The Commercialization of Decision-Making: Towards a Regulatory Framework to Address Machine Bias over the Internet

By Dipayan Ghosh, Harvard Kennedy School

Abstract

The consumer internet has exacerbated the discrimination problem. The business model that sits behind the front end of the internet industry is one that focuses on the unchecked collection of personal information, the continual creation and refinement of behavioral profiles on the individual user, and the development of algorithms that curate content. These actions all perpetuate the new pareto optimal reality of the commercial logic underlying the modern digitalized media ecosystem: that every act executed by a firm, whether a transfer of data or an injection of content, is by its nature necessarily done in the commercial interests of the firm because technological progress has enabled such granular profiteering. This novelty in the media markets has created a tension in the face of the public motive for nondiscriminatory policies; where adequate transparency, public accountability, or regulatory engagement against industry practices are lacking, it is directly in the firm's interest to discriminate should discriminatory economic policies suit its profit-maximizing motive. This paper discusses this technological development and offers policy responses to counteract these breaches against the subjects of internet-based discrimination.

Introduction: The Centrality of the Consumer Internet

The importance of the consumer internet in the context of the modern media ecosystem is unquestionable. Economic opportunities in housing, employment, and other objects of the consumer marketplace; national political concerns and the systemized dissemination of political communications; and social interactions that mirror or conversely define our sociocultural norms: these are all clear and evident results of the growth and present breadth of influence of the consumer internet.

The consumer internet is comprised of the firms that operate over the internet and interface directly with consumers—Facebook, Apple, Google, Netflix, Spotify, and Amazon among them. Consistent across the sector is a set of practices—constituting the business model that sits at the heart of the internet—driven by (1) the development of tremendously engaging platforms that surface the ranked content that the firms predict consumers most wish to see and will therefore engage with; (2) the uninhibited collection of the consumer's personal information all to the end of generating behavioral profiles on the consumer that record the consumer's likes, dislikes, preferences, interests, routines and behaviors; and (3) the refinement of highly sophisticated but equally opaque algorithms that curate content to fuel the first practice and target ads by taking advantage of the second one. This economic engine, consistent across the consumer internet, is depicted in Figure 1.

To be sure, there are two caveats to clear before moving forward. First, while this business model is in clear use within the walled gardens of such firms as Facebook and Amazon, each firm adopts it in its own way, using its own proprietary processes and propensity for personal data collection along with its singular value proposition for the consumer market, to take advantage of the profits the general business model can yield. And second, this business model is utilized to varying degrees by the subject firm; that is to say that there may be other core practices and contributions to company revenue that are also critical to the subject firm and operated in parallel to the consumer internet offering. To mention a few examples, Amazon and Google are market leaders in the provision of cloud computing services; Apple's core revenue is generated from the sale of consumer device technologies; and Netflix maintains an order-by-mail video rental service that has relatively little to do with the aforementioned business model leaving aside the agglomerations of personal interests derivable from physical rentals. But to reiterate, it is the set of practices leading to the business model illustrated above that constitutes what I mean by the "consumer internet"—and which I hope to scrutinize further and critique in this essay. This is particularly because it is this business model that has instigated and perpetuated the negative externalities that we care about protecting the public from today and in the way forward—precisely because the business model has promoted an insidious economic logic that aligns the interests of nefarious actors with those of the internet platform firms.

To ask *why* these firms have uniformly adopted this business model is pertinent. As others have discussed, the internet industry is one that operates in a free commercial zone—it is, in other words, a radically free market that favors and rewards open capitalism.[1] In such jurisdictions as the United States, we continue to lack a federal standard on privacy and most other public interest concerns that would otherwise concern the firms in this sector. This fundamental lack of consumer and citizen rights in the United States has enabled the internet firms to have a free pass to take advantage of the free market zone. And take advantage they have, just as suggested by the institutional directive within Facebook to "move fast and break things."

This lack of a regulatory regime has in turn meant that these firms have developed in a manner practically independent and uncaring of the public interest save when it serves their commercial interests. As with any business, the public interest need not be considered from their perspective; only the shareholders need be served. It is thus effectively unnatural to ask a chief executive in the industry to bend the knee to consumers; in the absence of meaningful economic regulations that target the capitalistic overreaches of the business model, nothing can save the public from the overreaches of the industry. And while public sentiment might swell to such a degree at times that it might appear the effective situation for the firms has changed for good because of ongoing public outcries, the lack of actual regulatory movement by the government equates to the free zone of commerce remaining intact. The public's memory is short; equivalently, the industry often moves directly back into the zone of commercial operation it did prior, perhaps under new disguises to protect itself from regulators. One could suggest that this is precisely what has happened in regard to the Cambridge Analytica disclosures of March 2018; while there was a cacophony of public outrage immediately following the whistleblower's revelations and the corresponding reports of sharing of 87 million Facebook users' information with the digital strategy firm engaged by Donald Trump's campaign for the presidency—outrage that led Mark Zuckerberg to testify before Congress mere weeks after the revelations—there is relatively little U.S.-led discussion now about what economic regulations should be passed to truly hold Facebook and like firms to account.[2] The stunted progress of the Honest Ads Act introduced by U.S. senators is 'Exhibit A.'[3]

It is due to this unrestrained progress of the business model—particularly the constant quest by firms like Facebook and Google to maximize through whatever means necessary and possible the amount of time users spend on the platforms—that the leading internet platforms have overtaken the media ecosystem in the United States.

The Science of Machine Bias

Robust discussion has developed in recent years, particularly since the boom of the big data economy, concerning the potential of machine learning algorithms to systematically perpetuate discriminatory results in various fields from medical science to educational opportunities.[4] Of primary concern is the development of machine learning models that engage in automated decision-making. While the methods underlying the application of machine learning are mostly taken from the traditional statistical literature and largely do not constitute mathematical novelty in general terms, cultural circumstances and advances in computing have popularized the term and expanded interest in the industry.

Supervised machine learning models are typically designed through a combination of human input and automated statistical analysis of a dataset. A dataset such as the demographic inferences Google draws over a class of users in a given region typically carries some implicit pattern; different classes of users might execute searches at various times of day, from various locations, and with varying frequencies—indicators called "features" since they are independent attributes associated with an instance, in this case an individual user. Machine learning models attempt to draw such relationships out of the data to develop inferences about the true nature of the population. A human—or in the case of unsupervised models, a machine—might code each user as participant in a particular class based on the user's individual features, inferred through analysis of the user's on-platform behavior, off-platform activity, and demographic data, thereby generating a set of "training data" that can be used to help the model learn how to classify future data points. A dataset including a population of such users might have some observable relationships consistent between certain classes of the population. These relationships are drawn into the machine learning "model" in the form of a set of decision rules—a series of inferences about the population developed from observation of the dataset that can then be implemented as a "classifier" of future objects subject to the model's classification regime. This implementation can then be executed on an automated basis such that when new observations come into view they can readily be analyzed and classified by the model.[5] The continual refinement of learning models through feedback from real world routines and behaviors is illustrated in Figure 2.

Taking YouTube as an example, we would regard the platform's video recommendation system by which the user is suggested a video to watch next as the model or classifier, which operates over a set of decision rules established by the machine learning model developed and refined by the company on an ongoing basis. The company's commercial objective is to engage the user,

thereby enabling it to collect increasing amounts of information about the user's habits and preferences, and to generate ad space that it can sell to the highest bidder interested in persuading a set of users. Particularly at the outset of algorithmic design, a team of humans might be employed to classify a global set of users into various categories for each feature. A feature concerning type of use of the platform, for instance, might include classes such as "channel operators", "power users", "frequent users" and "occasional users." Additional features might include demographic details, information pertaining to the user's historical use of the platform including which videos the user watches and which channels he or she has subscribed to, information pertaining to the use of other Google services, and position in the relationship graph network among others. YouTube then might train a model that analyzes how the existing "observed" data points concerning the company's users were classified. This analysis of observed data points is used to develop and train, on an ongoing basis, a set of decision rules that constitute the classifier model that can determine based on statistical analysis which class new data points—new users of YouTube, for example—should enter. Finally, this algorithmic inference then determines what videos the platform will recommend to the user. Feedback loops incorporating accuracy of predictions (whether reported by the user or inferred by the platform due to a user's disengagement or other negative behavior) can be used to refine the model over time. This in turn leads many users down a path of watching a long series of highly engaging videos described by some as going down the YouTube "rabbit hole."

