American Philosophical Quarterly
Volume 61, Number 4, October 2024

AUTOMATED INFLUENCE AND VALUE COLLAPSE

Resisting the Control Argument

Dylan J. White

ABSTRACT Automated influence is one of the most pervasive applications of artificial intelligence in our day-to-day lives, yet a thoroughgoing account of its associated individual and societal harms is lacking. By far the most widespread, compelling, and intuitive account of the harms associated with automated influence follows what I call the *control argument*. This argument suggests that users are persuaded, manipulated, and influenced by automated influence in a way that they have little or no control over. Based on evidence about the effectiveness of targeted advertising as well as empirical results about the nature of attentional control, I provide reasons to reject this argument. In turn, I use C. Thi Nguyen's theory of *value collapse* to develop a new account of the harmfulness of automated influence.

KEYWORDS automated influence, attention, control, value collapse, autonomy, persuasion

1. Introduction

One of the most pervasive applications of artificial intelligence (AI) in our day-to-day lives is the online world of automated influence. Benn and Lazar (2022) define automated influence as the use of AI "to collect, integrate and analyse people's data, and to deliver targeted interventions based on this analysis, intended to shape their behaviour for exogenous or endogenous ends" (p. 127). These interventions may take the form of targeted advertising, recommendations (for music, movies, who to date, where to eat, etc.), and digital nudges, as well as more subtle and opaque forms such as the ordering of search results or the news feed on your Facebook.[1]

Philosophers, policymakers, and psychologists alike have expressed concern about the moral and psychological harms of automated influence (Twenge, 2017; Aylsworth, 2020; Milano et al., 2020; Susser & Grimaldi, 2021; Benn & Lazar, 2022; Burtell & Woodside, 2023; White, 2024). The varieties of automated influence are said to manipulate, persuade, and influence individuals and societies in undesirable ways, leading to hyper-consumerism, degraded attention spans, threats to individual and collective autonomy, increased polarization and radicalization, and more. Despite a growing and cross-disciplinary literature on automated influence, a thoroughgoing account of exactly *why* automated influence is harmful and *how* it causes these harms is

still wanting. A common way of getting at the wrongness of automated influence is to appeal to the *persuasive design* of the attention economy (Eyal, 2013; Williams, 2018). On this account, users are at the mercy of the *persuasive design* of technologies, apps, and platforms that enable the collection of user data, making attentional control extremely difficult (if not impossible), behavior modification extremely likely, and putting individual and collective autonomy in peril (Williams, 2018; Aylsworth & Castro, 2021; Bhargava & Velasquez, 2021). This data is then used to deliver ads and recommendations that, similarly, have a controlling effect. Throughout, I refer to this as the *control argument*.

On the other hand, as Tim Hwang (2020) argues, despite what big tech, advertisers, and the general public may think, targeted advertising mostly fails to work, suggesting that users' attention is not reliably captured in the ways that the *control argument* takes for granted. For example, many reported ad-clicks are fraudulent or accidental, and studies suggest that online ads have little or no effect on the vast majority of users (Hwang, 2020). Although the *control argument* is both compelling and intuitive, it is ultimately incompatible with what we know about the ineffectiveness of large swaths of automated influence. If we are to better assess (and hopefully, address) the worries associated with automated influence, we must resolve this tension. We need a better framework for understanding the harms associated with automated influence.[2]

I develop such a framework using a suitably updated version of C. Thi Nguyen's (2020; 2021) theory of *value capture*, and subsequently, *value collapse*. Alone, Nguyen's (2021) *value collapse* argument is subject to similar criticisms of empirical inadequacy that I outline for the *control argument* below. I update Nguyen's argument with empirical evidence that suggests we have good reason to accept a version of this view. In developing this argument, I avoid the problems associated with the *control argument*, while also accounting for the measured ineffectiveness of targeted advertising. Further, through refining Nguyen's theory of *value capture* and *value collapse*, I demonstrate how these concepts can be used to develop a better understanding of automated influence.

2. The Control Argument

2.1 Automated Influence

To understand the *control argument*, we must first understand how the online world of automated influence operates. Ads, product and content recommendations, search results, and so on, are all delivered at breakneck speeds online. As Hwang notes:

> The entire process of putting out a request for bids, making the bids, evaluating the bids, and delivering the advertisement takes place in under a hundred milliseconds (. . .) This happens millions and millions of times across the internet every second, without ceasing and largely without hiccups. (Hwang, 2020, pp. 19–20)

This is only made possible through automation, relying on the "interaction between algorithms to make the discrete choices to bid on available blobs of advertising inventory" (Hwang, 2020, p. 20). These "blobs of advertising inventory" are generated through the attention economy—defined as the market where "consumers give new media developers their literal attention in exchange for a service" (Castro & Pham, 2020, p. 2).[3] I return to how exactly attention is commodified below. For now, it is only necessary to understand that automated influence is made possible by relying on YouTube videos, social media posts, search platforms, and more to capture the attention of users, collect data about them, and based on profiles constructed using that data, deliver ads, recommendations, and so on.

Through the collection, integration, and analysis of user data, AI powered recommender systems, targeted advertising, and

search engines help us to navigate our way through the "functionally infinite spaces of our digital infrastructure" (Benn & Lazar, 2022, p. 127). However, as noted above, automated influence has also been charged with eroding user autonomy in various ways. It is to these concerns that we now turn.

2.2 The Control Argument

Though pervasive, the *control argument* is rarely stated explicitly. Rather, those who either implicitly or explicitly endorse a version of *control argument* rely on empirically suspect ways of construing attention, vague metaphors, and anecdotal examples to illustrate their basic claim that users are at the mercy of, and capable of being controlled by, the *persuasive design* of various apps, platforms, recommender systems, etc. (Williams, 2018; Castro & Pham, 2020; Aylsworth & Castro, 2021; Bhargava & Velasquez, 2021). For example, Aylsworth & Castro (2021) liken the phenomenon of control to the words of comedian Esther Povitsky: "I wish I could read. I really do. I try to read. I buy books. I open books. And then I black out and I'm on Instagram and I don't know what happened" (Aylsworth & Castro, 2021, p. 1). Others have compared aspects of automated influence and the attention economy to "digital heroin" (Kardaras, 2016), and suggested that *persuasive design* 'hacks' the brain (Lustig, 2017).

To give some shape to the *control argument*, it can be stated as such:

P1. Attention[4] can be automatically captured by bottom-up stimuli.

P2. This attention capture is automatic in the sense that it cannot be prevented from happening.

P3. The *persuasive design* of automated influence and the attention economy automatically captures, and so controls, attention in this way.[5]

P4. The sustained loss of control over our attention is harmful.

C. Automated influence (and the attention economy) are harmful because the *persuasive design* and overall effectiveness of these technologies reliably and frequently cause us to lose control of our attention.

