreased the target's support for the candidate by an even more significant margin [36]. Flores and Coppock demonstrate similar effects using advertisements tailored to English- versus Spanish-speaking audiences [28].

2.2 Digital Nudges

A second form of automated influence that has generated considerable interest amongst both researchers and policymakers is digital nudging [16, 102]. Following Thaler and Sunstein, nudges are "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" [89]. The insight behind nudging is that decision-making is deeply influenced by decision-making *contexts*—by the way options are presented and arranged. Thus, by making subtle changes to these contexts, or "choice architectures," it is possible to influence the choices people make.

Thaler and Sunstein describe a variety of nudges designed to gently coax people toward individually or socially beneficial decisions, such as placing healthy foods at eye-level in cafeterias while placing junk foods slightly out of reach (to encourage healthy eating), enrolling people in social programs by default and allowing them to opt out, rather than the reverse (to encourage, for example, saving for retirement), and so on. While these examples are relatively low-tech, there are digital parallels. BJ Fogg conceptualized digital systems as "persuasive technologies,"[3] and called for increased attention to "the design, research, and analysis of interactive computing products created for the purposes of changing people's attitudes and behaviors" [29]. Unlike brick-and-mortar cafeterias, digital environments are *personalizable*—they are "adaptive choice architectures" that can be tailored to each individual user [85]. Leveraging the same data-rich profiles advertisers use, the designers of digital systems can deploy "nimble, unobtrusive and highly potent" digital nudges, which Karen Yeung calls "hypernudges" [107].

Like their analog counterparts, digital nudges come in many different forms and are designed to serve a number of different ends. We illustrate by way of two examples: (1) personalized health and sustainability nudges and (2) dark patterns.

2.2.1 Health and Sustainability Nudges. Smartphones, wearables (e.g., fitness trackers and smart watches), smart home devices, and other internet of things (IoT) technologies have become pervasive, enabling both constant monitoring of our activity and delivery of personalized nudges designed to guide it. Toward what ends is an open question [80]. Two application domains where research and development around digital nudges have drawn significant enthusiasm and attention are health and sustainability [61].

Health nudges promise to help people eat less, exercise more, stop smoking, and take their medicine. Like a health professional—say, a nutritionist or exercise coach—who is always on hand to offer advice and encouragement tailored to each person's needs, desires, and tendencies, these automated influence technologies attempt to deliver nudges optimized for each individual's specific health conditions and attuned to the particular kinds of interventions they are likely to find most persuasive. Importantly, automated personalization of nudges means interventions can target not just broad and "nonmodifiable" demographic characteristics, such as age, gender, and socioeconomic status, but more specific and predictive "modifiable" factors, such as available social supports, stress levels, and health literacy [82].

Research exploring the effectiveness of these techniques is wide-ranging, utilizing a variety of proxies to measure success. Using survey instruments, some researchers point to "end-user appreciation" (i.e., stated preference) of targeted interventions over generic ones as evidence of effectiveness [100]. More sophisticated, lab-based randomized control trials analyzing a combination of automatically detected and user-logged diet and exercise data, find preliminary evidence that personalized interventions increase exercise [75]. A review of studies investigating the effects of tailored messaging on user-reported physical activity finds some evidence that tailored messages yielded better results than generic messages [47], as did a review of studies that measured the effects of tailored diet interventions on actual weight loss and weight maintenance [2].

Digital nudges designed to encourage more environmentally-friendly behavior have also attracted interest. Researchers have explored how effectively digital nudges can curtail aggressive driving (a significant contributor to excessive fuel consumption) [8, 30, 83, 93], encourage adherence to detour suggestions that reduce fuel consumption [106], and reduce car usage overall, in favor of more environmentally sustainable options [9], finding in many cases that such nudges generate meaningful behavior change.

2.2.2 Dark Patterns. Finally, designers, researchers, and policymakers have become increasingly concerned about a more troubling approach to constructing digital environments known as "dark patterns." These are user interface design strategies that "knowingly confuse users, make it difficult for users to express their actual preferences, or manipulate users into taking certain actions" [51], or alternatively, cases where "designers use their knowledge of human behavior (e.g., psychology) and the desires of end users to implement deceptive functionality that is not in the user's best interest" [34]. A well-known example is free trial memberships that require credit card information up front, and then automatically convert to paid memberships at the end of the trial period, taking

[3] Though many might be more aptly called "manipulative technologies."

advantage of the fact that many will likely forget to cancel. Harry Brignull, who coined the term dark patterns, calls this "forced continuity" [10]. Or consider retail websites that nudge shoppers into making quicker (and thus less deliberative) decisions by displaying countdown timers or misleading stock reports, like "Only 1 left!"

Dark patterns emerged, Narayanan et al. argue, from the integration of nudge research with long-standing deceptive retail practices, and their adoption has been driven by the technology industry's imperative to achieve market growth at all costs [66]. Although there is only limited empirical research on dark patterns, existing findings are illuminating. To understand the scale of these practices, researchers at Princeton University and the University of Chicago analyzed 11K shopping websites and discovered 1,818 individual instances of dark patterns [55]. And in what appears to be the only controlled experiment yet conducted to test the effectiveness of dark patterns, researchers recruited unsuspecting experimental subjects into an elaborate, large-scale simulation of dark patterns tactics, finding that common dark patterns can significantly impact people's choices [51]. Importantly, dark patterns have been deployed in a variety of application domains beyond e-commerce. Researchers have identified dark patterns designed to encourage users to disclose personal information ("privacy dark patterns") [13], dark patterns in games [108], and dark patterns in human-robot interaction [46].

2.3 Recommender Systems

Both targeted advertising and digital nudges are intended primarily to influence. That's their purpose. Recommender systems differ in that they are designed, first and foremost, to help people sort through the vast quantities of information, entertainment, products, and services available online. Recommender systems evolved "in parallel with the web," becoming increasingly important for navigating it as information and other content proliferated online [6]. Today, they are one of the central mechanisms for organizing online content, sorting everything from Google search results to Amazon product suggestions, Facebook timelines to New York Times news feeds, Spotify playlists to Netflix movie recommendations [17]. But *which* items are surfaced and *how* they are sorted—which websites are indexed by a search engine, for example, and how product recommendations are arranged on a page—can powerfully influence the people who rely on these systems to navigate digital life [39].