Many have described machine learning models—and more generally, algorithmic processes—as fair, or at the least, fairer than a human would be in making the same decision. This idea has been wholly rejected by most. While theoretically algorithms could be designed in a manner that is contextually "fair," one question that naturally arises is what fairness (even in context) should actually mean; different parties might have different definitions in practice, and even with consensus on the meaning of fairness, machine learning models have been shown to discriminate. Another concern is that it has proven to still be difficult to design machine learning algorithms in a manner that foresees all potential forms of fairness and preempts them through reorientation of the algorithm. In the case of YouTube, for example, reports suggest that the recommendation algorithm has had a longstanding tendency to suggest users watch conspiracy-laden videos including the "Momo" hoax that targeted children[6] and the "flat-earther" myth.[7]

At issue is the propensity for most machine learning models to discriminate; in fact, this is precisely what they are meant to do: discern the characteristics of an incoming data point and infer, based on its features, which class it belongs to. Presumably, such models are used to give potentially different treatment to data points that occur in different classes. In the case of YouTube, established sports fans might consistently be recommended to watch videos related to sports; those interested in foreign policy might be subject to recommendations to watch political videos.

Title VII of the Civil Rights Act in the United States offers protection from unfair decisions made on the basis of any protected categories, including race, gender, pregnancy, religion, creed, veteran-status, genetic testing status, ancestry, and national origin. (Importantly, political discrimination is not included here.) Various laws institutionalize further protections, among them the Age Discrimination and Employment Act, which states that employers cannot terminate an employee simply because of age; there must be some substantiation that the employee no longer can work effectively. Similarly, the Americans with Disabilities Act prohibits employer discrimination against those individuals who can work effectively despite their disabilities. Various state laws go further than the federal laws and institute further protections from discrimination, particularly through added protected classes and other expansions including, for instance, new lower age thresholds to trigger the age discrimination law.

Developing civil rights jurisprudence carries two principal mechanisms for protection from discriminatory outcomes: disparate treatment and disparate impact.[8] In a typical disparate treatment case, a potential employer might suggest that the candidate should not be hired because he or she is the member of a protected category. This sort of determination would. Would amount to an intentional violating decision to discriminate against the candidate because of his or her protected class status.

But in the realm of machine bias, disparate impact cases are typically of greater concern because of the manner in which learning algorithms engage in automated classifications over which decisions—which could be vitally important to the data subject in question—are automatically applied and implemented against many data subjects together according to a set of rules contained in the model. Disparate impact cases typically refer to instances in which a particular decision has greater resulting impact on a protected group than on the rest of the population. Harmful disparate impact can trigger an investigation against the liable party. And a decision such as a hiring policy might be "facially neutral"—where the decision rule does not appear to be discriminatory on its face—but if when carried out in practice it results in a harmful disparate impact against a protected group then civil rights protections may be triggered.

It may be the case that a learning model used to classify users in a consumer internet application—for instance, in

the context of identifying the consumer group at which to target an ad campaign that includes a set of political messages—might attempt to maximize clickthrough rates or some other engagement or revenue metric applied by the platform firm. The learning model might identify characteristics in regard to a number of signals (or in this case, features) about the messaging and the advertiser's intended target audience, for instance that political ads feature men and masculine themes, as well as issues that may appeal to certain socioeconomic classes, and that the geographic region the advertiser wishes to target is in the Midwest. In such cases, it is likely that the algorithm will determine that the target audience that will yield greatest engagement for the advertiser and the platform is some group that is primarily male, wealthy, and Midwestern—which, it could be said, is a necessarily harmful discriminatory targeting practice given that certain protected classes are not included in the target audience. All that said, such targeting is likely not illegal for several reasons. First and foremost, there might be no civil rights laws that covers the content of the ad campaign in question since American laws primarily cover various economic opportunities but not social or political ones. Second and perhaps more critically, it might be the case that, even if the ad content is covered by civil rights laws and pursues a discriminatory execution of dissemination that prevents certain protected classes from seeing it, the classifier was technically "fair." In such cases if a suit is pursued then the platform firm that enabled the targeting may have to respond to the question of why the algorithm screened out an inordinate proportion of, say, women. Should the firm be able to offer a justifiable business reason then it could be adjudged that it did not engage in unfair discriminatory practices leading to harmful disparate impact.[9]

Broadly, the utilization of learning models can produce discriminatory outcomes through two main means: the nature of the training or input data, and the design of the learning algorithms themselves. Underlying each of these primary themes is a more human concern: that the data miner him- or herself could be (intentionally or unintentionally) biased and carry that bias into the programming of the model and analysis of the data.

Discriminatory Concerns Related to Training Data

There is a longstanding refrain in the field of computer science: "garbage in, garbage out." Machine learning models are "trained" through the analysis of the aforementioned training data, which in supervised learning schemes might be classified by humans. Data points—such as a typical Google user—has a set of attributes about his or her use of the company's platforms that can be used to classify the user into certain audience clusters. Inferences about new users to the company's systems are then made by the learning model. But as was discussed in a recent White House report, poor design of training data can promote discriminatory outcomes.[10] There are two primary mechanisms by which flaws in training data can perpetuate discriminatory decision-making.

The first is in the process by which the data is organized. Historical datasets on which training data is based typically come with certain mutually exclusive class fields as discussed above in the YouTube example—but the selection of class fields and attribution of data subjects to them occurs at the hands of humans in supervised learning premises. The individuals who organize these class fields—the data engineers responsible for development of learning models—attempt to define a paradigm through the identification of class fields that they believe most fairly and effectively reflects the situation of the real world. For instance, it might be the case that to make determinations about the creditworthiness of a loan applicant, credit agencies decide that it is most critical to understand his or her net worth, demographic information, profession, education level and relate categories—but that it is less important or particularly difficult to include information related to the individual's personal life goals, trustworthiness, and commitment to paying back the loan. This can germinate a form of bias in the designation of class fields, as such determinations to include and exclude certain categories could diminish the chances of a positive decision for certain demographic groups while elevating the chances of others. The creditworthiness example can be translated readily to the consumer internet context: firms continually refine ad-targeting algorithms so as to advance the commercial interests of the advertisers by offering them maximized bang for buck with the data that they have at hand. Whether the advertiser is a credit or housing or employment agency or another client, the tendency for all the parties at hand will be to promote profits over protecting the consumer's interest given the lack of any sort of legitimate nonpartisan scrutiny over firms in the digital advertising and consumer internet sectors.

The second major family of discrimination concerns that might arise from poor design of training datasets is attributable to the data that populates the datasets itself. Two main problems can be responsible for this: incorrect data and selection bias. In the first case, data might be outdated or otherwise contain inaccuracies about the population that perpetuates bias since the incorrect data is used to train the classifier model. For instance, if loan payback periods are incorrectly reported to be longer for some individuals than others, then those individuals might be adversely affected by the decisions executed by the resulting model trained on the inaccurate dataset. The second case, selection bias, is often subtler and involves the collection of data that is not representative of the population which, if used to train the resulting learning model, projects the inferences learned from the

biased training set on current decisions, likely resulting in biased decisions. A simple example of biased input data occurred in the case of the StreetBump application developed in Boston; the mobile application was designed to enable residents to report the occurrence of potholes to the app developer and the idea was seen as so successful in enabling crowdsourced reports that the municipality engaged the developers to know when and where to dispatch repair teams. After some time of use, however, it was found that repair teams were disproportionately dispatched to wealthier and younger neighborhoods—parts of the city that presumably had more people who owned smartphones and greater local propensities to participate in the crowdsourcing functionalities offered through the application. The city was, in other words, receiving a biased selection of the data; a truly representative set of data would report relative frequencies of potholes across neighborhoods in the city in proportion to their true occurrence. The repair service dispatching decision process thus could only produce biased results without some counteractive measure to replace a more representative sampling of data or tweak the algorithm such that it could correct for the direct harms that came to the neighborhoods that were less well-off.[11]

Related to the concerns around bias emerging from the training data is the capacity for learning models to suggest discriminatory decisions based on such biased datasets. Training data might only contain information at a level of granularity that disadvantages certain groups. Such issues around the granularity of the datasets in question lead to such potentially discriminatory practices as redlining, in which certain inferences are drawn about individual neighborhoods—inferences that are extended to advise decisions made about any residents living in those neighborhoods. If the data suggests that on average a certain zip code earns relatively little the inference could be that it will therefore yield low click-through rates on ads and eventual purchases of interesting market opportunities—and thus anyone living in that neighborhood could be subject to a discriminatory outcome that may constitute a harmful disparate impact upheld by the courts should the harm occur in regard to, for instance, a housing opportunity.

