Barriers for Academic Data Science Research in the New Realm of Behavior Modification by Digital Platforms

Travis Greene[1], **David Martens**[2], **and Galit Shmueli**[1,*]

[1]National Tsing Hua University, Institute of Service Science, Hsinchu, 30013, Taiwan
[2]University of Antwerp, Department of Engineering Management, Antwerp, 2000, Belgium
[*]corresponding author(s): Galit Shmueli (galit.shmueli@iss.nthu.edu.tw)

ABSTRACT

The era of behavioral big data has created new avenues for data science research, with many new contributions stemming from academic researchers. Yet, data controlled by platforms has become increasingly difficult for academic researchers to access. Platforms now routinely use behavior modification techniques to manipulate users' behavior, leaving academic researchers further isolated in conducting important data science and computational social science research. This isolation results from researchers' lack of access to human behavioral data, and crucially, to both the data on machine behavior that triggers and learns from the human data and the platform's behavior modification mechanisms. Given the impact of behavior modification on individual and societal well-being, we discuss the consequences for scientific knowledge creation and the roles and responsibilities of academic data scientists. Finally, we consider how current and future regulatory actions may enable academic data scientists to conduct meaningful research on platform environments in the realm of behavior modification.

Platform-based Data Science Research

A vibrant and growing community of data science academics is involved in methodological and applied research related to behavioral big data (BBD). BBD refers to very large and rich multidimensional data sets on human and social behaviors, actions, and interactions, which have become available to companies, governments, and researchers[1]. The availability of BBD is the result of our extensive use of digital technologies in almost every facet of our daily lives. Such data are captured and collected by internet and social media platforms, mobile apps, internet-of-things (IoT) gadgets, and more. They are the "digital breadcrumbs" we leave behind as we go about our lives online.

We use the term *digital platform* (or simply *platform*) to refer to multi-sided markets aimed at the commodification of user activities, web, and app content that rely on network effects and reduced transaction costs[2,3]. Platforms are implemented as programmable digital architectures geared towards the systematic collection, algorithmic processing, circulation, and monetization of user data. One way in which this user data, mostly BBD, is leveraged by platforms is by predicting user characteristics and behavior and shaping user actions through personalized interventions[4]. For example, Facebook uses BBD to predict and modify a user's likelihood of clicking a push notification[5]. BBD has been successfully used for prediction in a wide range of applications: What you "like" on Facebook has been shown to be predictive not only of your IQ, political preference, and openness[6], but also your product interest and even financial default behavior[7].

Academic and industry researchers acquire and analyze BBD for purposes of extracting knowledge, making scientific discoveries, and for developing and evaluating new or existing data science methodologies. The combination of "behavioral" and "big" creates challenges and opportunities for both applied and methodology-developing data science researchers[8,9].

The present work examines how the increasing reliance on behavior modification techniques by platforms, especially reinforcement learning-based (RL) algorithms, impacts the practice of academic data science research. RL is a third paradigm of machine learning (ML)—different from supervised and unsupervised approaches—which adopts Markov decision processes to model the sequential interactions of a learning agent and its environment (i.e., platform users)[10]. RL agent-induced changes in user behavior represent a novel form of *behavior modification* (BMOD), classically defined as "an observable, replicable and irreducible component of an intervention designed to alter or redirect causal processes that regulate behavior"[11]. BMOD techniques derive from principles of behaviorist psychology and include nudging, herding, and operant conditioning, among others[4]. Such techniques, when integrated into digital platforms, are commonly termed *persuasive technology*[12], but are often no longer "observable, replicable and irreducible." Behavioral interventions range in their transparency: some are theoretically observable to users (e.g., chatbots, suggestions by recommender systems, and app notifications) while others are less so[13] (e.g,

A/B testing, feed filtering and comment moderation on social networks[14], and deceptive interface design choices[15]). Platforms employ BMOD to provide personalized services, increase user engagement, "hook" users by habit formation[16], generate further BBD, and more.

Potentially dangerous intended and unintended social and political outcomes arise from platforms employing BMOD. BMOD powerfully combines multiple cognitive, social, and algorithmic biases[17] resulting in filter bubbles[18], polarization[19], malicious third parties taking advantage of platform BMOD algorithms (e.g. via "adversarial attacks"[20] and social bots[21]), manipulation of public opinion ("sowing discord"[22]) and elections[23], and participation in social movements[24]. Compounding these problems, tangled network dynamics make it difficult to ascribe responsibility to a single actor, human or otherwise[25].