Versions of this argument can be found throughout large swaths of the academic and popular literature on automated influence and the attention economy, as well as throughout recent policy recommendations (Turel & Oahri-Saremi, 2016; 2018; Bermúdez, 2017; Wu, 2017; Williams, 2018; Castro and Pham, 2020; Aylsworth and Castro, 2021; Bhargava and Velasquez, 2021; Rieser and Furneaux, 2022; Rose-Stockwell, 2023; UNESCO, 2023; U.S. Surgeon General's Advisory, 2023). The argument above constitutes a reconstruction of common assumptions and arguments provided throughout these various accounts.

To sum up, the *control argument* suggests that it is the *persuasive design* of the technologies and platforms associated with automated influence that, by supposedly manipulating and controlling our attention, threaten the autonomy and self-determination of users. Aylsworth & Castro, for example, argue that targeted advertising and associated technologies such as smartphones:

> pose a distinct threat to our rational capacities because an effect of the addiction is susceptibility to having one's attention hijacked at frequent intervals, interrupting one's ongoing tasks (Aylsworth & Castro, 2021, p. 4).

Similarly, James Williams (2018) refers to the burdens of 'impossible self-regulation' and the tendency to "*lose control over* one's attentional processes" (Williams, 2018, p. 15, italics in original) because of the tools of the attention economy; the tools that enable the pervasiveness of automated influence. We have all experienced something like this. We have a goal (such as writing a paper), but we get distracted by an app, a social media platform, a game, and the like, that seems to

hijack our attention. It is the pervasiveness of experiences such as these that makes the *control argument* seem compelling and intuitive.

The *control argument*, however, takes what may be a compelling and intuitive metaphor—control—for making sense of the psychological effects of automated influence too far. A useful analogy can be made to the concept of 'resources' in psychology and cognitive science. The metaphor of 'resources' has been used to explain various psychological phenomena such as attention (Wu, 2017; Williams, 2018) and self-control (Baumeister et al., 1998), but they have continually failed to shed light on their explanatory target. The "ego-depletion" model of self-control, for example, has faced many problems, from unfounded conclusions that glucose is the relevant resource necessary to exert self-control, to a failure to consider the roles that motivation and value play in our self-control decisions (Inzlicht & Schmeichel, 2012; Inzlicht et al., 2013;Sripada, 2020). Williams (2018) repeats many of these mistakes by applying the "ego-depletion" model of self-control to the problematic technology use associated with automated influence and the attention economy. According to Williams, digital technologies have placed further strain on the "finite resource" that is our ability to exert willpower and self-control (Williams, 2018, p. 22). However, using resources to explain self-control have proven largely ineffective at specifying a mechanism, explaining and predicting behavior, and suggesting interventions to improve self-control (Inzlicht & Schmeichel, 2012; Inzlicht et al., 2013). Indeed, as the psychologist David Navon (1984) argues, although resources may be useful metaphors, they are ultimately misleading, and akin to a theoretical soup-stone—unnecessary and providing no actual explanatory power. Just as resources are empirically inadequate ways of explaining more nuanced and complex psychological phenomenon, I suggest that the concept of *control* is similarly not up to the task of explaining the nuances of the psychological effects of automated influence. As Navon notes, resources were originally proposed not as a "construct whose usage presupposes its possible existence as a mental entity, but rather as an intervening variable" (1984, p. 231). The construal of resources as an actual 'mental entity' precipitated the problems of the "ego-depletion" model of self-control. We should be careful not to make the same mistakes when it comes to the *control argument*.[6]

Ultimately, the *control argument* paints the user of these technologies as at the mercy of *persuasive design*, rendering them incapable of exerting control over where they allocate attention and the decisions they are likely to make. As Tobias-Rose Stockwell notes of the harmful effects of the attention economy: "We know it, but we cannot stop" (Stockwell, 2023).

The *control argument* makes a number of assumptions about the nature of attention, two of which I highlight here. The first is an empirical assumption. The *control argument* takes for granted that the attention 'capture' that occurs through the platforms and apps associated with the attention economy is automatic in the sense that "[a]ttempts by a subject to prevent an automatic process from proceeding are not successful" (Yantis and Jonides, 1990, p. 122). To borrow a phrase from Hannah Pickard (2022), the *control argument* suggests that "if people could stop using, they would. But they can't, which is why they don't" (p. 326). According to the *control argument*, undesirable behavior modification at the hands of automated influence is inevitable because automatic processes prevent voluntary control. The assumption that attention can be automatically and reliably captured in this way suggests a reliance on a strict dichotomy of top-down and bottom-up attention. According to the *control argument* bottom-up attention is captured in a way that

makes top-down control (defined as the explicit and voluntary goals of the subject) extremely difficult, if not impossible. However, as psychologists and neuroscientists have recently demonstrated, this strict dichotomy between top-down and bottom-up attention is a false one (Awh et al., 2012; Todd and Manaligod, 2018; Shomstein et al., 2022). I return to this below (Sec 3.1).

The second assumption of the *control argument*, though largely implicit, is a conceptual one that would appear to suggest that any attempt to normatively evaluate where individuals allocate their attention is a lost cause. After all, if you cannot control where you allocate your attention, you cannot be held responsible for it. As Regina Rini (2023) notes: "Consider how quick you are to judge a parent fixated on their handheld screen and failing to notice their child." Whether warranted or not, any such similar judgments about where attention gets allocated would be meaningless if we accept the *control argument*. As a key goal of many who advance the *control argument* is to normatively assess the responsibility individuals have for where they allocate their attention, there is an inherent and unresolvable tension present. Others bite the bullet, and absolve the individual user of responsibility. Stockwell suggests: "This isn't your fault. It's by design. The digital rabbit hole you just tumbled down is funded by advertising, aimed at you" (Stockwell, 2023). Despite the massive amounts of money being leveraged to capture your attention however, we should question whether or not such control is actually possible.

Ultimately, control may be a useful metaphor for conceptualizing some of the moral and psychological harms that we see associated with automated influence and the attention economy, but metaphors are often oversimplifying, preventing us from seeing the nuances that may be necessary to effectively understand and address a given problem. In order to address the problems associated with these AI enabled platforms, technologies, markets, etc., we will need not just useful metaphors, but an understanding of the problem in all its nuance and complexity.

3. What We Get Wrong About Automated Influence

3.1 Attentional Control

For decades now, much empirical work on attention has been carried out based on a model of attentional control that divides that control into top-down (or endogenous), and bottom-up (or exogenous). Top-down control is construed as voluntary, dictated by the current, explicit goals of the subject, whereas bottom-up control is determined by the physical salience of the environment (Posner, 1980; Jonides, 1981; Corbetta et al., 2002; Beck & Kastner, 2009). This way of thinking can be found, to varying degrees, in the writing of many philosophers on automated influence. Rieser and Furneaux (2022) suggest that "[e]xogenous attentional control (. . .) accounts for how salient external stimuli redirect human attention away from cognitive processes that are currently focused elsewhere" and go on to characterize exogenous (bottom-up) control by external stimuli as "independent of the mental states of the users" (Rieser & Furneaux, 2022, pp. 3–4). Other accounts suggest that the bells and whistles of the attention economy 'hijack' our attentional control (Aylsworth & Castro, 2021) and that our attention is stolen by screens that "literally seize scarce mental resources" (Wu, 2017).