Algorithmic Design

Machine learning algorithms carry the bias contained in data inputs and reflect those biases as the model learns based on the makeup of the training data. But critically, there are additional concerns that can result from the mechanics of traditional statistical analysis as well.

Foremost is the common fallacy in statistical analysis that correlation necessarily implies causation. We know this not to be true; it might be the case, for example, that certain racial groups have higher education levels than others, but this does not suggest that certain races are more intelligent or hard-working than others. Though this issue has been surfaced with machine learning models, there are mechanisms to curtail its prevalence proactively, in much the same way that certain explanatory variables are excluded from regression models because they are redundant or misleading.

Perhaps more deeply concerning, a poorly designed machine learning model—or one that is ill-equipped to fully handle the problems of discrimination, especially in fields that are not subject to strict regulations like personal finance or housing—may drift over time in such a way that perpetuates biased outcomes for marginalized people. This problem is distinct from the initial training of a model; indeed, trained models implemented in the consumer internet industry are refined on an ongoing basis so that they reflect the user's desires to the greatest degree possible. But what happens when an algorithm exceeds its intended purview and presumes things about us as individuals or as a population that just are not true—or even worse, encourages engagement of our less virtuous tendencies? There is a widely known statistical concept that describes a related tendency: "confirmation bias," whereby the model—or its designer—finds what might be expected given cultural norms, instead of the reality. The broad propensity for machine learning to "drift" in such directions presents a veritable thicket of concerns regarding bias. For instance, a model might learn from original training data that has been carefully engineered and monitored by the data miner to limit occurrences of unfair discrimination—but at their heart, learning models are designed to cut corners, to efficiently make decisions and determinations about a population in a way that approximately understands the true nature of the real world and reflects that in its algorithmic design, and as such, they are designed to discriminate. This natural tendency for them to attempt to find ways to discriminate in whatever legal manner possible organically forces them to tend toward overstepping the boundaries that have been set for them through secondary backdoors, and this enforces within the model an economic logic that drives them to acquire new behaviors through novel discoveries about the real world. But what happens when those so-called "discoveries" that advise the decision-making algorithm are outsized or otherwise biased? This is the type of model drift—through ongoing observation of the real world—that can engender discriminatory behavior. It is this characteristic of machine learning that can cause models to systematically feed voter suppression content to underrepresented minorities or send nationalistic groups down hateful pathways on social media. A generic conclusion depicting how this might work is illustrated in Figure 3; while the learning model might treat representative cases across a sample population by developing a reasonably accurate decision system for the majority, it might not reflect the particular situation of the minority.

An additional concern that can subvert antidiscrimination efforts is the organic generation of so-called proxies as the model is trained from the input data. It may be that a machine learning model is designed to exclude the use of any protected class data in the course of statistical analysis so as to explicitly protect against discriminatory outcomes against those protected classes. Models might learn, however, that there are alternative "proxies" that are equivalently descriptive of the protected class categorization as the protected class data itself. For instance, an algorithm prevented from accessing race information pertaining to the population might determine that some combination of other class fields—such as location of residence and name—might be used in tandem to generate through the back door an understanding of the individuals' racial group category. Further, such inferences might be completely non-transparent to the model's engineers, since they typically occur silently premised on the data already provided as input to the model, and proxies are not proactively reported to the designers as they are generated by the learning algorithm in the course of maintaining and updating the classifier model.

And as a final note, there is a robust active conversation in regard to what should be considered "fair" in the first place. Should fair mean whatever is lawful—and correspondingly that everything outside the reach of the law is on-limits and therefore fair? That is essentially how the industry today operates—and it is the underlying free market economic design of the United States that in fact enables and encourages such capitalistic "innovations" as discriminatory decision-making executed by artificial intelligence so long as it does not constitute harmful disparate impact in the areas of industry protected by federal civil rights law. In this way, the vast majority of the consumer internet's industrial activity falls directly outside the purview of federal laws in the United States—unless of course the business activity concerns American civil rights laws as has been suggested by the American Civil Liberties Union about a narrow sliver of Facebook's advertising platform.[12]

***

These harmful effects are supercharged when it becomes the direct commercial interest of the party developing the learning model to develop a classifier that maximizes revenue. In such an environment, potential discriminatory outcomes are a mere afterthought.

The Radical Commercialization of Decision-Making

One could question whether or not the fact that the internet firms have overtaken and now define the western media ecosystem is in fact a negative thing; perhaps it is for the best in that it breaks the centralization of the creation of content. A truly social platform elevates not necessarily the content generated by actors in the mainstream media like mainstream newspapers but rather those issues and elements raised, reported and reposted by the common user, and particularly a mix of those posts that are (1) predicted to be interesting to the user in question and (2) which have received wide circulation. (Atop these factors are more including the explicitly expressed preferences of the user, who might for instance choose to see the News Feed in chronological order, obviating some of the concerns recently associated with social media platforms.) Thus the traditional central power of large media companies—that epitomized by say the Hearst Corporation among other examples of the twentieth century—is somewhat diminished by the nature of the internet and the internet platforms themselves as social media receives more attention from the younger generations of the population than traditional forms of print media that also offer access to the news. And in fact, most of the appeal of social media originates from its capacity to connect us with issues and ideas that matter in our individual lives—issues that would not appear on traditional media formats at all—more so than the more abstracted concerns of the mainstream media.

Where power has waned amongst the producers of news media, however, the power of the internet platforms has quietly emerged—albeit in very different form. While power for traditional media firms lies largely in defining and producing content for broad dissemination and consumption, internet firms in large part do not participate in content production. Google's value proposition is instead focused on offering the efficient and effective classification and searchability of content (including news); for Facebook, it is offering seamless connection and engagement across the user's friend graph; and for Twitter, it is the attribution of ideas and engagement against them by the broader user population.

But it is not only provision of these services in broad terms that distinguishes and strategically separates Google and Facebook from their competition—if that were the case then there would be far more effective competition against these firms. A key part of their ongoing commercial strength in fact lies in their first-mover advantage[13] in seizing the reins of the consumer internet business model premised on the creation of advertising exchanges at a time when we also saw the coinciding rise of capacity in two technologies: data storage and computing. Just as Google and Facebook settled on their advertising-based business models these two technologies surpassed a key threshold that triggered the rise of the "big data" economy.

It is the combination of these technologies—the novelty of the targeted advertising regime created and commercially promoted across the media ecosystem alongside the coinciding rise of big data capacity—that, along with their nominally unique consumer services,

set them on their historic trajectory. What has emerged, though, is a commercial regime underlying the entire consumer internet that is algorithmically trained for the maximization of monetary opportunity subject to few constraints.