Our central claim is that while BMOD used by platforms has a potentially enormous impact on individuals and societies, academics are effectively barred from investigating it. Researching the nature, causes, and effects of scientifically and socially important topics, and developing and evaluating data science methodology for conducting such research, is increasingly difficult for anyone except researchers at platforms. Without access to data on both the *human behaviors* and related *machine behaviors*, and sufficient access to (or information on) the causal BMOD mechanisms, academic research becomes both more complicated and constrained[26], possibly leading to disciplinary stagnation. The resulting academic isolation threatens to stifle valuable contributions to public policy and regulation at a time when evidence-based, not-for-profit stewardship of global collective behavior is more important than ever[27].

The role of academic data science research in advancing knowledge

Academic data science researchers contribute to advancing scientific and practical knowledge in different ways. Methodological data science contributions might use a dataset for illustration, but the dataset could be interchangeable. In contrast, an applied data science contribution is dataset/application-dependent, where the methodology can be interchangeable. We focus on data science academics working on platform-related methods or applications.

Methodologists develop, propose, and study methodologies applied to behavioral and social data related to platforms. These include a range of methods for the entire "data pipeline," from study design and data acquisition, all the way to deployment. Examples include data acquisition policies[28], methods for handling missing values in predictive tasks[29], adjusting for self-selection in large-scale human impact studies[30], and identifying heterogeneous treatment effects[31]. Further examples include evaluating the impact of behavioral data size and richness on predictive ability[32]; proposing and comparing model explainability approaches[33] and alternatives[34]; developing sampling designs from networked data[35] and unbiased treatment effect estimators in adaptive experiments[36]; and translating legal notions of discrimination into automated algorithmic solutions[37].

The applied academic data science community advances knowledge of algorithms for analyzing, predicting, and modifying human and social behavior on platforms, and on the implications of applying such methods. This is done by applying data science models, algorithms, and approaches for purposes such as marketing and risk management[6,32,38,39]. Some work has uncovered the ability of predictive algorithms, such as predictive text in search queries, to capture surprising and more truthful responses than traditional data collection methods such as surveys[40]. Other work has studied the appropriateness of predictive text in email[41]. Key platform functionality can rely on image recognition and word embedding algorithms. A study showing discriminatory behavior of facial recognition algorithms[42] has led to improvements of such algorithms. A related area of research has developed techniques to de-bias word embeddings[43], which can exacerbate gender stereotypes when used in predictive text (i.e., auto-complete) applications. Such work has led to new knowledge in the behavioral and social sciences, including political science[44], sociology[45,46], and psychology[47]. Besides these more established disciplines, data science academics can also play a role in advancing the state of knowledge of *machine behavior*[48], an emerging field focusing on the empirical study of behaviors of "intelligent machines" used in automated decision-making.

Industry-academia collaborations as knowledge enablers

The academic community's rich contributions have often resulted from collaborations with "traditional" industries that collect BBD, such as telecoms, retailers, insurance companies and healthcare organizations, and more recently with platforms such as search engines, social media, dating websites, and messaging apps. These collaborations have helped academics identify relevant data science challenges and opportunities in real settings. And importantly, they have also provided them access to such data (via purchase, collaborative studies, or otherwise). Nevertheless, while the amount of data collected has grown exponentially, academic data scientists have access to a far smaller *proportion* of existing data than ever before[49].

In typical collaborations, academics obtain a one-shot cross-sectional large behavioral dataset from the company. For example, using a large dataset of call detail records of customers of a telecom company during a given period, indicating who called whom, researchers showed the ability to predict product interest[38] and churn[50]. In the case of an intervention, the data might include multiple snapshots in time (e.g. before the intervention, during the intervention, and after the intervention). One example is a controversial collaboration between Facebook and academic researchers, where a large-scale randomized experiment on Facebook users revealed that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness[51]. Another example is the collaboration between academic

researchers and a major online dating site, where a large-scale randomized experiment revealed that anonymous browsing harms users' matching ability, especially for women, by eliminating the anonymous user's ability to send a "weak signal" to potential mates by visiting their profile pages[45].