Recent empirical work, however, demonstrates that this way of explaining the harms associated with the attention economy and automated influence, are less and less plausible. Psychologists and neuroscientists have demonstrated that the strict theoretical dichotomy posited between top-down and bottom-up attentional control is a false one (Awh et al., 2012; Todd & Manaligod, 2018; Shomstein et al., 2022). Awh and colleagues (2012) single out selection history (including

selection reward) as one aspect of attentional control that cannot be accounted for by either top-down, voluntary, goal-directed attentional control or bottom-up capture of attention by physically salient stimuli. In other words, there are effects of an individual's past experiences that influence the landscape, or priority structure of the subject's selection biases. Similarly, Todd and Manaligod (2018) propose that the theoretical framework of a priority state space (PSS) should replace the top-down versus bottom-up dichotomy. The PSS accounts for associative and statistical learning (similar in some ways to Awh et al.'s (2012) selection history), semantic association, and motivational and affective salience that do not fit neatly into the strict theoretical dichotomy that has dominated much attention research. The PSS necessarily and correctly complicates the endogenous/exogenous distinction of attentional control by emphasizing the complex interactions of various sources of salience that do not neatly fit this divide.

Philosophers have also begun to question this strict theoretical dichotomy between top-down and bottom-up attentional control (Jennings, 2020; Watzl, 2023). Ganeri (2017) writes: "The purported distinction between endogenous and exogenous which cognitive psychologists help themselves to brings with it far too many theoretical presuppositions to be helpful in the analysis of attention . . ." (p. 63). Accordingly, I suggest that the supposed persuasion and manipulation said to occur through automated influence does not occur in the way proposed by the *control argument*. The switch from thinking about attentional allocation as a strict dichotomy between top-down and bottom-up control to thinking in terms of priority structure maps and the plethora of influences on attentional control drastically changes how we should conceptualize the impact of automated influence. It cannot be described purely in terms of bottom-up capture or hijacking, rather we must consider the variety of influences such as affective and motivational states that dictate where we allocate our attention. We cannot rely on explaining the effects of the attention economy in terms of the bottom-up capture mediated by *persuasive design* alone. This does not mean that automated influence is not morally and psychologically harmful, but it does suggest that these harms do not occur in the way that many researchers suggest.

3.2 The Ineffectiveness of Targeted Advertising

The *control argument* also overestimates the ability of automated influence to persuade us to purchase advertised goods and to control and manipulate our behavior. In other words, automated influence often does not work. As Hwang (2020) argues, although targeted advertising is the "dark beating heart of the internet," with over 80 percent of Google's annual revenue coming from ads, 90 percent of Meta's, and Amazon and Microsoft similarly making billions every year from targeted advertising, the vast majority of these ads do not work. A recent report suggests that about a third of display-ad clicks alone are fraudulent (clicked by bots) or accidental (The Global PPC Click Fraud Report, 2020–21; ANA Programmatic Media Supply Chain Transparency Study, 2023). Moreover, ads appear to work on only a very small percent of the population, with one study from 2009 showing that around 8 percent of all users are responsible for 85 percent of all advertisement click-throughs (Hwang, 2020). The majority of these advertisement click-throughs come from loyal customers to the advertised product, meaning that the click-throughs came from users who would likely have purchased anyway. Younger users also seem to be much less susceptible to targeted advertisements than older users. As Hwang (2020) notes, a 2013 study of more than a million customers found that online ads had little or no effect on users between the ages

of 20 and 40, a demographic that represents a large proportion of internet users.

The selection history effect proposed by Awh et al. (2012) above may go some way towards explaining the ineffectiveness of targeted advertising. Targeted advertising is ubiquitous. Without an ad blocker, and in many instances even with one, it is near impossible to spend time online without being bombarded by targeted ads. Banner ads, pop-up videos, sponsored search results and more are now common place and have been for quite some time. So, even if targeted advertising once saw a marked improvement in the performance of the advertising industry, the now ever-present and expected nature of these distractors may make them easier to ignore, or at least to notice and quickly move on. Indeed, as Hwang (2020) notes, when banner ads first launched in 1994, they had an impressive click-through rate of 44 percent. Today, those click-through rates are less than 1 percent (Hwang, 2020). This discrepancy can be explained, at least in part, by the experimental work of Awh et al. (2012). They show that when specific distractors are associated with specific target positions or goals, the experience of these distractors in the past will drive more efficient orienting of attention. This suggests that constant exposure to the same, or similar, distractors (such as targeted ads) may eventually lead to these distractors being less effective. This does not necessarily mean that the distractors are not still potentially harmful or detrimental in some way, but in the case of targeted advertisements, it does suggest that they are not capable of the kind of control and manipulation of behavior assumed by the *control argument*.

3.3 Persistent Evidence of Harms

Despite the clear evidence that targeted ads are not as effective as we are led to believe, there remains good evidence that there are still harms associated with automated influence. Among the harms often associated with automated influence are the current adolescent mental health crisis, the deterioration of the autonomy of individuals, the exploitation of people's psychological vulnerabilities, and the undermining of collective autonomy and democracy (Twenge, 2017; Castro & Pham, 2020; Bhargava & Velasquez, 2021). Recent work has demonstrated the ability of reinforcement learning (RL) recommender systems to reliably manipulates users' values and opinions "as part of a policy to increase long-term user engagement" (Evans & Kasirzadeh, 2021). Similarly, Carroll, Dragan, Russell, and Hadfield (2022) show that recommender systems that are trained via long-horizon optimization have direct incentives to manipulate the values and preferences of users so that they are easier to satisfy, and moreover, that they are often successful. They write:

> . . . certain preferences are easier to satisfy than others, leading to more potential for engagement —this could be because of availability of more content for some preferences compared to others, or because strong preferences for a particular type of content lead to higher engagement than more neutral ones. (Carroll et al., 2022, p. 1)

Carroll et al. (2022) do not give specific examples of this but they are not hard to imagine. The phenomena of radicalization on platforms such as YouTube (Tufekci, 2018; Alfano et al., 2020) are, in part, the result of RL recommender systems learning that extremist, polarizing, or controversial content generates more engagement than neutral content. In this way, simple measurable metrics (clicks, watches, Likes, etc.) become proxies for discerning, and perhaps shaping, the preferences and values of the user. Twitter takes normally hard to measure aspects of friendship and complex desires associated with social engagement and boils them down to easily satisfied (or at least, easily measurable) metrics such as Followers, Likes, and

Retweets (Nguyen, 2020). Even though the attention economy does not manipulate or control us by hijacking, or causing us to *lose control over*, our attention, and even though targeted advertising is not reliably effective, automated influence still appears to have an effect, sometimes negative, on users and more collectively, on society as a whole. How, then, are we to make sense of these effects? Here, C. Thi Nguyen's (2020; 2022) notions of *value capture* and *value collapse* are enlightening.