It is throughout the three pillars of the aforementioned business model that describes the consumer internet's practices that advanced machine learning systems are implemented for gains in profit—and equivalently, it is throughout each of these core practices that there is tremendous capacity for discriminatory results pushed onto the individual consumer. On a continuous basis, algorithms are trained to understand the consumer's preferences, beliefs and interests all of which are shuffled into the individual behavioral profile; keep the user engaged on the platform by understanding and ranking all content existing in the realm of posts that could be populated in the user's News Feed; and push digital advertisements at the user with which he or she will be likely to engage. In a sense, then learning algorithms are continuously and ubiquitously used by the firms leading the internet industry to infer as best as possible what the individual's true nature is and what arrangement of content should be pushed at the individual to maximize profits for the service operator.

I describe this as the "commercialization of decisions"—and it is radical because of its continual engagement and refinement, and its total ubiquity across the sector. All decisions made by learning algorithms in the context of the consumer internet are now necessarily commercialized in light of the combined strengths of supercharged big data technologies and platform power. That is to say that each decision made by a consumer internet learning algorithm—be it over determination of what content to push at the user or inference of the user's character, or some other narrower practice—is incentivized by the pursuit for profits; there is currency tied to every decision-making process that occurs in the industry no matter how impactful or important it is. This is a critical distinction from prior times: the commercialization of decision-making has inseminated novel opportunities to disseminate any sort of speech—whether organic, commercial, or otherwise nefarious in nature—and inject it throughout the modern media ecosystem.

We have thusly moved on from the formative "public good" conceptualization at the inception of the world wide web; we are in the age now of the "commercial good"—explicitly, of the firms leading the industry. The media ecosystem of the twentieth century, in contrast, did not involve the commercialization of fine-grained information dissemination. This was perhaps true even in the early stages of the internet through the turn of the millennium. But now algorithmic developments including the deployment of sophisticated learning models by the most cash-rich firms in the world—alongside their data-gathering practices and advantageous pseudo-monopolistic positions in a market with a paucity of true or would-be competitors—have collectively introduced a vicious situation by which commercial operators have the opportunity to initiate, advertise, and host a market for commercialized information dissemination in such a way that it is those willing to pay-to-play in this commercial regime who exclusively have the capacity to push information at the individual.

That is not to say that this power of decision-making was to an extent true of past instantiations of the American media ecosystem as well. The prior world dominated by broadcasting, radio, and print materials too had the capacity to produce and perpetuate bias. But there were some key differences. Their reach was not as granular or personalized because of the nature of the technology in question; a consumer internet laden with learning algorithms evolving and operating over corporate servers and producing results within milliseconds on the Search results page generates different impacts entirely. Furthermore, these more traditional past instances of the media ecosystem were heavily regulated either directly by the government or indirectly through combination of measures instituting industry-wide transparency and public accountability.[14] Examples include federal election regulations for the broadcast and radio formats as well as journalistic standards across the news media. Thus, overall their capacity to engage in unfair practices leading to potential consumer harms was constantly policed. While they did nevertheless have tremendous power –these formats collectively constituted the media ecosystem—they experienced continual pressure and possessed limited capacity to perpetuate damaging impacts on the public.

Individual capacity to determine what we will see and be subject to has been holistically undermined and diminished by the consumer internet firms. Whereas the individual's consumption of information in decades past was one of open space or human thought it has now been invaded by a silent form of commercial speech in that the content displayed before us at the call of the firm responsible for populating the results page. Each time we open the laptop or checks the phone and utilizes the services central to information consumption today we are subjected to an array of information preselected and ordered for us at the determination of a mercenary machine that works for the profit of Facebook or Google, with nothing else trained into its decision modeling besides profit maximization. Scholars contend that human minds were not meant to deal with this kind of ease: instead we are biologically trained to see a wide unlabeled array of content and contend with its merits and demerits to the end of deciding for ourselves whether we shall support and take up the opinion-driven arguments or objective information contained therein.[15]

This is Thoreau's civil disobedience; but the Twitter feed has subverted the very concept of civil disobedience and subjugated the human interest to such an extent that it is not only our democratic processes and progress but our moral humanity itself that is currently under direct and immediate threat perpetrated by the consumer internet's business model. It is that which is in the crosshairs of the modern commercialized information dissemination system in America.

To examine this conundrum from a different angle, it is the 'third layer' of the infrastructure of the media system that has now been radically industrialized. The other two—the first being the physical network infrastructure and the second the content—already were in decades past. The third is the content dissemination network—but it could be said that the third layer of the infrastructure never should have been a free market in the way it is now in the first place. Leaving aside whether and how much the first two layers should have been opened to the industry at all and inspecting only the third, we can note that the industrialization of the dissemination layer clearly subverts the consumer's interest if left to the free market, given the observable negative externalities including the perpetuation of the disinformation problem and the wide spread of hate speech over these platforms.

Nissenbaum argues the approach to consumer privacy protection undertaken by the Federal Trade Commission and Department of Commerce is dangerous, noting that the U.S. government's "interest has been limited…by a focus on protecting privacy online as, predominantly, a matter of protecting consumers online and protecting commercial information: that is, protecting personal information in commercial online transactions. Neither agency has explicitly acknowledged the vast landscape of activity lying outside the commercial domain."[16] Nissenbaum's reference is to the manner in which U.S. governmental agencies focus not on privacy concerns at large as and when they occur across society including governmental agencies and regulated entities like hospitals and banks, but rather only on those occasions when the data transfer affects "consumers"—those individuals party to some monetary transaction in the marketplace. Based on the discussion above we can extend Nissenbaum's point to the lack of effective oversight over the commercialization of decision-making—precisely because the narrow and independently minor decisions made using the classifier models developed by learning algorithms do not necessarily have dollars attached to them. But they are nonetheless designed in such a way as to yield the greatest possible profit margin for the service operator—and even perpetuate provably discriminatory decisions against individuals and classes of individuals so long as doing so remains non-transparent to the public and is aligned with the profit motives of the platform firm.

To that end, the collection of personal information is ubiquitous and its transfer amongst firms involved in the digital media ecosystem multidirectional. Indeed, the *modus operandi* of leading internet firms is to at once be at the center of and reach its tentacles throughout the commercial information sharing network stretching across the digital ecosystem. Firms like Google accordingly utilize a multitude of technologies and technological protocols to collect personal data, including over its own platforms, as well as through web cookies and physical equipment technologies deployed throughout the world. Critically this information is maintained by the firm and others like it within the company's walled gardens—its proprietary systems so that Google can maintain hegemony over the knowledge of the customer's individual profile for content-targeting purposes. Further, the firm "leases" the information out in anonymized formats—enabling advertisers to target certain classes of the population at will. Sometimes, the advertiser might inject its own data into Google's advertising platform, encouraging Google to help it reach audience segments to a remarkable degree of precision. This bidirectional relationship is critical to the functionality of the consumer internet—and operates as the grease at the joints of an industry enabling the aforementioned radical commercialization.

Bias in the Consumer Internet

The commercialization of decision-making in the consumer internet plays out in various ways potentially detrimental to marginalized groups including protected classes of the population. When the markets elevate currency over values the resulting economic logic tends toward enabling the pursuit of highest profit margin at the expense of any other concern, particularly if it is an unpoliced one extant over a largely unregulated market. Machine learning is the tool that enables the collation and exploitation of information, thus reducing transaction costs even further—with the profits generated thereof typically being drawn up by the industrial entities responsible for implementing the learning models in integrated manner.

Indeed the internet is effective as a means for communication—to the extent it is now humanity's social medium of choice—because it reduces costs of transaction in the exchange of information relative to the communication media of the past which typically did not enable personalization of rendered services nor collection of information on the consumer in the first place; the internet thus enables a two-sided exchange in a manner we had no capacity for in years past.[17]

But it is precisely this reduction of transaction costs that has enabled discriminatory outcomes that disfavor marginalized communities, particularly in the United States where the internet is in such wide use, the internet industry has such tremendous political power, and our

demographic heterogeneity and national political economic tradition and trajectory are such that the capacity for internet-enabled discrimination has been supercharged.

In this part we discuss a non-exhaustive set of common practices in and features of the internet industry that illustrate its capacity for discrimination though they have nevertheless reduced transaction costs for individual consumers.