Industry-academia collaborations that involve conducting randomized experiments require the ability to intervene in the platform according to best research practices. Yet research choices such as what is measured, who and what is manipulated, and how the experiment is deployed are governed by practical business and deployment considerations. Complete academic freedom in the deployment and intervention, even in a simple randomized experiment, such as an A/B test, might already be practically infeasible. For example, the deployment of an A/B test will be limited to a sample of users from a certain area or background based on the industry partner's reach and business considerations; the experiment would be limited to a very short, specific time period (e.g., holiday season) which in itself might impact the study generalizability. This issue of limited access and platform-driven constraints becomes amplified with multiple interventions, and even more extreme in the case of BMOD that involves ongoing personalized interventions. Platform insiders face a similar issue when trying to evaluate policies for RL-based personalized systems—without access to "live" data and real interactions, platform data scientists must resort to clever ways of repurposing logged offline data[52]. What is already difficult for company *insiders* thus becomes virtually impossible for academic *outsiders*.

Researchers who do not have collaborations with platforms resort to collecting BBD from websites and online forums (e.g. via web scraping and APIs), using existing public data repositories (e.g. UCI Machine Learning Repository, Kaggle, ImageNet), obtaining "grey-market" datasets[53], or using online labor markets, such as Amazon Mechanical Turk. Researchers turn to online labor markets for tasks that range from labeling data through experiments and surveys on crowd workers, to collecting data on crowd workers' social media posts with their consent. Both Facebook and Twitter, for instance, offer such an API. These APIs make it easy to retrieve data in an ethical (and legal) manner. However, APIs come with restrictions: Twitter provides both public and premium APIs, which vary in the number of tweets and number of days and years you can go back to retrieve tweets. In the wake of the Cambridge Analytica incident, Facebook has become extremely strict in providing access to its API, even for academic researchers[54], thereby (ironically) leading to less transparency about the value and use of their BBD.

The second author experienced this hurdle first-hand. In a joint study with political scientists, Facebook "like" data was obtained, with consent, from over 6,500 Flemish (Belgian) persons between March and June, 2018, using the public Facebook Graph API[55]. Participants were additionally asked to complete 12 survey questions about their media consumption and political preferences. Subsequent data science analyses found novel insights from this BBD: alternative rock music is predictive for being a left voter, whereas right voters seem to prefer techno music[44]. Followup research questions by reviewers naturally emerged, which required either asking participants additional questions or repeating the experiment in a different country or after a certain political event. Unfortunately, given the company's new policies, none of these research extensions were possible. Starting from August 2018, Facebook no longer provided us with access to the Facebook "like" data through their API[56]. A change to Facebook's data sharing policy thus effectively curtailed this research project.

High caliber industry-academic partnerships, such as Harvard's Social Science One, encounter major difficulties as well. While Facebook committed to sharing more than a petabyte of data with researchers through this initiative, internal disagreements and issues resulted in the initiative's co-founder leaving the project[57]. Although the Social Science One project eventually culminated in a large amount of Facebook data being released for research purposes, these data appear to permit only "broad group-level or time series trends and relationships of interest"[58]. We address the tension between data granularity and privacy and its relation to BMOD research in the Recommendations section.

In summary, academics face major obstacles in obtaining observational BBD and conducting experiments in collaboration with platforms. Crucially, however, these obstacles are multiplied as platforms increasingly deploy BMOD strategies on their users. We describe these new challenges in the following section.

Academic Data Science Research in the Realm of Behavior Modification

From collecting passive footprints to modifying user behavior

Cross-sectional datasets are useful for purposes of developing and evaluating predictive and descriptive analytics. They are also useful for simple intervention methods, such as A/B tests or factorial design experiments. However, the BBD industry, now dominated by platforms such as Google, Facebook, Amazon, Uber and their Chinese counterparts Baidu, Tencent, etc., has moved to a new stage of action, beyond mere prediction and A/B testing. A more interactive mode of user engagement is now used, where humans' feedback and reactions are the trigger for subsequent platform actions, creating human-to-machine, machine-to-human, and human-to-human feedback loops and chains of actions, reactions, and interactions. For example, a user's posts on Facebook might determine which news items the platform displays to him/her. The user's reaction to the news recommendation (clicking, sharing, or posting a reaction) then leads to further actions by the platform, as well as by the user's platform "friends," and so on.