4. Value Collapse

4.1 From Value Capture to Value Collapse

According to Nguyen, *value capture* occurs when an agent enters a social environment that presents external, simplified, and quantified expressions of value and "the simplified value *takes over* as the primary guide in my practical reasoning" (2020, p. 202). The world of automated influence can be considered one such environment. What is valuable is distilled down into simple metrics such as views, Likes, Retweets, clicks, and the like, the analysis of which is used to fuel targeted advertising, recommender systems, and search engines which, in turn, capture values associated with certain material goods, content, and so on. Instances of value capture abound. Take Nguyen's (2020) example of the FitBit. A FitBit measures the number of steps that you take per day. Without context (i.e., diet, the activity that produced those steps, time spent outside, etc.), steps per day are at best a poor metric for overall health. However, because steps per day are a clear, measurable, and easy to understand metric, it becomes convenient, and perhaps even enticing, to associate health, a multi-faceted, complex concept and value, with steps per day. In this way the value of health becomes *captured* by the metric of steps per day. Similarly, a user may join Twitter with the goal of forming friendships and connecting with strangers, but come to obsess over numbers of Followers and Retweets that are incapable of capturing the richness and subtlety of something like friendship. With these examples in mind, Nguyen lays out a four step-process that outlines how value capture occurs:

1. Our values are, at first, rich and subtle.
2. We encounter simplified (often quantified) versions of those values.
3. Those simplified versions take the place of our richer values in our reasoning and motivation.
4. Our lives get worse. (Nguyen, 2020, p. 201)

The conclusion that our lives get worse is based on the assumption that values are naturally "rich and inchoate" (Nguyen, 2020, p. 201), and are better when they are so. For example, the value of health, although more complex and multi-faceted than steps per day tracked on a FitBit, and thus harder to assess and maintain, is nonetheless better when conceived of in all its richness and subtlety. Health cannot be meaningfully distilled into steps per day. Without the contextual considerations of sleep, nutrition, habits and routines, stress and anxiety, etc., steps per day may be a good measure of nothing other than steps per day. Even then, many users cheat on 'exergame' apps, incentivized by the enticing metrics to devise ways of driving up the step counter without actually exercising (Lee and Lim, 2017). In this way, capturing rich and subtle values such as health and friendship in overly simplified metrics makes them ultimately harder to obtain. This then leads to *value collapse*.

Value collapse can be viewed as the result of sustained *value capture* and can similarly be outlined in a four-step process. Here, values are intimately related to attention.

1. Values drive attention.
2. Explicit values place clear boundaries on the attention.
3. Experiences that might drive us to improve our value formulation typically fall outside those boundaries.

> 4. In the grip of explicit values, we won't be motivated to pay attention to things that might inform us of gaps in our explication of value. (Nguyen, 2022).

In value collapse it becomes easy to dismiss anything that falls outside of these clear and explicit values, leading to a kind of "inattention feedback loop." Values are not only captured by these quick-and-easy metrics, but they are collapsed because it becomes increasingly difficult to attend outside of those values.

Nguyen's claims, on their own, are subject to the same criticism of empirical underdetermination as the *control argument*. That values drive attention in a way that places such clear boundaries on the attention should not be taken for granted. There are, however, good empirical reasons to accept such an argument. A number of experiments have demonstrated the phenomenon of value-driven attention capture (Anderson et al., 2011; Anderson and Yantis, 2013; Anderson and Kim, 2019). In one experiment (Anderson et al., 2011), participants (unknowingly) learned to associate a particular target (a color) with a reward (money). On subsequent tasks, the presence of that color, though having nothing to do with the new target, significantly delayed the ability of the participant to correctly identify the new target; in other words, to stay on task. Further, the closer the target appears to the high-value distractor, the greater the delay. The authors conclude that these findings "establish that nonsalient,[7] task-irrelevant stimuli previously associated with reward slow visual search" (Anderson et al., 2011, p. 10369), and so argue that value-driven attention capture is distinct from the well-established roles of ongoing goals (top-down) and salience (bottom-up) in attentional control.[8]

Admittedly, although value and reward may be somewhat conflated in the above study, and although these experimental conditions do not adequately reflect the online world of automated influence, these results suggest evidence in favor of Nguyen's *value collapse* argument in the context of automated influence. The frequent and often potent rewards associated with the attention economy are well-documented (Twenge, 2017; Williams, 2018; Lindström et al., 2021). To the extent that these rewards come to be valued by the user, and this reward (whether associated with status, professional success, attracting a partner, etc.) is much greater than what a participant receives in an experimental setting, these results support the idea that what comes to be valued through automated influence does indeed place boundaries on the attention.

Other recent work (Berkman et al., 2017) also supports this way of construing a kind of value-attention feedback loop. Attention shapes self-control and adaptive choice by dictating what options enter the choice set for the subject at any given moment, "foregrounding their salient attributes" (Berkman et al., 2017, p. 423). In other words, what we attend to will affect our values and subsequent self-control decisions. This relationship is not a one-way street, however. Just as what we attend to will shape our values (by gating our choice set, the information we are exposed to, etc.), our values similarly shape what we choose to attend to. In the context of automated influence, as our values become more and more shaped and entrenched by the algorithmic machinations of its associated technologies, we may become more likely to direct our attention towards the bells and whistles, the simple metrics of success, the recommended content, and so on, that are indicative of automated influence. Thus, I suggest pernicious effects of the attention economy and automated influence are best explained as contributing to a kind of *value collapse*.

The argument advanced here is not that it's impossible to expand one's values when they are in the throes of *value collapse* because of an inability to exert attentional control,

but rather that expanding on and improving one's values becomes increasingly *unlikely* because one is not *motivated* to seek out the kinds of things that lie beyond the captured values. This avoids the problems associated with underselling the ability of users to exert attentional control, while also explaining how automated influence can be detrimental to autonomy even though targeted advertisements are largely ineffective. Moreover, *value collapse* can explain how the effects of automated influence are harmful at both the level of engagement/data collection, and the level of delivery. Let's take each of these levels in turn to see the advantages of explaining the harmful effects of automated influence in terms of *value collapse*.