Targeted Advertising Platforms

The creation of the commercial regime underpinning the consumer internet economy—targeted advertising—has enabled both intentional and unintentional discriminatory outcomes. Typically, ad targeting regimes take advantage of the commercial interests of two types of parties: the advertisers that wish to communicate their products and services to consumers and persuade purchasing decisions as possible; and the platforms and publishers that have access to consumer attention and therefore own ad space.[18]

Usually, platforms also possess and analyze large, refined stores of information on consumers. The raw data might include data collected about the consumer's "on-platform" activity including what products, social posts and search results displayed on the platform in question the consumer interacts with; the consumer's "off-platform" web activity pertaining to activity on third-party websites, including mouse clicks, browsing pathways, and content consumed; location information shared with the platform via the consumer's smartphone should the consumer have opted into location sharing with the platform service (or through other means in certain cases[19]); location and behavioral data collected through other device technologies such as beacons and routers that interact with the consumer's devices in the physical world; data purchased from or voluntarily shared by third parties such as data brokers and advertisers; and many others.

Advertising platforms—including those implemented and hosted by Facebook, Google, and Twitter—take advantage of such data collection regimes to infer behavioral advertising profiles on each user participating on the company's internet-based services. Those behavioral profiles are maintained by the platform firms and largely remain non-transparent to third parties. But should advertisers such as apparel designers and retail banks wish to target certain audience segments—young people of a certain income in Manhattan and San Francisco, for example—the platform firm typically analyzes its data stores, and determines which grouping of consumers in the target geography would be most likely to purchase the advertiser's wares. It is this determination of who should go into the targeted audience segment that clearly has the capacity to engender harmful disparate impact. A recent suit put forth by the U.S. Department of Housing and Urban Development illustrates this tension clearly: the used its authority under the Fair Housing Act to allege that Facebook enables harmful disparate impact in making available housing opportunities because of the way that advertisers can target certain groups according to their membership in various consumer classes—including protected classes such as race and gender.[20]

Perhaps most dangerously, civil rights laws in the United States only cover certain key areas that are absolutely critical to maintaining a modicum of economic fairness—including in housing and employment. Unfortunately, such protections against a commercial operator enabling disparate impact in the majority of other areas does not necessarily trigger a civil rights violation despite the clear discriminatory outcomes that can arise from only certain marginalized communities being subject to shady scams or, conversely, more mainstream communities exclusively being pushed very favorable ads such as investment opportunities.

Meaningful Social Interaction

Consumer internet firms deal in a novel form of currency: the collective combination of the user population's personal information and attention. By raking as much of this as possible and amassing it to generate collated ad space that can be sold off to the highest bidder via intelligent auctions for the purpose of enabling targeted commercial speech, the internet companies maximize their value proposition to businesses that wish to advertise back at the consumer. It is a vicious cycle that takes advantage of other efficiencies as well—in particular, the need to continually engage users such that they spend as much time on the platform as possible and furthermore engage with it to the greatest possible extent.

In 2018, Facebook chief executive Mark Zuckerberg proclaimed that his company would institute new changes to the algorithm driving the social media network's core News Feed service that ranks the universe of content available to a given user in the home screen; he noted that the company would now focus on promoting "meaningful social interactions."[21] That is not to say that this was not always in the company's designs: he discussed how recent events had illustrated more clearly that there was too much passive interaction with content, particularly posts shared by "businesses, brands, and media."

What does meaningful social interaction really entail? Conveniently for Facebook, it is a metric that if effectively maximized can contribute to the two resources it principally cares about: the consumer's attention and personal information. Effective meaningful social interaction would keep users on the platform because

if done right it would connect users to more personal social content that they actually want to see—and if they engage more with such content then Facebook will know it and thereby know the user better such that ads can be disseminated more efficiently at them and ad space can be increased.

This is where the power of commercial machine learning—and resulting machine bias—come in. There is no scientific way to determine what types of content matter for an individual user; it is nigh impossible for a machine to infer precisely what the individual consumer truly cares about. Only broad inferences can be drawn—but it might be more difficult to infer what academic subjects and scholars resonate for an individual, or which particular players on a team he or she likes, or which shade of blue he or she likes the most. This is the fallacy of data, and by extension, learning models; it is used to estimate the real feature but cannot ever offer a precise representation of the real world, and yet it is readily used to make determinations about what the individual actually cares about in the real world. Thus, the leading consumer internet companies' quest toward enabling meaningful interaction—whether in the context of a search engine or e-commerce platform or social media network—is flawed at best.

The industry's use of highly sophisticated artificial intelligence systems including neural networks for real time analysis of user behaviors—in conjunction with social science research conducted within the industry itself—powers the refinement of the models used to rank such features as the News Feed. But regrettably, such systems have the propensity to supercharge the deployment of assessments about the individual in ways that implicate the individual's interest. If a user does not interact with some mundane piece of content because it does not personally resonate at a social or intellectual level, the platform must reorient its assessments about that individual user. It is this dynamic that has led Facebook down the path of grouping individuals by political allegiances and which has caused YouTube to be unable to screen certain inappropriate videos.

When looking through the lens of discriminatory practices, the platforms are designed to necessarily make assumptions about the nature of the individual based on the individual's demographic profile—including protected classes such as race and gender but also more precise ones including interests in certain forms of ethnic culture, music, and other instances of intellectual content. This is an online commercial landscape in which disparate impacts can run riot—where only certain marginalized classes are shown (or not shown) certain forms of content. And even if the content does not trigger civil rights protection in the United States, there are other manners in which it might damage the economic prospects of the individual. If Facebook decides that an individual is likely more interested in basketball than microeconomics, for example, it might be the case that that individual is never subjected to content that would encourage better practices around personal finance, better awareness of the political state of the nation, and better awareness of broader economic opportunities that might be available should the user know where to look for them.

Whether such ranking models are fair or not all depends in the end on the design of the algorithm that maximizes so-called meaningful social interaction, defined and algorithmically trained to service the commercial objectives of the platform operator.

The Initial Pursuit of High-Value Customer Audiences

There is a tradition in Silicon Valley, particularly in the consumer internet industry, whereby fledgling firms tend to serve those niches that are already well-off first; should they be able to prove the efficacy of the business by serving those high-value customer segments then they might receive investment funds to tackle broader growth as well. Indeed, companies leading the sector have variously been party to such practices: Facebook first invited only Harvard students to participate on the network[22]; Airbnb initially served only those cities where real time hotel prices were high[23]; and Gmail's beta version was distributed first to a few hundred opinion leaders and those friends they wished to invite to use the service as well.[24]

Needless to say, such communities—namely, elite universities, would-be hotel patrons in rich cities, and public intellectuals—are not representative of any community beyond the elite and tend to deprioritize or exclude marginalized communities that are most often subject to harmful discrimination. Nonetheless, it is through the observation of these initial groups' interactions with the platforms that computer engineers attempt to design the form their platforms will take at steady state. This culture of serving the privileged first and rolling out consumer products to the rest of society should the product gain in popularity is seemingly part and parcel of the investment culture that bleeds through the venture capital industry.

But it is a culture that considers the desires of lower socioeconomic classes last. And when overlaying the development and refinement of learning models over this conundrum, the potential for machine bias leading to disparate impact becomes resoundingly clear. The argument could be made that the platform firms protect against this potential harm in various ways—for instance by protecting against in-built propensities for learning models to perpetuate biased outcomes—but the fact remains that the design of the platforms necessarily must favor the elite and wealthy first. In a capitalistic regime favoring free markets no other approach would be viable

for venture capitalists and founders; if they do not take advantage of the economic opportunity of serving the well-off first then the competition will eventually do so and overtake them. In fact it could be said that if at any point a platform such as Facebook were to lose high-value users to the competition then the company would have to either acquire those competitors or reorient the ways in which the fundamentally platform works so as to increase the probability that high-value users might come to the platform. Indeed, this is exactly the strategic circumstance Facebook finds itself in now as it considers how to reclaim the high customer lifetime values at hand with respect to the young users who opt for non-Facebook internet-based services.