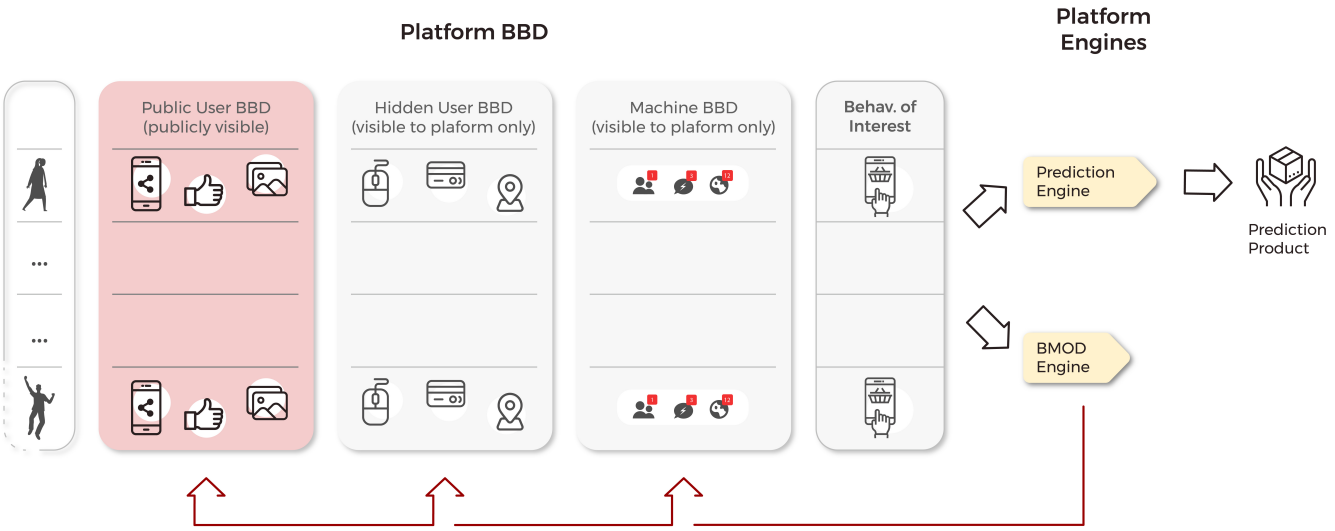

Figure 1. Human users' behavioral data and related machine data used for BMOD and prediction. Rows represent users. Academic researchers typically can access only publicly-visible user data.

Personalized interventions on platforms are increasingly governed by adaptive data-driven (i.e., RL) algorithms rather than by fixed business rules or managerial decisions[59]. For example, recommender systems used by popular commercial platforms such as TikTok, Pandora, Instagram and YouTube require vast amounts of interactive BBD—based on BMOD—in order to select an optimal recommendation policy for a given user[60,61]. LinkedIn uses multi-armed bandits (MABs), a simplified type of RL, for automating ad placement decisions[62] and Yahoo! uses it for personalized news recommendations[63]. RL formalizes human-machine interaction histories as sequences of state, action, reward "trajectories" generated in *online* (i.e., learning while interacting in real-time with users) or *offline* (i.e., learning from previously collected user interactions) contexts[10]. The RL agent's goal is to intervene in its environment of human users to learn an optimal *policy* maximizing the accumulation of designer-specified rewards (e.g., clicks). Although platforms have only recently begun to apply RL to interact with and control digital environments of human users, researchers have demonstrated empirically that RL agents can induce humans to exhibit a variety of "target behaviors" selected for by algorithm designers[64]. A growing literature now details how RL-based systems can fail—often due to poorly specified reward functions—and how these failures may have social consequences[65].

Understanding what these algorithms do, how to improve them, where potential harms lie, and which scientific questions can be answered in this new realm are critical areas for academic contributions. While industry has excellent data scientists and research groups, their incentives and goals (e.g. optimizing algorithmic performance) are different from those of academics. It is therefore crucial to have diverse academic research communities closely involved[27]. Yet, with this new mode of operation, many data science academics find themselves in new, isolated terrain. Figure 1 illustrates this new realm.

New needs for conducting scientific research

Conducting meaningful scientific research in the new realm of behavior modification requires direct access not only to the platform's human data, but crucially to strategies and behaviors of the algorithms interacting with the humans. The strategic use of algorithms to modify human behavior in a dynamic fashion means that studying human and/or algorithm behavior in this ecosystem requires access not only to the human BBD, but also to the related *machine BBD*—the machine's triggering and resulting behaviors.