4.2 Value Collapse at the Level of Engagement

In order to understand the harmful effects of automated influence, we need to understand much more than just what happens at the instant the ad is delivered to a user. In order for automated influence to operate, data must first be collected to construct user profiles that are then used to target ads. In order for an attention economy to exist, attention must first be made legible (Scott, 1998; Fourcade & Healy, 2017; Hwang, 2020). As Hwang remarks:

> The need to create a liquid market in human attention influences the architecture of the social spaces of the web. Commodification requires attention to be legible: in other words, the internet must structure "engagement" in a way that is easy and accurate to measure. (Hwang, 2020, p. 115)

The very fact of being able to intelligibly speak of an "attention economy" demands that attention is distilled into a metric that is easy and accurate to measure. This pursuit of legibility is the first step towards *value capture*, and subsequently, *value collapse*.

At the level of engagement/data collection, this pursuit of legibility affects what is valued about the user, and by the user. These are, ultimately, the same thing: the thin metrics that are used to measure behavior, preferences, and values online. Just as we are better off to consider values in all their richness and complexity, the same goes for behavior. As a recent Whitepaper from the AI Objectives Institute (AIOI) remarks:

> Idealized markets and democratic systems assume that humans buy or vote in ways that reflect their long-term interests and values. To the extent that this form of sovereignty is true, human behavior helps to keep our institutions aligned. Yet even if people know what their long-term interests and values are, even if they are materially secure, there is no guarantee that they will choose actions that will be in service of those same values, (AIOI Whitepaper, p. 12, 2023)

In other words, humans are not always perfect rational actors. We act in ways that do not always perfectly reflect our preferences and values, so a snapshot of behavior at time *t* does not necessarily carry with it the weight of inferential power necessary to effectively deliver recommendations and ads that accurately reflect a user's values. One could argue that behavioral metrics are continually being collected across various domains which can help alleviate this problem, but the kinds of behavioral metrics used throughout the automated influence infrastructure are themselves poor representations of the richness and complexity of behavior. Likes, clicks, views, and Retweets can only capture so much. These Likes, clicks, and views then are not only poor metrics of behavior online, but cannot account for individuals as complex social beings that exist outside of an online environment.

A specific example comes from the world of mobile health apps, specifically those designed for helping people face mental health challenges. As O'Brien and colleagues note, the success of users who use mobile health apps is often evaluated according to user

engagement, typically operationalized as frequency and duration of app use, behavioral interaction with the app (for example downloads, clicks) and popularity (for example user reviews, ratings). Usage data is assumed to capture different types and depths of app engagement, yet misses cognitive and emotional responses to the app and is disconnected from behavior change in real-world settings (O'Brien et al., p. 1, 2020).

In this sense, what the purveyor of the app has deemed valuable about the user and what they use to drive further development, to make recommendations (and perhaps, to deliver ads), has been *captured* by the impoverished metrics used to measure success and engagement on the app. This *value capture* fails to account for the behavioral complexity of its users and the different experiences that people may have on the app.[9] The cognitive and emotional responses of the users on the app cannot be boiled down to or meaningfully inferred from thin metrics such as downloads and clicks. The same goes for automated influence writ large, including the construction of user profiles for targeted advertising we have discussed throughout. These profiles will necessarily be impoverished representations of the behaviors and values of users because the quantifiable metrics used to construct them are designed for legibility, not to account for the complexity of human behaviors and values.

Now we turn briefly to what Nguyen has more front-of-mind when discussing *value capture*. Namely, that those simplified metrics come to replace *our own* values. Above, we saw examples of how this could happen: step-counts on FitBit may come to act as stand-in for health in your reasoning and motivation, even though step-counts alone cannot account for the complexity of individual health. Similarly, you may come to think that having many Likes and Followers on Twitter means that you are having interpersonal success. One of the core ways that this is accomplished is through what Nguyen (2020) calls gamification. As Nguyen (2020) remarks: "Games can present us with a fantasy of value clarity. And if we are too seduced by that fantasy, we may be moved to oversimplify our own values" (p. 194). In gamification, the game like metrics that oversimplify our own values come to act as incentives, and so take the place of our real values in our reasoning and motivation. Because values drive attention and we are motivated by the value clarity on offer by these platforms, we will become less and less likely to attend to experiences, options, values, and the like, that fall outside of these simple metrics. In other words, our values *collapse.* It is not that we *cannot* pay attention to other things, it is that we are not *motivated* to pay attention to other things. Note how distinct this is from the *control argument* which suggests that we are at the mercy of the *persuasive design* that causes of to *lose control over* our attentional resources. Yes, the design of simple game-like metrics is capable of seducing us, but not of the kind of control that this argument suggests. The *value collapse* account does not hinge on a false theoretical-dichotomy between top-down and bottom-up attention. Rather, in order to make sense of where attention gets allocated, we must take into account motivational and affective salience, as the PSS model of attentional control (Tood & Manaligod, 2018) and other similar priority structure accounts that take values and attention to be deeply intertwined do (Awh et al., 2012; Watzl, 2023).

At the level of engagement, then, values are *captured* by the thin, quantifiable metrics employed to assess user behavior. These metrics are used by developers, advertisers, and the like, to develop new products, make recommendations, and deliver ads. Despite what they may think, however, these metrics are not necessarily reliable indicators of long-term user interests and values. Nonetheless, under the right conditions (i.e., gamification), these easy-to-measure quantifiable values

may come to replace the more rich and subtle values of users. This leads to a kind of value *collapse*, wherein attending outside of these constructed, simplified values becomes increasingly difficult. Crucially, this account avoids many of the problems associated with the *control argument*.

4.3 Value Collapse at the Level of Delivery

In a similar yet distinct fashion, these phenomenon also occur at the level we more immediately associate with automated influence, the level of delivery. It is at this level that recommendations are made and ads are delivered with the goal of influencing our behavior. This is generally accomplished by tightly associating the thing being sold and/or recommended with some generally desired character trait, social status, or lifestyle. Waide (1987) captures this phenomenon nicely:

> In order to increase sales, the advertiser identifies some (usually) deep-seated non-market good for which people in the target market feel a strong desire. By 'non-market good' I mean something which cannot, strictly speaking, be bought or sold in a marketplace. Typical non-market goods are friendship, acceptance and esteem of others (Waide, 1987, p. 1).

Thus, the values associated with things like friendship and acceptance become *captured* by the advertised good. In addition to driving sales, strengthening the association between a particular brand and a value may be the goal of much advertising.

As Hwang (2020) notes: "Brand advertising (. . .) is less about the immediate purchase and more about shaping the public's associations with a brand and differentiating it from its competitors" (p. 79). For example, Nike may invest in advertising not primarily to boost sales (although of course this is also a desired outcome) but to maintain its status as a "cool" brand in the public eye. If this works well enough, then someone looking to buy a "cool" pair of sneakers, or more generally, someone with the desire to be "cool," may unreflectively think of and purchase Nike products. Similarly, Apple may invest in advertising to strengthen the brands association with professional success and ingenuity, making someone with these values more likely to buy Apple products. What is happening in these situations is a kind of *value capture*. Complex and rich values such as professional success, social acceptance, ingenuity, and even "coolness," are simplified in a way that makes them, in theory, easier to measure and perhaps, to achieve. This is, of course, an illusion, but if the advertisement indirectly causes an individual to buy into this illusion, then it has accomplished its goal. The advertisement need not lead to a click or a sale for this *value capture* to occur.