People might assume that the recommender systems organizing their social media feeds, surfacing interesting news articles, and serving up streaming TV shows are guided by some objective, measurable conception of relevance. And while that was more or less true in the early years of recommender systems, things have since changed behind-the-scenes. According to anthropologist Nick Seaver, the effectiveness of recommender systems was originally measured against a straightforward metric called root mean squared error (RMSE): "a recommender system predicts how users will rate items, and it is judged by how accurate its predictions were" [81]. In other words, an "effective" recommender system was defined as one that successfully delivered what a user actually wanted to read or watch on TV. Over time, however, as the business model of the internet changed, and as researchers found it increasingly difficult to improve predictive accuracy, "RMSE was dethroned as the paradigmatic measure of success" and replaced by measures of *engagement*—a successful recommendation defined as one that keeps user attention on the platform [81].

Of course, there is nothing wrong, in principle, with trying to engage people. The reason this history is relevant to the present discussion is that it illustrates a potential disjuncture between the real and perceived organizing principles behind many content recommender systems. Technology companies describe these systems, and users perceive them—if they notice them at all[4]—as engines of user satisfaction that deliver to people the content they want. When in reality, these systems may be trying to influence them.

Many product recommender systems operate according to a similar logic [19]. When someone visits Amazon's website they may perceive its product recommendations as neutral, objective suggestions. In fact, Amazon prioritizes its own private label products, encouraging shoppers to buy the items most profitable to the company [24]. Google does the same when presenting its search results [68]. Both companies face regulatory scrutiny because of the perceived unfairness of these practices [32]. A report by the UK government's Centre for Data Ethics and Innovation argues that, "As with personalised advertising, recommendation systems can be optimised to serve different business goals" [17].

Empirically measuring the effectiveness of recommender systems thus requires, first, clarifying what goals—*whose* goals—the system is trying to accomplish. Are the content recommendations made by a social media algorithm "effective" when they are maximally engaging (keeping user eyeballs glued to screens), or when they deliver information and entertainment users find useful, enriching, and informative? Are "effective" product recommender systems those that maximize profits or those that help users find and purchase the things they most want, need, and value?

3 THE ETHICS OF INFLUENCE

To be influenced is neither good nor bad—it is a pervasive, unavoidable condition. Every person's beliefs, desires, emotions, and behavior are shaped by incalculably many influences, some (like those discussed here) intentionally inflicted, others the product of accident or circumstance. We can be influenced in ways that make us (individually and collectively) better or worse off. And being an independent decision-maker—which many view as a normative ideal—does not mean being free from influence; it means understanding, more or less, how we are influenced, being able to critically reflect on the reasons motivating us, and ultimately endorsing our choices.

From this vantage point, some automated influence technologies are obviously nefarious. Dark patterns, for example, as their name suggests, seem indefensible on any imaginable ethical grounds. Others, though—e.g., targeted advertising, some forms of digital nudging, and especially recommender systems—are more complex. On one hand, they undoubtedly provide certain benefits. As we've seen, recommender systems help us navigate otherwise overwhelming gluts of information, products, and online content. Digital nudges

[4]In a 2018 survey, Pew Research found that more than half of Facebook users reoprted they did not understand why they were shown the particular content in their feeds. See https://www.pewresearch.org/fact-tank/2018/09/05/many-facebook-users-dont-understand-how-the-sites-news-feed-works/

can encourage us to behave in ways that are healthier and more environmentally sustainable. Targeted advertising surfaces products relevant to our preferences and interests, and more importantly, it pays for other services people value.

On the other hand, many worry that along with whatever good they do, these influence technologies can also cause a variety of harms. The data collection that fuels them raises significant privacy concerns. The way they influence people could threaten individual autonomy. Targets of automated influence can suffer economic and epistemic harms. Determining if and when automated influence is morally acceptable requires taking all of these issues into account, and *measuring* the morally-relevant impacts of these technologies requires asking questions about them that go beyond how effectively they deliver on their promises—beyond questions about click-through rates and conversions, to questions about the effects of these technologies on privacy, autonomy, and other ethical values.

In what follows, we briefly discuss each of these issues. In the final section we consider how empirical scholarship can help determine whether, in each case, the good outweighs the bad.

3.1 Privacy

Influence technologies run on personal information. Targeting advertisements, fine-tuning nudges, and filtering recommendations requires information about the people on the receiving ends of these systems. For this reason, these technologies have drawn the attention of researchers and policymakers concerned about privacy [3]. Surveying research on the ethics of targeted advertising, for example, Varnali finds that "the literature unanimously acknowledges the fact that the technology that allows tracking individuals as they surf the Internet and process this data to single-out and deliver personalized ads has unprecedented potential to violate privacy" [99].

But privacy is a contested—perhaps an "essentially contested"—concept, which is to say, there is disagreement about both the harms privacy violations entail and about the steps necessary to avoid them [63]. In US law and policy, privacy is generally understood as an individual's right to control their personal information [103], a right which is operationalized in the form of so-called "notice and consent" procedures [86]. According to this approach, data collection violates privacy when data collectors fail to solicit consent before capturing and analyzing personal information. By contrast, Europe's General Data Protection Regulation (GDPR) allows for data collection when users consent[5] *or* when it is necessary to fulfill a "legitimate interest." Whether or not collecting personal information for the purposes of ad targeting and related practices meets this "legitimate interests" provision is the subject of debate [37].

Some believe these approaches fundamentally misunderstand privacy. Philosopher Helen Nissenbaum, for example, argues that privacy, generally, should be understood not in terms of individual consent or organizational interests, but rather as the flow of information in accordance with shared, context-dependent social norms—what she calls privacy as "contextual integrity" [69]. On this view, the data collection that fuels targeted advertising threatens privacy, because it tends to involve data flowing inappropriately. Information input into search engines, shared on social media, or gleaned from online shopping is repurposed, often without the data subject's knowledge, for targeting [70]. As Nissenbaum and others point out, this doesn't mean ad targeting *necessarily* violates privacy. Targeting technologies could, for instance, process information locally on individual machines, rather than transmitting personal data to third party aggregators [90]. Whether digital nudges and other automated influence technologies violate or respect privacy rests on similar questions.

3.2 Autonomy

Concerns about privacy draw attention to the data fueling automated influence, raising questions about how it is acquired. But even if all of those worries were resolved, and people were assured that information about them was captured and analyzed appropriately, questions would remain about the influence itself. Chief among them would be concerns about its effects on individual autonomy.

Autonomy is the capacity for independent decision-making [77], and respect for autonomy is a core liberal democratic value [18]. Certain kinds of influence threaten this capacity. For example, coercing people—i.e., *forcing* them to act a certain way—deprives them of autonomy [104]. As does manipulating people, which interferes with their ability to reflect on and reason about their options [84]. However, not all influence is corrosive to autonomy. Persuasion—offering reasons or incentives to act a certain way—respects autonomy, because it allows the target of influence to consider their options and decide for themselves [87]. Automated influence invites worries about autonomy to the extent that it is manipulative or coercive, rather than persuasive.