Public Policy Interventions to Counter the Spread of Machine Bias

Experts contend that model designers can protect against bias through development of technologies that check the representative nature of the training data and fairness of the outcomes. But while one absolutely can engineer such technological solutions to counter the overreaches of learning models, what forces companies to be fair when it is in their commercial interests to discriminate, even unfairly so, as long as the discrimination is not illegal? I would thus suggest a slightly different remedy: implementation of such technologies by the industry backed up by accountability forced on the industry through smart and earnest governmental regulation.

Machine learning technologies have come to the fore because of their tremendous efficiency. No longer do we require humans to monitor traffic systems to infer areas of congestion and manage the network; Uber, Waze and Lyft can accomplish the task much more effectively on an algorithmic basis. No longer do we need news editors to determine what information should or should not go front and center before our individual attention; Facebook, Apple and Twitter can infer who we are, what we want to see, and route the relevant content to us. No longer do we need to ask the contracting expert what flooring suits our apartment the best; Amazon will find out for us and assure it arrives post haste. And no longer do we need to rely on the guidance counselor to help decide where to apply and what college to ultimately attend; Google can address all our concerns.

As machine learning algorithms and artificial intelligence system become more ingrained in our daily lives and influence our behaviors throughout the day, so too does humanity necessarily become increasingly dialogical with the machine underpinning the consumer internet. Society's observable actions and behaviors are actively feeding the decisions executed by the machine that sits quietly behind the internet, and the corresponding commercially-driven decisions in turn influence our actions and outlooks in the real world. Beyond the obvious questions this conundrum presents in regard to individual autonomy, psychological dependence, mental health, and the broader concern of empowering a civilization-wide overdependence on machine technologies and implicit bias against sentient real-world expertise, however, is the apparent reality: machines are discriminatory by design. Indeed, the more discriminatory they can be—the more incisive their predictions about individual behaviors and the collective outlooks of population classes—the more they add to the industry's pocketbook. This is ad targeting and content curation 101: if a machine can understand your mind, it is doing the job Facebook designed it to do. But in the course of so doing, the machine is bound to make frequent mistakes; there is no real-time learning system that can effectively model the human psychology without making mistakes along the way—and it is that in noise that pervades the system where harmful bias lurks.

There is no reasonable solution, then, but to utilize the full agency of the public interest to intervene and clarify for commercial entities what is right and what is wrong. Without making such rules of the road explicit, it is in the industry's interest to breach the public interest so long as it is legal to do so and unintelligible to the public; if Instagram leaves such opportunities on the table, Snapchat will pounce—and vice versa. They both thus have to take such opportunities up unless the consumer market reproaches them through expression of collective sentiment in the marketplace or unless the government intervenes. And consumer outrage expressed through purchasing behaviors will take too long or have minimal long-term impact in a space that offers little transparency. We need look no further than the voluntary reforms instituted by Facebook since the Cambridge revelations; while it has ceased certain activities, firms not under the public eye have taken them up, taking advantage commercial zones of operation cast aside by Facebook on the back of vociferous public advocacy.

We can conversely inspect the industry's actions in the face of governmental inquiries. If the industry earnestly wished to protect against these harms, why would it not wish to submit to governmental review and sectoral oversight? It is a problem of the interests of private commerce versus the interests of human rights. The culture engendered by the Facebook cultural insignia "move fast and break things" necessarily implicates machine bias and other challenges wrought by the radical capitalism seen in this industry. The industry's tendency is accordingly not to take on challenges presented by algorithmic design earnestly until it becomes popular to do so—by which time it is too late; the consumer internet industry's systems may have by that time contributed inordinately to systemic bias, prominent as it is in the American media and information universe.

A novel approach for governmental intervention should include the following, offered in increasing order of political difficulty given the inevitable policy pushbacks each measure would face.

- *Federally funded research into techniques to protect against algorithmic bias.* Computer scientists have developed novel techniques to protect against machine bias in recent years.[25] But as a general matter these approaches are variously applicable only to certain types of models or are otherwise not always feasible because of cultural norms that dictate companies will fail to pause to question whether their models are fair before deployment, or because of other practical hang-ups in the sector. More robust research is needed to develop more industry-grade mechanisms to help protect against machine bias. Further research is also needed to develop greater understanding regarding the impact of computing machines on society, and what public policy measures should be taken to counter industrial overreaches and contain harms. A good start has come from the Defense Advanced Research Projects Agency's Explainable Artificial Intelligence program[26]; the government should channel further resources to such pursuits.

- *Federally endorsed multi-stakeholder standards development toward a guiding framework for ethical artificial intelligence.* Mathematicians and computer scientists have variously come together with ethicists and philosophers from across the industry, civil society, and academia to produce a slew of ethical codes for artificial intelligence and machine learning in recent years. The fact these discussions exist is a positive development. But we must assure history does not regard them as fluff. One commonly cited framework[27] highlights five key principles: "responsibility" such that those with grievances in regard to an algorithmic outcome have redress with a designated party; "explainability" so that the algorithms and data used to develop them can easily be explained to the public or those subject to their decision-making; "accuracy" so that the model's errors can be identified and proactively addressed; "auditability" so that third parties including public interest agents can investigate the algorithms and assure their integrity; and "fairness" so that the models do not perpetuate biased outcomes. This represents a start to developing a comprehensive set of principles on the governance and execution of fair machine learning models. The government should work with these and more stakeholders to coordinate a multi-stakeholder conversation concerning the development of an ethical framework for artificial intelligence. These conversations should be focused on the particular issue of the technological nature of the algorithms and data inputs themselves—leaving other important but less relevant contemporary conversations regarding the technology industry and the governance of artificial intelligence to the side. The government can use the National Institute of Standard and Technology's Cybersecurity Framework developed under the auspices of the Obama administration as a blueprint for how such multi-stakeholder guidance can come together.[28] Of particular importance throughout the process will be the assurance that public interest advocates are represented.[29] Such conversations can focus on the industrial use of artificial intelligence over the internet to maintain group focus while also addressing internet algorithms' outsize influence over the information ecosystem.

- *Industrial auditing and oversight of high-impact commercial internet algorithms backed by governmental enforcement to assure fairness.* Consumer internet firms extensively implement machine learning models to drive growth, engagement, behavior profiling, and revenue collection and management—among many other activities. These have a tremendous impact on public interests from fairness to democratic process and should be subject to general governmental oversight in some capacity. A model like that settled by the Federal Trade Commission with Facebook[30] and Google[31] through consent orders nearly ten years ago may be appropriate, whereby in response to industry overreaches the agency settled new conditions with each company, including the ongoing auditing of their practices with regard to maintaining consumer privacy. This condition from the consent orders effectively enforced a sea change on the companies; it forced them to install what are now known throughout the industry as privacy program teams—staff that are charged by the company to work with product managers and engineers to understand every single proposed product innovation, including the most minor of features, and help the subject firm coordinate a cross-functional decision as to whether the proposed changes would harmfully implicate the user's privacy or not. The personnel in the privacy program teams interact with external professional auditing consultants who

verify the integrity of the privacy practices of the subject firm, and develop periodic reports shared with the federal regulators that help affirm that the subject firm's privacy program is effectively protecting users from privacy overreaches. It could be argued that, in light of the reality that PricewaterhouseCoopers failed to find Facebook's missteps that were eventually revealed by Christopher Wylie, these types of setups are bound to fail.[32] They can suffer, for example, from the traditional auditor's paradox, by which the auditing firm becomes close and collegial with the subject firm and fails in its role as an independent review agency working in earnest for the public interest. Culturally there can be a lack of incentive to report concerns accurately, largely because sharp criticisms will be seen by the subject firm. This is where the government can come in by holding all parties accountable. As the U.S. government pursues actions against the internet industry on the basis of further breaches of privacy, security, public trust, and algorithmic integrity, it should consider mechanisms to additionally force the companies to work with independent external auditors to assure internet-based artificial intelligence systems are not implicating public interests.