Automated influence also need not have a widespread effect at the individual level to lead to the kind of "inattentional blindness" associated with *value collapse*. As noted above, by many measures, targeted advertising is relatively ineffective, and so accounts that rely on caching out the harms associated with automated influence in terms of behavioral modification and control seem, at face value, to be at a loss. However, behavior modification can occur without the kind explicit control discussed by Williams (2018) and moreover automated influence can be relatively ineffective at the individual level while still having outsized aggregate effects. Benn & Lazar (2022) refer to these aggregate effects of automated influence as "*stochastic* manipulation" (p. 141). As they argue, online targeted advertising is not much more effective than other forms of advertising and even in cases where it is, "it's hard to get too riled up about being nudged into consuming a little more than your budget allows or spending more time than you think you should staring at a screen" (Benn & Lazar, 2022, p. 141). However, at the aggregate level of *stochastic* manipulation, these effects of automated

influence on individuals, however rare and ineffective, can have widespread and harmful effects nonetheless. As they suggest: "From the perspective of each individual consumer, choosing one product rather than another may make little difference. But at the aggregate level, the inevitability that digital platforms will shape our purchasing choices can lead to serious anticompetitive results" (Benn & Lazar, 2022, p. 142). Benn & Lazar are rightfully worried about the consolidation of power that results from these anticompetitive results, but what I want to emphasize here are how these results contribute to *value collapse*.

There are only a few major digital platforms that hold sway over the online advertising marketplace, and to the extent that we attend to these ads (this can be relatively little but still have structural effects), certain values will be *captured* by them. The anticompetitive result of these aggregate effects is that what is valuable and how to achieve certain values becomes more and more dictated by those who have consolidated power in these spaces. Our values (i.e., what it means to be healthy, successful, happy, and attractive) are *collapsed* into the advertised or recommended good. To return to Waide (1987), he suggests: "In some cases, the product actually gives at least partial satisfaction to the non-market desire—but only because of advertising. (. . .) We become enforcers for the advertisers" (p. 74). In this way, even relatively ineffective targeted advertising and their associated marketing claims (wear Nikes if you want to be "cool") can become self-fulling if the ads reach enough people. The ability of AI to deliver targeted ads at an unprecedented scale ensures that they do.

The effects of *stochastic* manipulation then have the similar effect of placing clear boundaries on the attention in the same way that clear, quantifiable metrics do at the level of engagement. As the values that are *captured* by automated influence come to dominate our field of view, both literally and figuratively, these values begin to drive our attention. In this way, behavior modification at the hands of automated influence is both possible and pernicious, but one needs to look at the aggregate level to see this due the relative ineffectiveness of targeted advertising at the individual level. Moreover, this behavior modification does not occur because of the ability of advertisers or tech companies to exert some kind of explicit control over individuals, but because of the aggregate effects that drive collective values and attention.

5. Conclusion

What I have argued here is that *value collapse* offers a promising framework for better understanding the moral and psychological harms associated with automated influence. The *control argument* paints a picture of automated influence as capable of exerting direct control over the behavior of individuals. However, as I have outlined, these accounts rest on false assumptions both about the nature of attention and control, as well as the effectiveness of automated influence, especially targeted advertising, to modify behavior at the individual level. Whereas the *control argument* is not specific about the nature of attention, the targets of automated influence, or how such control happens, the account developed here is empirically responsible and clear about why automated influence is morally and psychologically problematic. Using Nguyen's (2020; 2021; 2022) frameworks of *value capture* and *value collapse* to account for these harms avoids these empirical pitfalls while being able to offer a convincing account of these moral and psychological worries despite the inability of automated influence to exert this kind of widespread direct control.

As noted above, because automated influence is one of the domains in which AI is the most pervasive in our lives, having a proper framework to understand any harms associated with automated influence is of

the utmost importance. This is especially true at our current moment when we are working to meaningfully regulate AI, and as these technologies continue to advance at unprecedented speeds. Amongst the ways that AI poses a threat to human-autonomy and self-determination, automated influence is foremost in its reach and ability to shape our day-to-day lives. This does not happen, however, through the direct control of individual behavior but through the subtler effects of *value capture* and *value collapse*.

Department of Philosophy
University of Guelph
Guelph, ON
N1G 2W1
dwhite11@uoguelph.ca

NOTES

Dylan J. White is supported in part by funding from the Social Sciences and Humanities Research Council.

1. Throughout, I primarily draw on examples of automated influence from targeted advertising, but the arguments advanced here apply to all forms of automated influence.

2. Although not the focus of the current essay, to the extent that large language models (LLMs) are poised to alter the landscape of automated influence (Susser and Grimaldi, 2021; Burtell and Woodside, 2023) by allowing for more effective targeted ads, user specific content, etc., developing such a framework now is of the utmost importance. However, the opposite could also be true; generative-AI powered ads could oversaturate an already overvalued market (Hwang, 2020; Ball, 2023). Even if we grant, however, that targeted ads, recommender systems, and so on will improve, I suggest that they will not improve because of their ability to literally *control* people, but because they may become more effective at shaping our values, motivations, and choice-architectures such that we will be increasingly likely to attend to what a given company, platform, etc., wants us to attend to.

3. However, it may be just as accurate to describe the attention economy as the market where users give their *data* in exchange for these services. Attempting to capture attention through *persuasive design* is a more-or-less reliable way of collecting such data. See Hwang (2020) for more.

4. Attention is never clearly defined by most proponents of the *control argument,* a problem I take up in detail elsewhere (White, 2024). In defining attention, I follow Wu (2023) in taking as my starting point a Jamesian *common ground* view; namely, that attention solves a *selection problem* for the organism in question. Attention, then, is selection for guiding action (including mental action). The *control argument* suggests that this selection for action is reliably and pervasively controlled by automated influence and the attention economy. For more on defining attention this way, and a brief overview of recent philosophical debates about attention, please see (White, 2024).

5. The targets that automated influence seek to control, though often left unspecified, are multifaceted—behaviour, attention, clicks, purchases, etc.,—and, often, inseparable. At the core of these, however, is attention. Therefore, ultimately, one can assume that it is the control over what gets attended to that automated influence and the attention economy targets, according to the control argument. Thank you to an anonymous reviewer for urging me to clarify this.

6. Elsewhere, I argue that recent policy recommendations that assume something like the *control argument* are unlikely to be effective at tackling the moral and psychological harms associated with automated influence and the attention economy (White, 2024).