A wealth of recent scholarship contends that targeted advertising can manipulate and even coerce. Zuboff argues that because it is practically impossible to evade the data collection driving targeted advertising, individuals are forcibly—coercively—subjected to it [112]. For Calo, insofar as targeted ads exploit cognitive biases, they involve a kind of "digital market manipulation" [14]. Susser, Roessler, and Nissenbaum argue that targeted advertising constitutes "online manipulation" whenever it influences people covertly—i.e., when individuals aren't aware that they are being influenced, toward what end, or how [87]. These worries are especially salient in the political context, where threats to voter autonomy can undermine the integrity of democratic processes [25, 114].

Nudging—digital and otherwise—has also raised autonomy concerns, though there is significant disagreement about how warranted they are [35, 72, 79, 84, 88]. Sunstein argues that nudges do not threaten autonomy, as long as (1) they do not foreclose options, (2) they do not impose significant costs on making any particular choice, and (3) they aim to further the target's interests [84]. However, it can be difficult to determine whose interests, exactly, nudges—especially digital nudges—serve. For example, unbeknownst to users, the nudges many mobile health ("mHealth") apps deliver are designed not just to improve user heath, but also to drive users to buy sponsored products and services. Which is to say, they "try to influence economic behaviours by way of merging health

[5]Though the consent procedures GDPR requires are more demanding than in the US context. See [37].

and commercial content" [78]. Similar worries have been raised about recommender systems [12, 60]. Ensuring that nudges are "transparent"—i.e., people understand that they are being nudged and toward what ends—can, to some extent, mitigate these concerns [35, 59]. But given our reliance on recommender systems, and the increasing sophistication of many digital nudges, concerns about their effects on individual autonomy are likely to grow.

3.3 Economic Harm

The manipulative influences that raise autonomy concerns have also raised concerns about economic harm. Targeted advertisements, recommender systems, and dark patterns that exploit either the "widespread human biases" identified by behavioral economists or the particular, "idiosyncratic" decision-making vulnerabilities of individuals, threaten not only people's independence but also their wallets.

From an economic perspective, there are two potential problems. First, dark patterns, product recommender systems, and manipulative targeted advertisements might induce buyers to purchase what they would not—on reflection—choose to buy, creating inefficiencies in the allocation of goods (i.e., distributing products to people who will not value them most) [110]. Second, manipulative influences might enable sellers to charge more for products than buyers would otherwise pay. Although sellers have always engaged in what economists call "first degree price discrimination" and technology firms refer to as "price customization"—i.e., tailoring prices to individual buyers, based on what is known about them—many worry that the surveillance and dynamic experimentation digital environments afford could supercharge these tactics [62]. When that happens, overall economic efficiency is reduced and sellers are able to "siphon rents," claiming more than their fair share of the surplus [101]. On the whole, as Calo writes, the rise of automated influence technologies mean "A firm with the resources and inclination will be in a position to surface and exploit how consumers tend to deviate from rational decisionmaking on a previously unimaginable scale. Thus, firms will increasingly be in the position to create suckers, rather than waiting for one to be born" [14].

Moreover, as Zarsky points out, these technologies do not need to "work"—in the sense of successfully influencing people to buy things they don't really want or to pay more for them than they otherwise would—to introduce inefficiencies into the market. Even completely ineffective targeted advertisements can irritate or stress consumers, leading them to waste time and energy seeking out ad-blocking technology, and generating general distrust toward sellers [110]. Similar concerns have been raised about product recommender systems. Many have complained, for example, that Amazon's product recommendations rely on fake reviews, wasting shoppers' time and leading to overall distrust in the site's suggestions [67].

3.4 Epistemic Harm

Beyond concerns about privacy, autonomy, and economic welfare, critical scholars have for many years warned about the grave epistemic harms that automated influence threatens—harms people suffer as *knowers*. Almost a decade ago, Eli Pariser warned that personalized digital environments seemed to be isolating people in hermetic "filter bubbles," only delivering news, media, and other content that confirmed their prior beliefs and aligned with their tastes, preferences, and values [74]. As Pariser pointed out then, this can be detrimental to both individuals and society. Individually, people's perspectives are narrowed, their intellectual horizons diminished, and their beliefs polarized. Collectively, people lose the sense of existing in a shared world, driving them to prioritize personal interests over common purpose.

Not everyone is convinced that filter bubbles are quite so hermetically sealed (at least not yet) [11, 27, 115]. And some argue that traditional "legacy media"—especially cable news—is a more significant driver of political polarization than digital technologies [4]. The rise of internet-driven political disinformation, conspiracy theories, and vaccine and other health misinformation, however, are—for many—reason for persistent concern. A report from the Data & Society Research Institute warns that digital technologies enable a combination of surveillance, targeting, and automation—in their words, a "digital influence machine"—that can be "weaponized" by malign actors [65], damaging both individual, autonomous and collective, democratic experience.

4 EMPIRICAL EVIDENCE AND ETHICAL VALUES

So far, we have (1) surveyed the landscape of automated influence technologies, illustrating how they work, some ends and purposes to which they are applied, and the empirical research that investigates whether and to what extent they effectively deliver on their promises, and (2) described some of the many ethical concerns about these technologies, raised by researchers and policymakers. We illustrate some of our findings in two tables. In Table 1, we point to the kinds of concrete research questions that motivate much empirical work on the effectiveness of targeted advertising, digital nudges, and recommender systems. In Table 2, we highlight empirical questions generated by the ethics and policy literature. While these sets of questions overlap in some places, we suggest that the issues raised by ethics and policy research—the morally relevant dimensions of these technologies, and their effects on individuals and society—are in need of more systematic empirical investigation.

4.1 Privacy

Consider empirical questions about the effects of automated influence on privacy. Drawing from the discussion about privacy in the previous section, we can see that determining whether or not—or to what extent—these technologies impact privacy rests on the following kinds of questions:

- "How much personal information is needed to effectively target advertisements or content recommendations?"
- "Do people report feeling in control of their personal information?"
- "Do individuals understand and consent to the data collection practices fueling targeted advertising and other influence technologies?"
- "To what extent are people capable of refusing/avoiding tracking?"
- "Do people perceive targeting as invasive? Do they report self-censoring in response to targeting?"