- *Radical data and algorithmic transparency for the public.* Centrally responsible for the exceedances of the algorithms underlying the consumer internet architecture is the lack of transparency into how they are developed, how the operate, and what they accomplish. Consumer transparency into this regime—through consumer understanding of what data corporate actors hold on them, how behavioral profiles are developed through inference, how machine learning models are used to develop such features as the News Feed and YouTube recommendation algorithms, and what the practical outcomes of these algorithms are—is critical to limiting the harmful discriminatory effects of the internet platforms. Indeed, many have attempted to develop tools to layer such transparency over the sector—including the political ad transparency projects led by ProPublica, Mozilla, and Who Targets Me, which were all stopped by Facebook in early 2019.[33] That Facebook was so determined to block the aforementioned services by tweaking its code is illustrative of the tension at the heart of true transparency measures: transparency breaks the impetus of the business model of internet companies like Facebook. These companies want to protect information pertaining to how their curation and targeting algorithms work for two primary reasons. First, helping the public peer into the targeting metrics pertaining to Facebook ad campaigns can shine a much-needed light on how this company's algorithms perpetuate bias including by feeding the filter bubble problem, stoking hateful conduct online and offline, and enforcing damaging disparate treatment and impact in areas including politics and media exposure. Second, exposing the design behind algorithms enables the company's competitors to understand important strategic elements of the commercial makeup of Facebook and adjust their strategies in real time to challenge the company's strength in the market. In other words, it is all a form of commercial protectionism. I suggest a novel regime—a radical form of transparency as I have discussed with colleagues in related work—that can truly hold the industry accountable for the negative effects pushed by its models onto the public. Such transparency would enable users—or at the least, governmental or nonprofit organizations working in the public interest—to see what inputs go into the development of algorithms developed in the internet industry, and what outcomes those algorithms produce.

In addition to these proposals, reforms concerning privacy and competition policy are much-needed and should be pursued as well. I discuss with colleagues what form such reforms could take in related work.[34]

Conclusions: An Ethical Approach in the Way Forward

The tide of public sentiment is closing on Silicon Valley internet firms. Over the past year, the Cambridge Analytica revelations, frequent disclosures about privacy and security breaches, and historic regulatory fines have demonized the sector and turned our attention toward Facebook, Google, and Amazon.

What distinguishes the consumer internet sector is that it is not subject to a rigorous regulatory regime like the telecommunications, healthcare, or financial industries are; the operation of online digital services over a physical infrastructure is still largely a novel practice as far as the laws are concerned, and the U.S. Congress has not yet acted. In this mostly regulation-less environment, these firms have had the opportunity to grow profits toward the combination of business practices that most effectively yields highest margins—in just the way that Karl Marx suggests capitalists would. These companies have, in the view of many scholars, subjugated the national public interest. The disinformation problem; the spread of hate speech; the persistence of extremist content; and the present concern of algorithmically-charged outcomes that perpetuate harmful bias: these negative externalities

are the symptoms of the commercial regime that sits behind the consumer internet, a silent machine that is designed algorithmically only to seek the highest possible profit without consideration of the public interest.[35]

Centrally concerning is the currency these firms deal in and the opaque mechanisms by which they rake it. Some industry executives suggest the services they offer are "free"—a misleading conjecture. True, consumers do not pay monetary fees for their services, but the most effective consumer internet firms develop as two-sided platforms that amalgamate a complex combination of user attention and data on the end-consumer side of the market, and translate it through an automated digital advertising exchange into monetary reward in the advertising market. Further, these firms have inordinate market power in the end-consumer market; Facebook for example has near-monopolies in traditional social media and internet-based text messaging, Amazon has a near-monopoly in e-commerce, and Google has near-monopolies in online video, email, and search. Thus, these firms are able to hoover currency in the form of attention and personal data on one side of the market and charge monopoly rents for it on the other side of the market.

This hegemony over the market has been shown to trod over the public interest. The industry's disincentive in protecting the public from such negative externalities as the disinformation problem is a mere symptom of its unwillingness to bend this highly profitable business model along with its use of social power to protect the business model from regulation through influence over policymakers. What the public now needs is a novel regulatory regime that can effectively rebalance the distribution of power between the industry, government, and citizen—a digital social contract. Jean-Jacques Rousseau suggested the danger of radical property rights—such as those that the capitalistic Silicon Valley now has over the individual's attention and personal data—when he noted we must "beware of listening to [the first man to claim property rights]. You are lost if you forget that the fruits of the earth belong to us all, and the earth itself to no one."[36]

I do not contend that we should abolish the industry's ownership of intellectual property in the consumer internet industry altogether—nor that we compromise the targeted advertising business model entirely. I would rather suggest development of a regulatory response that effectively responds to the capitalistic overreaches of the business model that sits behind the consumer internet. And this must include measures that can effectively hold the industry's artificial intelligence platforms accountable, including through transparency that would enable public visibility into the darker effects of learning models implemented by the industry that systematically make decisions that are not in the interests of the individual.

Free market capitalism is the principal hallmark of the American approach to national economic design, but the government has never hesitated to strike down the market when its practices have implicated the nation's commitment to democracy. This is the very situation we now find ourselves in with respect to the internet.

---

[1] Robin Mansell, New visions, old practices: Policy and regulation in the Internet era, *Journal of Media & Cultural Studies*, Volume 25, Issue 1, 2011.

[2] Cecilia Kang and Sheera Frenkel, Facebook Says Cambridge Analytica Harvested Data of Up to 87 Million Users, *The New York Times*, April 4, 2018.

[3] S.1989, Honest Ads Act, 115th Congress, 2017-2018.

[4] See e.g. Solon Barocas and Andrew Selbst, Big Data's Disparate Impact, California Law Review, Vol. 104, No. 3, June 2016.

[5] Yoshua Bengio and Yann LeCun, Scaling Learning Algorithms towards AI, in *Large-Scale Kernel Machines*, MIT Press, 2007.

[6] Amanda Sakuma, The bogus "Momo challenge" internet hoax, explained, *Vox*, Mar 3, 2019.

[7] Ian Sample, Study blames YouTube for rise in number of Flat Earthers, *The Guardian*, Feb 17, 2019.

[8] Steven Willborn, The Disparate Impact Model of Discrimination: Theory and Limits, *American University Law Review,* Vol. 34, No. 3, Spring 1985.

[9] Susan Grover, The Business Necessity Defense in Disparate Impact Discrimination Cases, *Georgia Law Review,* Vol. 30, No. 2, Winter 1996.

[10] Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights, *Executive Office of the President*, May 2016.

[11] When Digital Dust Is Gathered, Constellation May Be Muddled, *National Public Radio*, April 13, 2013.

[12] Facebook agrees to sweeping reforms to curb discriminatory ad targeting practices, *American Civil Liberties Union*, Mar 19, 2019.

[13] Thomas Eisenmann, Internet companies' growth strategies: determinants of investment intensity and long-term performance, *Strategic Management Journal*, Oct 30, 2006.

[14] See e.g. Advertising and disclaimers, Federal Election Commission.

[15] Cass Sunstein, #Republic, Divided Democracy in the Age of Social Media, *Princeton University Press*, 2017.

[16] Helen Nissenbaum, A Contextual Approach to Privacy Online, *Daedalus*, Fall 2011.

[17] B. Mahadevan, Business Models for Internet-Based E-Commerce: An Anatomy, California Management Review, July 1, 2000.

[18] Robbin Lee Zeff and Bradley Aronson, Book Advertising on the Internet, *John Wiley & Sons*, 1999.

[19] Joseph Cox, I Gave a Bounty Hunter $300. Then He Located Our Phone, *Vice*, Jan 8 2019.

[20] The Secretary, Department of Housing and Urban Development, on behalf of Complainant Assistant Secretary for Fair Housing and Equal Opportunity v. Facebook, Charge of Discrimination, U.S. Department of Housing and Urban Development.