The use of BMOD by platforms can interfere with research efforts by masking, changing, and even overriding effects of interest. For this reason, there is a growing niche of research focused on developing unbiased training and evaluation procedures using ideas drawn from causal inference[36,66]. As noted, in observational studies, users' BBD alone is insufficient for answering causal questions about human behavior and possible feedback loops between human behaviors and algorithm learning and actions. For example, an experiment evaluating the effect of anonymous browsing on matching outcomes in online dating found that anonymous browsing reduces matches by masking visits of anonymous browsers to profile pages[45]. Yet, if the site's recommender system uses the anonymous users' browsing information, it might recommend the anonymous users to those whose profiles they stealthily visited, thereby countering the researchers' intervention.

To assess clashes and interactions between the researchers' interventions and the platform BMOD, researchers would

therefore require sufficient information about the BMOD mechanism itself. Such information would include what user data the BMOD has access to, the RL algorithm type and parameters (states, rewards, policy type, etc.), and what types of actions the BMOD can initiate. For example, studying the effect of an intervention on social media requires not only a user's behavior (e.g., dwell time, posts, likes) but also the machine's notifications, recommendations, ads, etc., which function as unobserved confounders of observed effects[26]. A good example comes from a recent study investigating "the effect of message ad content on subsequent customer engagement" on Facebook which required the authors to make educated guesses about how Facebook's newsfeed Edgerank algorithm worked in order to to adjust for newsfeed algorithm's effect. The authors admit that a "perfect" solution to the selection problem is unlikely to be achieved without full knowledge of Facebook's targeting rule[67]. In reality, sufficient information about the BMOD mechanisms is unlikely to be available—even to the platform data scientists. While explaining the difficulties involved in developing an "offline" evaluation method for YouTube's RL-based recommender system, YouTube's own data scientists write, "there are multiple agents in our system, many of which we do not have control over[61]." Therefore, given the unavailability of the needed information about the BMOD mechanism, academics need access to the BMOD system in order to have any hope of "reverse engineering" the parts of the system relevant to the study of interest.

Access to machine data and the BMOD mechanism is also needed for methodological research: for studying and improving existing BMOD mechanisms. This includes developing BMOD-related methods such as "interpretability/explainability" algorithms (e.g., "why you're seeing this ad") used in developing "counterfactual" explanations of predictions from supervised models[68] and from RL systems[69]. Access to user data, machine data, and the BMOD mechanism is also needed for developing new methods to be deployed on platforms, which are likely to interact with platform BMOD. One example is developing text mining algorithms (e.g. for sentiment analysis or anomaly detection) that would be influenced by predictive text options displayed to the user while typing, thus inserting "anchoring bias"[13] into the observed data. Another example is developing link prediction—such an algorithm would be affected by a platform's existing link prediction BMOD, its feed filtering BMOD, and potentially other BMOD (e.g. due to integration of networks, such as LinkedIn and Office 365[70].) A researcher's knowledge of BMOD mechanisms could be helpful in deciding which questions are potentially answerable on that platform. BMOD information could also help researchers evaluating the historical performance of algorithms by guiding them in choosing periods of data where comparable BMOD mechanisms were in place. For example, professional networking platforms routinely send users notifications encouraging them to fill out missing "job skills" on their profiles. These personalized nudges create feedback loops affecting other BMOD mechanisms: adding data in response to the messaging BMOD leads the platform's job recommender system to generate new recommendations. A researcher comparing such recommender systems with batches of historical data from different time periods, without accounting for such BMOD actions, might then come to a conclusion about their relative performance for the wrong reasons[71]. In effect, the evaluation reveals more about the system's underlying data-generating dynamics than the algorithm's performance. This issue similarly applies to the case of the dating site experiment mentioned earlier. If researchers could verify the BMOD mechanism (i.e., the recommender system proposing matches to users) recommended "anonymous browsers" to users, they might adapt their experimental design accordingly, temper their causal claims, or abandon the idea altogether.