Table 1: Typical empirical questions exploring the effectiveness of automated influence technologies

Mode of Influence	Application Domain	Typical Empirical Research Questions
Targeted Advertising	Commerce	"Are users presented with targeted advertisements more likely to report an intention to purchase the advertised product or service, relative to generic controls?"
		"Are users presented with targeted advertisements more likely to click on them, relative to generic controls?"
Digital Nudges	Health	"Do users report higher physical activity levels after receiving targeted digital nudges, relative to controls?"
	Sustainability	"Do car tracking technologies report better fuel consumption when drivers receive targeted digital nudges, relative to controls?"
Recommender Systems	Social Media	"Do targeted recommendations drive users to spend longer on the platform?"

Table 2: Examples of empirical questions generated by concerns raised in the ethics and policy literatures

Ethical Value	Mode of Influence	Relevant Empirical Questions
Privacy	Targeted Advertising	"Is targeting data collected with the consent of data subjects?"
		"Do people think targeting data is collected in ways that are contextually appropriate?"
Autonomy	Health Nudges	"Are people aware that they are being nudged?"
		"Do people understand why and how they are being nudged?"
Economic Harm	Dark Patterns	"Do particular dark patterns cause people to pay more than they otherwise would for specific products?"
Epistemic Harm	Recommender Systems	"Are people susceptible to conspiracy theories or other forms of disinformation in proportion to the relative share of news and other media they receive via targeted recommendations?"
Epistemic Harm	Recommender Systems	"Is public debate healthier/more robust in places where rates of e.g., social media usage, are lower?"

- "When asked, do people think targeting information is collected in ways that are contextually appropriate?"

Related questions have been the subject of some highly significant empirical research, though not necessarily in the context of automated influence. For example, Acquisti et al. analyze people's sense of and ability to control their personal information [1], McDonald and Cranor investigate the costs (and opportunity costs) of privacy consent procedures [58], and Martin and Nissenbaum use survey instruments to explore how people understand privacy in relation to context [54]. This work can serve as an important model for future efforts. As new technologies for delivering automated influence are developed and existing technologies evolve, research into their effects on privacy will need to evolve along with it.

4.2 Autonomy

As we saw in the previous section, there are urgent, unanswered questions about *how* automated influence technologies impact individual decision-making—especially, the impacts of such influence on individual autonomy. Whether or not particular vehicles of automated influence threaten autonomy depends on questions such as:

- "Are the targets of influence aware that they are being targeted?"
- "Do the targets of influence know *why* they are being targeted—i.e., do they understand the particular ends to which the influence is designed to steer them?"
- "Do the targets of influence understand how the technologies targeting them work?"
- "What characteristics do influence technologies target? Do they seek to exploit people's weaknesses and vulnerabilities?"
- "How difficult is it for targets of influence to resist advertisements, nudges, or other appeals?"
- "If targets of influence are briefed about having been targeted, do they report feeling used or exploited?"

Bridging philosophy and HCI, Calvo et al. offer a framework for understanding where autonomy is implicated in the design of digital systems [15]. Some preliminary survey research has explored user

awareness of dark patterns [52] and user experience of dark patterns as manipulative [33]. Mathur et al. raise related questions about how to empirically measure the extent to which digital influences impact autonomy, again in the context of dark patterns research [56]. We echo their call for the development of new empirical strategies that shed light on these questions, and suggest that such strategies would be valuable for exploring the autonomy impacts of a range of automated influence technologies beyond dark patterns.

4.3 Economic Harm

Of all the ethical issues raised in this paper, questions about economic harm would seem to be the most amenable to empirical measurement. For example, one might ask:

- "Do targets of influence actually pay more for goods or services than non-targeted control groups? If so, how much more?"
- "Do targets of influence return purchases or report regretting purchases at a higher rate than control groups?"
- "How much time and effort do targets of influence spend trying to avoid targeted appeals?"

Nevertheless, the law and policy literature investigating issues such as algorithmically-enabled price discrimination and the effects of digital technologies on consumer purchasing tends to focus on the *capabilities* of technologies to inflict related harms, and the economics literature relies almost entirely on theoretical modeling rather than empirical analysis. As we saw in the section on targeted advertising, above, a small number of empirical studies investigate the effects of targeted ads on how much individuals report they would pay for a particular product, which could speak to worries about price discrimination and the "siphoning" of rents by sellers. But overall there appears to be a significant gap in the literature: in the words of one scholar, there is "no consistent analytical framework and scant systematic empirical evidence about platform markets and consumer welfare effects" [21].

4.4 Epistemic Harm

Finally, as discussed in the previous section, some empirical research explores the extent to which content recommender systems isolate people in polarizing filter bubbles (mostly finding insufficient reason for significant concern). And major efforts have been undertaken to empirically study both the strategies utilized by malign actors to spread disinformation across social networks and ways to mitigate that spread.[6] Additional research is needed to understand the effects of these influences on individuals and society, exploring:

- "Are people susceptible to conspiracy theories or other forms of disinformation in proportion to the amount of time they spend online, or in proportion to the relative share of news and other media they receive via targeted recommendations?"
- "Are people more prone to confirmation bias/resistant to countervailing evidence the more time they spend online, or the more news and other media they receive via targeted recommendations?"
- "Is public debate healthier/more robust in places where rates of social media usage are lower?"

While some research in psychology and media effects investigates related questions (and AI ethics researchers should aim to build more bridges with these fields), there is little consensus about their answers. A major review of the online disinformation literature concludes that understanding "the effects of exposure to information and disinformation on individual beliefs and behaviors" is a "key remaining research question" [91].

5 CONCLUSION

Our goals in this paper have been twofold: (1) to draw attention to the landscape of automated influence technologies, to empirical research that investigates them, and to ethics and policy scholarship that raises concerns about their effects, and (2) to demonstrate that while these literatures overlap in places more work is needed to meaningfully integrate them. Measuring the effects of influence on individual decision-making is already methodologically challenging, but equal attention should be paid to determining *which effects* are measured.

For empirical research to inform ethics and policy debates it has to describe and measure the specific morally and legally relevant phenomena at their center, which requires carefully conceptualizing and defining empirical research problems through an ethics and policy lens. Put another way, normative (i.e., "values") issues ought to drive the construction of empirical questions, rather than the reverse. As we discussed in the previous section, examples of such work exist—especially in more developed areas of policy research, such as privacy studies. Critical and empirical scholars develop projects together from the ground up, translating normative questions into empirical ones. These collaborations offer models for future efforts. We hope to have demonstrated the urgency of this work, and to motivate and guide future research exploring automated influence at the intersection of ethics and empirical science.