7. Here, nonsalient means the participant had no prior disposition to attend to one particular color or target over another.

8. They go to suggest that "value-driven attentional capture may play a key role in a variety of clinical syndromes in which both attention and reward have been critically implicated" (Anderson et al., 2011, p. 10370). Although the harms associated with automated influence are not best understand as any kind of clinical syndrome, attention and reward are certainly implicated in these harms.

9. Not all instances of *value capture* such as this are necessarily harmful. In fact, as a recent study suggests, the widespread use of mobile health apps as well as wearable and ambient biosensors "have set the stage for the development of multimodal artificial intelligence solutions that capture the complexity of human health and disease" (Acosta et al., 2022, p. 1).

REFERENCES

Acosta, Julián N., Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. 2022. "Multimodal Biomedical AI," *Nature Medicine*, vol. 28, 1773–1784.

"AI Objectives Institute Whitepaper—A Research Agenda for the Production of a Flourishing Civilization." 2023. *AI Objectives Institute*. https://aiobjectives.org/whitepaper

Alfano, Mark, Amir Ebrahimi Fard, J. Adam Carter, Peter Clutton, and Colin Klein. 2021. "Technologically Scaffolded Atypical Cognition: The Case of YouTube's Recommender System," *Synthese (Dordrecht)* vol. 199, no. 1–2, 835–858.

"ANA Programmatic Media Supply Chain Transparency Study—First Look." 2023. *Association of National Advertisers*. https://www.ana.net/miccontent/show/id/rr-2023-06-ana-programmatic-transparency-first-look

Anderson, Brian A., Patryk A. Laurent, and Steven Yantis. 2011. "Value-Driven Attentional Capture," *Proceedings of the National Academy of Sciences—PNAS* vol. 108, no. 25, 10367–10371.

Anderson, Brian A., and Steven Yantis. "Persistence of Value-Driven Attentional Capture," *Journal of Experimental Psychology. Human Perception and Performance* vol. 39, no. 1, 6–9.

Anderson, Brian A., and Haena Kim. 2019. "On the Relationship Between Value-Driven and Stimulus-Driven Attentional Capture," *Attention, Perception, and Psychophysics*, vol. 81, 607–613.

Aranda, Julie H., and Safia Baig. 2018. "Toward "JOMO": The Joy of Missing Out and the Freedom of Disconnecting," Proceedings of the 20th Internatinoal Conference on Human-Computer Interaction with Mobile Devices and Services, 1–8.

Awh, Edward, Artem V. Belopolsky and Jan Theeuwes. 2012. "Top-Down Versus Bottom-Up Attentional Control: A Failed Theoretical Dichotomy," *Trends in Cognitive Sciences*, vol. 16, no. 8, 437–443. https://doi.org/10.1016/j.tics.2012.06.010

Aylsworth, Timothy. 2020. "Autonomy and Manipulation: Refining the Argument AgainstPersuasive Advertising," *Journal of Business Ethics* vol. 175, no. 4, 689–699.

Aylsworth, Timothy, and Clinton Castro. 2021. "Is There a Duty to be a Digital Minimalist?" *Journal of Applied Philosophy*, vol. 38, no. 4, 662–73. https://doi.org/10.1111/japp.12498

Ball, James. 2023. "Online Ads Are About to Get Even Worse." *The* Atlantic.https://www.theatlantic.com/technology/archive/2023/06/advertising-revenue-google-meta-amazon-apple-microsoft/674258/

Beck, Diane M., and Sabine Kastner. 2009. "Top-Down and Bottom-Up Mechanisms in Biasing Competition in the Human Brain," *Vision Research,* vol. 49, no. 10, 1154–65. https://doi.org/10.1016/j.visres.2008.07.012

Benn, Claire, and Seth Lazar. 2022. "What's Wrong with Automated Influence," *Canadian Journal of Philosophy* vol. 52, no. 1, 125–148.

Berkman, Elliot T., Cendri A. Hutcherson, Jordan L. Livingston, Lauren E. Kahn, and Michael Inzlicht. 2017. Self-Control as Value-Based Choice. *Current Directions in Psychological Science,* vol. 26, no. 5, 422–28.

Bermúdez, Juan Pable. 2017. "Social Media and Self-Control: The Vices and Virtues of Attention," in *Social Media and Your Brain*, ed. Prado, C. G.. (Praeger).

Bhargava, Vikram R., and Manuel Velasquez. 2021. "Ethics of the Attention Economy: The Problem of Social Media Addiction," *Business Ethics Quarterly,* vol. 31, no. 3, 321–59.

Björn, Lindstrom, Martin Bellander, David T. Schultner, Allen Chang, Phillipe N. Tobler, and David M. Amodio. 2021. "A Computational Reward Learning Account of Social Media Engagement," *Nature Communications*, vol. 12, 1311.

Burtell, Matthew and Thomas Woodside. 2023. "Artificial Influence: An Analysis of AI-Driven Persuasion," https://doi.org/10.48550/arXiv.2303.08721.

Carroll, Micha, Anca Dragan, Stuart Russel, and Dylan Hadfield-Menell. 2022. "Estimating and Penalizing Induced Preference Shifts in Recommender Systems," Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland.

Castro, Clinton, and Adam K. Pham. 2020. "Is the Attention Economy Noxious? *Philosophers' Imprint*, vol. 20, no. 17.

Corbetta, Maurizio, and Gordon L. Shulman. 2002. "Control of Goal-Directed and Stimulus- Driven Attention in the Brain," *Nature Reviews. Neuroscience*, vol. 3, no. 3, 201–215. https://doi.org/10.1038/nrn755

Evans, Charles and Atoosa Kasirzadeh. 2021. "User Tampering in Reinforcement Learning Recommender Systems," https://doi.org/10.48550/arXiv.2109.04083.

Eyal, Nir. 2013. *Hooked: How to Build Habit-Forming Products* (Penguin Canada).

Fourcade, Marion, and Kieran Healy. 2017. "Seeing Like a Market," *Socio-Economic Review*, vol. 15, no. 1, 9–29.

Ganeri, Jonardon. 2016. *Attention, Not Self* (Oxford University Press).

Hwang, Tim. 2020. *Subprime Attention Crisis: Advertising and the Time Bomb at the Heart of the Internet* (New York, Farrar, Straus and Giroux).

Inzlicht, Michael. & Schmeichel, Brandon. J. 2012. "What Is Ego Depletion? Toward a Mechanistic Revision of the Resource Model of Self-Control." *Perspective on Psychological Science*, vol. 7, no. 5, https://doi.org/10.1177/1745691612454134

Inzlicht, M., Schmeichel, Brandon. J., & Macrae, C. Neil. 2013. "Why Self-Control Seems (but may not be) Limited." *Trends in Cognitive Sciences*, vol. 18, no. 3), 127–133. https://doi.org/10.1016/j.tics.2013.12.009

Jennings, Carolyn Dicey. 2020. *The Attending Mind* (Cambridge University Press).