[21] Adam Mosseri, Bringing People Closer Together, *Facebook Newsroom*, Jan 11, 2018.

[22] Alexis Madrigal, Before It Conquered the World, Facebook Conquered Harvard, *The Atlantic*, Feb 4, 2019.

[23] Jessica Salter, Airbnb: The story behind the $1.3bn room-letting website, *The Telegraph*, Sep 7, 2012.

[24] Chris Welch, Gmail is 10 years old today, *The Verge*, Apr 1, 2014.

[25] See e.g. Nicholas Diakopoulos, Accountability in Algorithmic Decision Making, *Communications of the ACM*, Feb 2016.

[26] David Gunning, Explainable Artificial Intelligence (XAI), Defense Advanced Research Projects Agency.

[27] Nicholas Diakopoulos et. al., Principles for Accountable Algorithms and a Social Impact Statement for Algorithms, *Fairness, Accountability, Transparency in Machine Learning*.

[28] Cybersecurity Framework, National Institute of Standards and Technology, U.S. Department of Commerce.

[29] Yochai Benkler, Don't let industry write the rules for AI, *Nature* 569, 161, 2019.

[30] Agreement Containing Consent Order, File No. 092 3184, *In the Matter of Facebook, Inc.*, Federal Trade Commission.

[31] Agreement Containing Consent Order, File No. 122 3237, *In the Matter of Google, Inc.*, Federal Trade Commission.

[32] PwC had cleared Facebook's privacy practices in leak period, *Reuters*, Apr 20, 2018.

[33] Jeremy Merrill and Ariana Tobin, Facebook Moves to Block Ad Transparency Tools — Including Ours, *ProPublica*, Jan 28, 2019.

[34] See e.g. Dipayan Ghosh and Ben Scott, Digital Deceit II: A Policy Agenda to Fight Disinformation on the Internet, *New America & Shorenstein Center on Media, Politics and Public Policy at the Harvard Kennedy School*, Sep 24, 2018; and Dipayan Ghosh, A New Digital Social Contract to Encourage Internet Competition, *Antitrust Chronicle, Competition Policy International*, Apr 2019.

[35] Joseph Nye, Protecting Democracy in an Era of Cyber Information War, *The Hoover Institution*, Nov 13, 2018.

[36] J. Rousseau, On the Origin of the Inequality of Mankind: The Second Part

Dipayan Ghosh is a Shorenstein Fellow and co-director of the Platform Accountability Project at the Harvard Kennedy School. He has served as an economic policy advisor in the White House during the Obama administration and a privacy and public policy advisor at Facebook.

Supporting Data

Figure 1. Consumer internet platforms engage in exchanges of personal information, content including personal posts and news, and dialogical feedback.

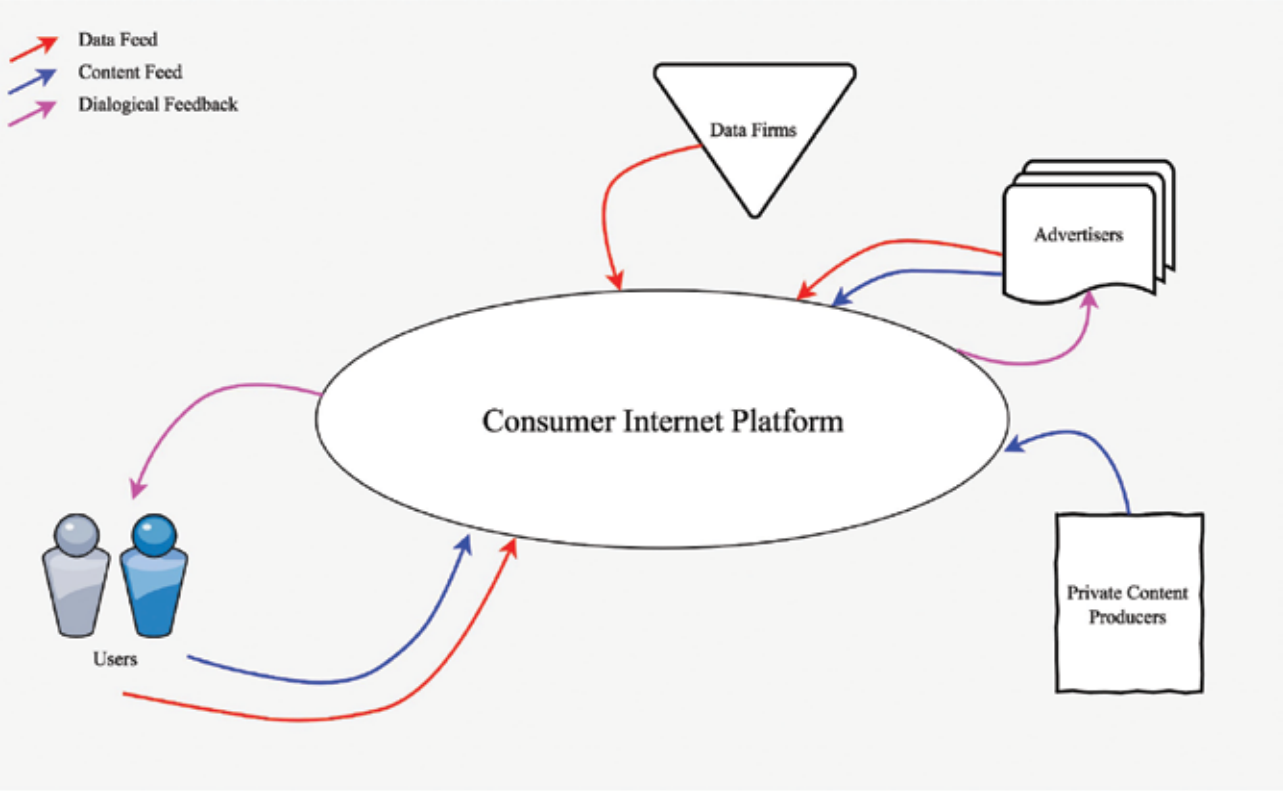

Figure 2. Commercial learning models designed to infer behavioral profiles and curate content are continually refined by consumer internet firms through feedback from real world routines and behaviors.

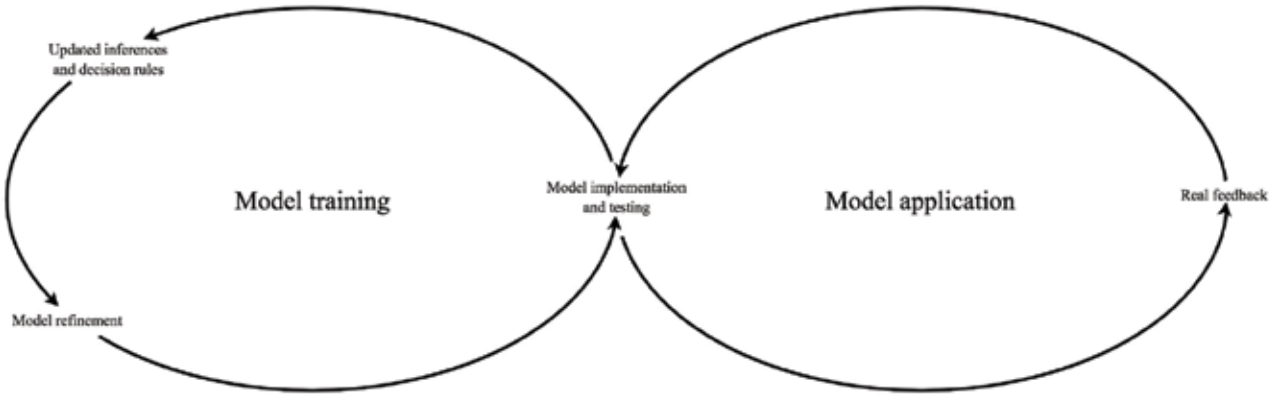

Figure 3. While learning models might efficiently design decision regimes for large populations, poorly designed systems may fail to detect that minority populations defined along protected class lines have a different nature, which can systemically perpetuate harmful discrimination.