[6] See, e.g., the work of the Data & Society Research Institute's "Media Manipulation Project," Harvard's Shorenstein Center on Media Politics and Public Policy, and the Oxford Internet Institute's "Computational Propaganda Project."

REFERENCES

[1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and Human Behavior in the Age of Information. *Science* 347, 6221 (2015), 509–514. https://doi.org/10.1126/science.aaa1465

[2] Rikke Aune Asbjørnsen, Mirjam Lien Smedsrød, Lise Solberg Nes, Jobke Wentzel, Cecilie Varsi, Jøran Hjelmesæth, and Julia EWC van Gemert-Pijnen. 2019. Persuasive System Design Principles and Behavior Change Techniques to Stimulate Motivation and Adherence in Electronic Health Interventions to Support Weight Loss Maintenance: Scoping Review. *Journal of Medical Internet Research* 21, 6 (2019), e14265. https://doi.org/10.2196/14265 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[3] David Beer, Joanna Redden, Ben Williamson, and Simon Yuill. 2019. *Landscape Summary: Online Targeting: What Is Online Targeting, What Impact Does It Have, and How Can We Maximise Benefits and Minimise Harms?* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819057/Landscape_Summary_-_Online_Targeting.pdf publisher: Centre for Data Ethics and Innovation.

[4] Yochai Benkler, Rob Faris, and Hal Roberts. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics.* Oxford University Press.

[5] Alexander Bleier and Maik Eisenbeiss. 2015. Personalized Online Advertising Effectiveness: The Interplay of What, When, and Where. *Marketing Science* 34, 5 (Sep 2015), 669–688. https://doi.org/10.1287/mksc.2015.0930

[6] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. 2013. Recommender Systems Survey. *Knowledge-Based Systems* 46 (Jul 2013), 109–132. https://doi.org/10.1016/j.knosys.2013.03.012

[7] Sophie C. Boerman, Sanne Kruikemeier, and Frederik J. Zuiderveen Borgesius. 2017. Online Behavioral Advertising: A Literature Review and Research Agenda. *Journal of Advertising* 46, 3 (Jul 2017), 363–376. https://doi.org/10.1080/00913367.2017.1339368

[8] Kanok Boriboonsomsin, Alexander Vu, and Matthew Barth. [n.d.]. Eco-Driving: Pilot Evaluation of Driving Behavior Changes among U.S. Drivers. ([n. d.]), 19.

[9] E. Bothos, Sebastian Prost, Johann Schrammel, K. Röderer, and G. Mentzas. 2014. Watch your Emissions: Persuasive Strategies and Choice Architecture for Sustainable Decisions in Urban Mobility. *PsychNology J.* (2014).

[10] Harry Brignull. 2013. Dark Patterns: inside the interfaces designed to trick you. https://www.theverge.com/2013/8/29/4640308/dark-patterns-inside-the-interfaces-designed-to-trick-you

[11] Axel Bruns. 2019. Filter bubble. *Internet Policy Review* 8, 4 (Nov 2019). https://doi.org/10.14763/2019.4.1426

[12] Christopher Burr, Nello Cristianini, and James Ladyman. 2018. An Analysis of the Interaction Between Intelligent Software Agents and Human Users. *Minds and Machines* 28, 4 (Dec. 2018), 735–774. https://doi.org/10.1007/s11023-018-9479-0

[13] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proceedings on Privacy Enhancing Technologies* 2016, 4 (Oct 2016), 237–254. https://doi.org/10.1515/popets-2016-0038

[14] M. Ryan Calo. 2014. Digital Market Manipulation. 82 (2014). https://doi.org/10.2139/ssrn.2309703

[15] Rafael A Calvo, Dorian Peters, Karina Vold, and Richard M. Ryan. 2020. Supporting Human Autonomy in AI Systems: A Framework for Ethical Enquiry. In *Ethics of Digital Well-Being: A Multidisciplinary Approach*, Christopher Burr and Luciano Floridi (Eds.). Springer Open, 31–54.

[16] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–15. https://doi.org/10.1145/3290605.3300733

[17] Centre for Data Ethics and Innovation. 2020. *Review of Online Targeting: Final Report and Recommendations.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/863030/CDEJ7836-Review-of-Online-Targeting-04022020-final.pdf

[18] John Philip Christman. 2009. *The Politics of Persons: Individual Autonomy and Socio-Historical Selves.* Cambridge University Press.

[19] Jennifer Cobbe and Jatinder Singh. 2019. Regulating Recommending: Motivations, Considerations, and Principles. *SSRN Electronic Journal* (2019). https://doi.org/10.2139/ssrn.3371830

[20] Julie E. Cohen. 2020. *The Emergent Limbic Media System.* Edward Elgar Publishing, 60–79. https://doi.org/10.4337/9781788972000.00010

[21] Diane Coyle. 2019. Practical Competition Policy Implications of Digital Platforms. *Antitrust Law Journal* 82 (2019), 835.

[22] Graham Dixon, Jay Hmielowski, and Yanni Ma. 2017. Improving Climate Change Acceptance Among U.S. Conservatives Through Value-Based Message Targeting. *Science Communication* 39 (June 2017), 107554701771547. https://doi.org/10.1177/1075547017715473

[23] C Doctorow. 2020. How to Destroy 'Surveillance Capitalism [Blog post]. OneZero.

[24] Renee Dudley. 2020. Amazon's New Competitive Advantage: Putting Its Own Products First. *ProPublica* (Jun 2020). https://www.propublica.org/article/amazons-new-competitive-advantage-putting-its-own-products-first

[25] European Commission. 2020. *Technology and Democracy: Understanding the Influence of Online Technologies on Political Behaviour and Decision-Making.* https://publications.jrc.ec.europa.eu/repository/bitstream/JRC122023/technology_democracy_final_online.pdf

[26] Ayman Farahat and Michael C Bailey. 2012. How effective is targeted advertising? (2012), 10.

[27] Richard Fletcher, Alessio Cornia, and Rasmus Kleis Nielsen. 2020. How Polarized Are Online and Offline News Audiences? A Comparative Analysis of Twelve Countries. *The International Journal of Press/Politics* 25, 2 (Apr 2020), 169–195. https://doi.org/10.1177/1940161219892768

[28] Alejandro Flores and Alexander Coppock. 2018. Do Bilinguals Respond More Favorably to Candidate Advertisements in English or in Spanish? *Political Communication* 35, 4 (Oct. 2018), 612–633. https://doi.org/10.1080/10584609.2018.1426663 Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2018.1426663.

[29] B.J. Fogg. 2003. *Persuasive Technology: Using Computers to Change What We Think and Do.* Morgan Kaufmann Publishers.