Jonides, John. 1981. "Voluntary Versus Automatic Control Over the Mind's Eye's Movement, in *Attention and Performance IX*, ed. Long, J. B., and Baddeley, A.D., (Lawrence Erlbaum Associates), 187–203.

Kardaras, Nicholas. 2016. "It's 'Digital Heroin': How Screens Turn Kids Into Psychotic Junkies." *The New York Post*, August 26. http://nypost.com/2016/08/27/its-digital-heroin-how-screens-turn-kids-into-psychotic-junkies/

Lee, Yeoreum, and Youn-Kyung Lim. 2017. "How and Why I Cheated on My App: User Experience of Cheating Physical Activity Exergame Applications," Proceedings of the 2017 Conference on Designing Interactive Systems.

Lustig, Robert. 2017. *The Hacking of the American Mind: The Science Behind the Corporate Takeover of our Bodies and Brains* (New York: Avery).

Milano, Silvia, Mariarosaria Taddeo and Luciano Floridi. 2020. "Recommender Systems and Their Ethical Challenges," *AI & Society*, vol. 35, 957–967.

Navon, David. 1984. "Resources—a Theoretical Soup Stone? *Psychological Review*, vol. 91, no. 2, 216–234. https://doi.org/10.1037/0033-295X.91.2.216

Nguyen, C. Thi. 2020. *Games: Agency as Art* (Oxford University Press).

———. 2021. "The Seductions of Clarity," *Royal Institute of Philosophy Supplement*, vol. 89, 227–255.

Nguyen, C. Thi. 2022. "Value Collapse," The Royal Institute of Philosophy Cardiff Annual Lecture 2022. https://www.youtube.com/watch?v=zt03qjTyefU

O'Brien, Heather L., Emma Morton, Andrea Kampen, Steven J. Barnes, and Erin E. Michalak. "Beyond Clicks and Downloads: A Call for a More Comprehensive Approach to Measuring Mobile-Health App Engagement, " *BJPsych Open* vol. 6, no. 5, 86.

Pickard, Hannah. 2022. "Addiction and the Meaning of Disease. In Heather, N., Field, M., Moss, A. C., & Satel, S., eds., *Evaluating the Brain Disease Model of Addiction*, 321–338. Routledge.

Posner, Michael I. 1980. "Orienting of Attention," *The Quarterly Journal of Experimental Psychology*, vol. 32, no. 1, 3–25.

Rieser, Lars, and Brent Furneaux. 2022. "Share of Attention: Exploring the Allocation of User Attention to Consumer Applications," *Computers in Human Behavior*, vol. 126, 107006–. https://doi.org/10.1016/j.chb.2021.107006

Rini, Regina. 2023. "Your Attention, Please!" *Times Literary Supplement*, July 7. https://www.the-tls.co.uk/articles/your-attention-please-afterthoughts-regina-rini/

Rose-Stockwell, Tobias. 2023. *Outrage Machine: How Tech Amplifies Discontent and Disrupts Democracy—and What We Can Do About It* (Hachette Book Group).

Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed* (Yale University Press).

Shomstein, Sarah, Xiaoli Zhang and Dick Dubbelde. 2022. "Attention and Platypuses," *Wiley Interdisciplinary Reviews. Cognitive Science*, e1600–e1600. https://doi.org/10.1002/wcs.1600

Sripada, Chandra. 2020. "The Atoms of Self-Control," *Noûs,* vol. 55, no. 4, 800–824. https://doi.org/10.1111/nous.12332.

Susser, Daniel, and Vincent Grimaldi. 2021. "Measuring Automated Influence: Between Empirical Evidence and Ethical Values," Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21.

"The Global PPC Click Fraud Report, 2020–21." 2021. *Search Engine Journal*. https://www.searchenginejournal.com/the-global-ppc-click-fraud-report-2020-21/391493/

Todd, Rebecca M., and Maria G. M. Manaligod. 2018. "Implicit Guidance of Attention: The Priority State Space Framework," *Cortex*, vol. *102*, 121–138. https://doi.org/10.1016/j.cortex.2017.08.001.

Tufekci, Zeynep. 2018. "YouTube, the Great Radicalizer," *New York Times*, July 20. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html.

Turel, Ofir, & Qahri-Saremi, Hamed. 2016. "Problematic Use of Social Networking Sites: Antecedents and Consequence from a Dual-System Theory Perspective. *Journal of Management Information Systems*, vol. 33, no. 4, 1087–1116. https://doi.org/10.1080/07421222.2016.1267529.

———. 2018. "Explaining Unplanned Online Media Behaviors: Dual System Theory Models of Impulsive Use and Swearing on Social Networking Sites. *New Media & Society*, vol. 20, no. 8, 3050–3067. https://doi.org/10.1177/1461444817740755.

Twenge, Jean. 2017. *iGen: Why Today's Super-Connected Kids Are Growing Up Less Rebellious, More Tolerant, Less Happy—and Completely Unprepared for Adulthood* (Simon & Schuster).

UNESCO. 2023. "Global Education Monitoring Report, 2023." *Technology In Education: A Tool on Whose Terms?* https://unesdoc.unesco.org/ark:/48223/pf0000385723

U.S. Surgeon General's Advisory. 2023. "Social Media and Youth Mental Health." https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf.

Waide, John. 1987. "The Making of Self and World in Advertising," *Journal of Business Ethics* vol. 6, no. 2, 73–79.

Watzl, Sebastian. 2023. "What Attention Is. The Priority Structure Account," *Wiley Interdisciplinary Reviews. Cognitive Science*, vol. 14, no. 1, e1632–n/a. https://doi.org/10.1002/wcs.1632.

White, Dylan J. 2024. "Paying Attention to Attention: Psychological Realism and the Attention Economy," *Synthese,* vol. 203, no. 2, 43. https://doi.org/10.1007/s11229-023-04460-4

Williams, James. 2018. *Stand Out of Our Light: Freedom and Resistance in the Attention Economy* (Cambridge University Press).

Wu, Tim. 2017. "The Crisis of Attention Theft—Ads That Steal Your Time for Nothing in Return. *Wired*, Business, April 14, 2017. https://www.wired.com/2017/04/forcing-ads-captive-audience-attention-theft-crime/.

Wu, Wayne. 2023. "On Attention and Norms: An Opinionated Review of Recent Work."*Analysis (Oxford)*. https://doi.org/10.1093/analys/anad056.

Yantis, Steven, and John Jonides. 1990. "Abrupt Visual Onsets and Selective Attention: Voluntary Versus Automatic Allocation," *Journal of Experimental Psychology. Human Perception and Performance* vol. 16, no. 1, 121–134.