[30] Jon Froehlich, Tawanna Dillahunt, Predrag Klasnja, Jennifer Mankoff, Sunny Consolvo, Beverly Harrison, and James A. Landay. 2009. UbiGreen: investigating a mobile tool for tracking and supporting green transportation habits. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*. ACM Press, Boston, MA, USA, 1043. https://doi.org/10.1145/1518701.1518861

[31] Alan S. Gerber, Gregory A. Huber, David Doherty, Conor M. Dowling, and Costas Panagopoulos. 2013. Big Five Personality Traits and Responses to Persuasive Appeals: Results from Voter Turnout Experiments. *Political Behavior* 35, 4 (Dec. 2013), 687–728. https://doi.org/10.1007/s11109-012-9216-y

[32] Shirin Ghaffary and Jason Del Rey. 2020. The Big Tech Antitrust Report Has One Big Conclusion: Amazon, Apple, Facebook, and Google Are Anti-Competitive. *Recode* (Oct 2020). https://www.vox.com/recode/2020/10/6/21505027/congress-big-tech-antitrust-report-facebook-google-amazon-apple-mark-zuckerberg-jeff-bezos-tim-cook

[33] Colin M. Gray, Jingle Chen, Shruthi Sai Chivukula, and Liyang Qu. 2020. End User Accounts of Dark Patterns as Felt Manipulation. *CoRR* abs/2010.11046 (2020). arXiv:2010.11046 https://arxiv.org/abs/2010.11046

[34] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, 1–14. https://doi.org/10.1145/3173574.3174108

[35] Daniel M. Hausman and Brynn Welch. 2010. Debate: To Nudge or Not to Nudge. *Journal of Political Philosophy* 18, 1 (Mar 2010), 123–136. https://doi.org/10.1111/j.1467-9760.2009.00351.x

[36] Eitan Hersh and Brian F. Schaffner. 2011. *When is Pandering Persuasive? The Effects of Targeted Group-Based Appeals.* SSRN Scholarly Paper ID 1901820. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=1901820

[37] Chris Jay Hoofnagle, Bart van der Sloot, and Frederik Zuiderveen Borgesius. 2019. The European Union General Data Protection Regulation: What It Is and What It Means. *Information Communications Technology Law* 28, 1 (Jan 2019), 65–98. https://doi.org/10.1080/13600834.2019.1573501

[38] Tim Hwang. 2020. *Subprime Attention Crisis: Advertising and the Bomb at the Heart of the Internet.* Farrar, Straus and Giroux.

[39] Lucas D. Introna and Helen Nissenbaum. 2000. Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society* 16, 3 (Jul 2000), 169–185. https://doi.org/10.1080/01972240050133634

[40] Elvira Ismagilova, Emma L. Slade, Nripendra P. Rana, this link will open in a new window Link to external site, and Yogesh K. Dwivedi. 2019. The Effect of Electronic Word of Mouth Communications on Intention to Buy: A Meta-Analysis. *Information Systems Frontiers; New York* (May 2019), 1–24. https://doi.org/10.1007/s10796-019-09924-y Num Pages: 1-24 Place: New York, Netherlands, New York Publisher: Springer Nature B.V.

[41] Maurits Kaptein. 2015. *Persuasion Profiling: How the Intenret Knows What Makes You Tick.* Business Contact Publishers.

[42] Maurits Kaptein. 2018. Customizing persuasive messages; the value of operative measures. *The Journal of Consumer Marketing; Santa Barbara* 35, 2 (2018), 208–217. https://doi.org/10.1108/JCM-11-2016-1996 Num Pages: 10 Place: Santa Barbara, United Kingdom, Santa Barbara Publisher: Emerald Group Publishing Limited.

[43] Maurits Kaptein, Panos Markopoulos, Boris de Ruyter, and Emile Aarts. 2015. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies* 77 (May 2015), 38–51. https://doi.org/10.1016/j.ijhcs.2015.01.004

[44] Maurits Kaptein and Petri Parvinen. 2015. Advancing E-Commerce Personalization: Process Framework and Case Study. *International Journal of Electronic Commerce* 19, 3 (Jul 2015), 7–33. https://doi.org/10.1080/10864415.2015.1000216

[45] Lennart J. Krotzek. 2019. Inside the Voter's Mind: The Effect of Psychometric Microtargeting on Feelings Toward and Propensity to Vote for a Candidate. *International Journal of Communication* 13, 0 (Aug. 2019), 21. https://ijoc.org/index.php/ijoc/article/view/9605 Number: 0.

[46] C. Lacey and C. Caudwell. 2019. Cuteness as a 'Dark Pattern' in Home Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 374–381. https://doi.org/10.1109/HRI.2019.8673274

[47] Amy E. Latimer, Lawrence R. Brawley, and Rebecca L. Bassett. 2010. A systematic review of three approaches for constructing physical activity messages: What messages work and what improvements are needed? *International Journal of Behavioral Nutrition and Physical Activity* 7, 1 (May 2010), 36. https://doi.org/10.1186/1479-5868-7-36

[48] Bruno Latour. 1992. *Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts.* MIT Press.

[49] Jill Lepore. 2020. *If Then: How the Simulmatics Corporation Invented the Future* (first edition ed.). Liveright Publishing Corporation.

[50] Shaohua Lian, Tingting Cha, and Yunjie Xu. 2019. Enhancing geotargeting with temporal targeting, behavioral targeting and promotion for comprehensive contextual targeting. *Decision Support Systems* 117 (Feb. 2019), 28–37. https://doi.org/10.1016/j.dss.2018.12.004

[51] Jamie Luguri and Lior Jacob Strahilevitz. 2019. Shining a Light on Dark Patterns. *SSRN* (2019). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3431205

[52] Maximilian Maier and Rikard Harr. 2020. Dark Design Patterns: An End-User Perspective. *Human Technology* 16, 2 (Aug 2020), 170–199. https://doi.org/10.17011/ht/urn.202008245641

[53] Miguel Malheiros, Charlene Jennett, Snehalee Patel, Sacha Brostoff, and Martina Angela Sasse. 2012. Too close for comfort: a study of the effectiveness and acceptability of rich-media personalized advertising. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 579–588. https://doi.org/10.1145/2207676.2207758

[54] Kirsten Martin and Helen Nissenbaum. 2016. Measuring Privacy: An Empirical Test Using Context to Expose Confounding Variablest. *Columbia Science and Technology Law Review* 18, 1 (2016), 44.

[55] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov 2019), 1–32. https://doi.org/10.1145/3359183

[56] Arunesh Mathur, Jonathan Mayer, and Mihir Kshirsagar. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. *arXiv:2101.04843 [cs]* (Jan 2021). https://doi.org/10.1145/3411764.3445610 arXiv: 2101.04843.

[57] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences* 114, 48 (Nov. 2017), 12714–12719. https://doi.org/10.1073/pnas.1710966114

[58] Aleecia M Mcdonald and Lorrie Faith Cranor. 2008. The Cost of Reading Privacy Policies. *I/S: A Journal of Law and Policy for the Information Society* 4, 3 (2008), 26.

[59] Christian Meske and Ireti Amojo. 2020. Ethical Guidelines for the Construction of Digital Nudges. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*. 10.

[60] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender systems and their ethical challenges. *AI SOCIETY* 35, 4 (Dec 2020), 957–967. https://doi.org/10.1007/s00146-020-00950-y

[61] Tobias Mirsch, Christiane Lehrer, and Reinhard Jung. 2017. Digital Nudging: Altering User Behavior in Digital Environments. In *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017)*. 15.

[62] Rafi Mohammed. 2017. How Retailers Use Personalized Prices to Test What You're Willing to Pay. *Harvard Business Review* (Oct 2017), 4.

[63] Deirdre K. Mulligan, Colin Koopman, and Nick Doty. 2016. Privacy Is an Essentially Contested Concept: A Multi-Dimensional Analytic for Mapping Privacy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2083 (Dec 2016), 20160118. https://doi.org/10.1098/rsta.2016.0118

[64] Anthony Nadler, Matthew Crain, and Joan Donovan. 2018. *Weaponizing the Digital Influence Machine: The Political Perils of Online Ad Tech*. 47 pages. https://datasociety.net/wp-content/uploads/2018/10/DS_Digital_Influence_Machine.pdf

[65] Anthony Nadler and Joan Donovan. 2018. Weaponizing the Digital Influence Machine. https://datasociety.net/library/weaponizing-the-digital-influence-machine/ Publisher: Data & Society Research Institute.

[66] Arvind Narayanan, Arunesh Mathur, Marshini Chetty, and Mihir Kshirsagar. 2020. Dark Patterns: Past, Present, and Future. *ACM Queue* 18, 2 (May 2020), 25.

[67] Nicole Nguyen. 2019. "Amazon's Choice" Does Not Necessarily Mean A Product Is Good. *BuzzFeed* (Jun 2019). https://www.buzzfeednews.com/article/nicolenguyen/amazons-choice-bad-products

[68] Jack Nicas. 2017. Google Uses Its Search Engine to Hawk Its Products. *The Wall Street Journal* (2017). https://www.wsj.com/articles/google-uses-its-search-engine-to-hawk-its-products-1484827203

[69] Helen Nissenbaum. 2010. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books.

[70] Helen Nissenbaum. 2011. A Contextual Approach to Privacy Online. *Daedalus* 140, 4 (Sep 2011), 32–48. https://doi.org/10.1162/DAED_a_00113

[71] Safiya Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

[72] Robert Noggle. 2018. Manipulation, Salience, and Nudges. *Bioethics* 32, 3 (2018), 164–170. https://doi.org/10.1111/bioe.12421

[73] Vance Packard. 1957. *The Hidden Persuaders*. D. McKay Co.

[74] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1118322

[75] Mashfiqui Rabbi, Angela Pfammatter, Mi Zhang, Bonnie Spring, and Tanzeem Choudhury. 2015. Automated Personalized Feedback for Physical Activity and Dietary Behavior Change With Mobile Phones: A Randomized Controlled Trial on Adults. *JMIR mHealth and uHealth* 3, 2 (2015), e42. https://doi.org/10.2196/mhealth.4160 Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.

[76] Sarah Rajtmajer and Daniel Susser. 2020. Automated Influence and the Challenge of Cognitive Security. In *Proceedings of the 7th Symposium on Hot Topics in the Science of Security*. ACM, 1–9. https://doi.org/10.1145/3384217.3385615

[77] Joseph Raz. 1986. *The Morality of Freedom* (reprinted ed.). Clarendon Press.

[78] Marijn Sax, Natali Helberger, and Nadine Bol. 2018. Health as a Means Towards Profitable Ends: mHealth Apps, User Autonomy, and Unfair Commercial Practices. *Journal of Consumer Policy* 41, 2 (Jun 2018), 103–134. https://doi.org/10.1007/s10603-018-9374-3

[79] Andreas T. Schmidt and Bart Engelen. 2020. The Ethics of Nudging: An Overview. *Philosophy Compass* 15, 4 (2020), e12658. https://doi.org/10.1111/phc3.12658

[80] Christoph Schneider, Markus Weinmann, and Jan vom Brocke. 2018. Digital Nudging: Guiding Online User Choices Through Interface Design. *Commun. ACM* 61, 7 (Jun 2018), 67–73. https://doi.org/10.1145/3213765

[81] Nick Seaver. 2018. Captivating Algorithms: Recommender Systems as Traps. *Journal of Material Culture* (Dec 2018), 16.

[82] Azizi Seixas, Colleen Conners, Alicia Chung, Tiffany Donley, and Girardin Jean-Louis. 2020. A Pantheoretical Framework to Optimize Adherence to Healthy Lifestyle Behaviors and Medication Adherence: The Use of Personalized Approaches to Overcome Barriers and Optimize Facilitators to Achieve Adherence. *JMIR mHealth and uHealth* 8, 6 (2020), e16429. https://doi.org/10.2196/16429 Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.

[83] Sjef Siero, Martin Boon, Gerjo Kok, and Frans Siero. 1989. Modification of driving behavior in a large transport organization: A field experiment. *Journal of Applied Psychology* 74, 3 (June 1989), 417–423. https://doi.org/10.1037/0021-9010.74.3.417 Num Pages: 417-423 Place: Washington, US Publisher: American Psychological Association (US).

[84] Cass R Sunstein. 2016. *The Ethics of Influence: Government in the Age of Behavioral Science*. Cambridge University Press.

[85] Daniel Susser. 2019. Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures. In *AIES: AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*.

[86] Daniel Susser. 2019. Notice After Notice-and-Consent: Why Privacy Disclosures Are Valuable Even If Consent Frameworks Aren't. *Journal of Information Policy* 9 (2019), 37. https://doi.org/10.5325/jinfopoli.9.2019.0037

[87] Daniel Susser, Beate Roessler, and Helen Nissenbaum. 2019. Online Manipulation: Hidden Influences in a Digital World. *Georgetown Law Technology Review* 4 (2019), 1–45.

[88] Henrik Skaug Sætra. 2019. When Nudge Comes to Shove: Liberty and Nudging in the Era of Big Data. *Technology in Society* (Apr 2019), S0160791X19300661. https://doi.org/10.1016/j.techsoc.2019.04.006

[89] Richard H. Thaler and Cass R. Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press.

[90] Vincent Toubiana, Arvind Narayanan, Dan Boneh, Helen Nissenbaum, and Solon Barocas. 2010. Adnostic: Privacy preserving targeted advertising. In *Proceedings Network and Distributed System Symposium*.

[91] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *William and Flora Hewlett Foundation* (2018).

[92] Zeynep Tufekci. 2014. Engineering the Public: Big Data, Surveillance and Computational Politics. *First Monday* 19, 7 (Jul 2014). http://firstmonday.org/ojs/index.php/fm/article/view/4901

[93] Johannes Tulusan, Thorsten Staake, and Elgar Fleisch. 2012. Providing eco-driving feedback to corporate car drivers: what impact does a smartphone application have on their fuel efficiency?. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*. ACM Press, Pittsburgh, Pennsylvania, 212. https://doi.org/10.1145/2370216.2370250

[94] J. Turow. 2012. *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth*. Yale University Press. https://books.google.com/books?id=rK7JSFudXA8C

[95] Joseph Turow, Michael X. Delli Carpini, Nora A Draper, and Rowan Howard-Williams. 2012. *Americans Roundly Reject Tailored Political Advertising—At a Time When Political Campaigns are Embracing It*.

[96] Joseph Turow and Chris Jay Hoofnagle. 2019. Mark Zuckerberg's Delusion of Consumer Consent. *The New York Times* (Jan 2019). https://www.nytimes.com/2019/01/29/opinion/zuckerberg-facebook-ads.html

[97] Ali A. Valenzuela and Melissa R. Michelson. 2016. *Turnout, Status, and Identity: Mobilizing Latinos to Vote with Group Appeals*. SSRN Scholarly Paper ID 2818475. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=2818475

[98] Jenny van Doorn and Janny C. Hoekstra. 2013. Customization of online advertising: The role of intrusiveness. *Marketing Letters* 24, 4 (Dec. 2013), 339–351. https://doi.org/10.1007/s11002-012-9222-1

[99] Kaan Varnali. 2019. Online behavioral advertising: An integrative review. *Journal of Marketing Communications* 27, 1 (Jan 2019), 93–114. https://doi.org/10.1080/

13527266.2019.1630664

[100] Lex van Velsen, Marijke Broekhuis, Stephanie Jansen-Kosterink, and Harm op den Akker. 2019. Tailoring Persuasive Electronic Health Strategies for Older Adults on the Basis of Personal Motivation: Web-Based Survey Study. *Journal of Medical Internet Research* 21, 9 (2019), e11759. https://doi.org/10.2196/11759 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[101] Gerhard Wagner and Horst Eidenmuller. 2019. Down by Algorithms: Siphoning Rents, Exploiting Biases, and Shaping Preferences: Regulating the Dark Side of Personalized Transactions. *University of Chicago Law Review* 86 (2019), 581. https://heinonline.org/HOL/Page?handle=hein.journals/uclr86&id=593&div=&collection=

[102] Markus Weinmann, Christoph Schneider, and Jan vom Brocke. 2016. Digital Nudging. *Business Information Systems Engineering* 58, 6 (Dec 2016), 433–436. https://doi.org/10.1007/s12599-016-0453-1

[103] Alan Furman Westin. 2015. *Privacy and Freedom.* IG Publishing.

[104] Allen Wood. 2014. *Coercion, Manipulation, Exploitation.* Oxford University Press.

[105] Tim Wu. 2016. *The Attention Merchants: The Epic Scramble to Get Inside Our Heads* (first edition ed.). Alfred A. Knopf.

[106] Wenzhen Xu, Yuichi Kuriki, Taiki Sato, Masato Taya, and Chihiro Ono. 2020. Does Traffic Information Provided by Smartphones Increase Detour Behavior?: An Examination of Emotional Persuasive Strategy by Longitudinal Online Surveys and Location Information. In *Persuasive Technology. Designing for Future Change*, Sandra Burri Gram-Hansen, Tanja Svarre Jonasen, and Cees Midden (Eds.). Vol. 12064. Springer International Publishing, Cham, 45–57. https://doi.org/10.1007/978-3-030-45712-9_4 Series Title: Lecture Notes in Computer Science.

[107] Karen Yeung. 2017. Hypernudge: Big Data as a Mode of Regulation by Design. *Information, Communication Society* 20, 1 (Jan 2017), 118–136. https://doi.org/10.1080/1369118X.2016.1186713

[108] José P Zagal, Staffan Björk, and Chris Lewis. 2013. Dark Patterns in the Design of Games. In *Foundations of Digital Games Conference, FDG 2013*. 8.

[109] Brahim Zarouali, Tom Dobber, Guy De Pauw, and Claes de Vreese. 2020. Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media. *Communication Research* 0, 0 (2020), 0093650220961965. https://doi.org/10.1177/0093650220961965

[110] Tal Z Zarsky. 2019. Privacy and Manipulation in the Digital Age. *Theoretical Inquiries in Law* 20 (2019), 32.

[111] Kaifu Zhang and Zsolt Katona. 2012. Contextual Advertising. *Marketing Science* 31, 6 (Nov 2012), 980–994. https://doi.org/10.1287/mksc.1120.0740

[112] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (first edition ed.). PublicAffairs.

[113] Mark Zuckerberg. 2019. The Facts About Facebook. *Wall Street Journal* (Jan 2019). http://ezaccess.libraries.psu.edu/login?url=https://search-proquest-com.ezaccess.libraries.psu.edu/docview/2170828623?accountid=13158

[114] Frederik J. Zuiderveen Borgesius, Judith Möller, Sanne Kruikemeier, Ronan Ó Fathaigh, Kristina Irion, Tom Dobber, Balazs Bodo, and Claes De Vreese. 2018. Online Political Microtargeting: Promises and Threats for Democracy. *Utrecht Law Review* 14, 1 (Feb 2018), 82. https://doi.org/10.18352/ulr.420

[115] Frederik J. Zuiderveen Borgesius, Damian Trilling, Judith Möller, Balázs Bodó, Claes H. De Vreese, Natali Helberger, Internet Policy Review, and Internet Policy Review. 2016. Should We Worry About Filter Bubbles? *Internet Policy Review* (2016). https://doi.org/10.14763/2016.1.401