While data on human and machine behavior and the underlying BMOD algorithms are invaluable for advancing data science, they are typically inaccessible to academic data scientists for a variety of reasons including consumer privacy, trade secrets, proprietary content, and political sensitivities[49]. Much of the source code and data used for training algorithms is proprietary[48]. Reverse engineering these systems to expose biases introduced by BMOD may be illegal in the US under the Computer Fraud and Abuse Act (CFAA)[8]. Recently, however, a federal court ruled that the violation of websites' terms of service for the purpose of researching algorithmic discrimination is not criminal[72]. Nevertheless, problems of access still pose barriers for academics interested in *platform governance*[73] and *algorithmic auditing*[14]. In some cases, academic isolation has forced researchers to engage in "guerrilla tactics" using novel but ultimately inefficient (and possibly illegal) methods, such as creating bots and fake user profiles to probe and interact with platforms in hopes of making the "black box" more transparent[74].

Lastly, a further set of research challenges arises as BMOD can potentially affect behavior outside of the platform environment, by spillover of the BMOD effect to other platforms or offline, and through third parties who purchase platform data and prediction products[4] (predictions of user behavior). Studying human and machine behavior under BMOD therefore requires richer BBD over longer periods and from sources outside the platform, especially as it is currently unknown to what extent BMOD methods influence user behavior or extend to cognition more generally[75].

Table 1 summarizes these key differences between conducting academic research with and without platform BMOD, under three scenarios: (1) using observational BBD, (2) conducting randomized experiments (e.g. A/B tests), and (3) employing research-driven BMOD. We can map each of the examples described earlier to a row in the table, based on the researcher's study type and the platform BMOD functionality. In general, the differences pertain to the levels of visibility, access, and control that academic researchers have to both human and machine BBD and to platform BMOD mechanisms, as well as the types of data science applications or methodologies that can be studied.

Table 1. Contrasting needed access, data, and academic self-reliance in studies involving platform BMOD vs. no BMOD.

Study Type	Platform	Research Topics	Data/Access Needed	Academic Dependence
Observational	No BMOD	Prediction & insight on human behavior	• User BBD	• Few publicly available datasets • Platform BBD subject to API & platform policies • Reproducibility depends on policy changes
Observational	BMOD	Prediction & insight on human behavior, on machine behavior, and on human-machine interactions	• User BBD • Machine BBD	• No publicly available data • Platform human BBD shared with select academics • Platform machine BBD is extremely limited
A/B Test	No BMOD	Effect of intervention on human behavior	• User-level treatment group, measured outcome & relevant BBD • A/B test implementation details	• No publicly available data • Collaboration with select academics • Reproducibility possible
A/B Test	BMOD	Effect of intervention on human behavior, machine behavior, & human-machine interactions	• User-level treatment group, measured outcome, & relevant BBD • Machine BBD • A/B test implementation details • BMOD mechanism information (identify interference with researcher's A/B test)	• No publicly available data • Collaboration with select academics • Reproducibility unlikely (requires same platform BMOD)
BMOD by Researcher	No BMOD	Effect of specific BMOD on human behavior and society	• User-level BMOD data (human + machine)	• No publicly available data • Reproducibility possible (researcher uses same BMOD)
BMOD by Researcher	BMOD	Effect of specific BMOD on human behavior, machine behavior, human-machine interactions, and society	• Researcher's BMOD data (human + machine) • Platform's BMOD data (human + machine) • Platform's BMOD mechanism information (identify interference with researcher's BMOD)	• No publicly available data • Reproducibility impossible

Table 2. New challenges facing academic data science researchers

Challenge	Description
More complex ethics reviews	IRB members may not understand the complexities of algorithmic methods, such as RL, increasingly used by platforms.
New publication standards	A growing number of journals and conferences require proof of impact in deployment, as well as ethics statements of potential impact on users and society.
Reproducibility	Research using BMOD data by platform researchers or with academic collaborators cannot be reproduced by the scientific community.
Corporate scrutiny of research findings	Platform research boards may prevent publication of research critical of platform and shareholder interests.

New challenges facing data science academics

While it might be theoretically possible for an academic to simulate the BMOD mechanism and resulting user and machine data[52,76], or create a mock version of the digital platform, both of these "solutions" are neither practical nor completely effective in capturing the scale and networked-nature of real platform behavior. Industry-academia collaborations have therefore become the only feasible way for academic researchers to access not only the BBD (of humans and machines), but also the BMOD that platforms apply to their users. Yet very few platforms are willing to grant academics access to their human BBD, machine BBD, manipulations, and even fewer are likely to apply the researchers' intended manipulations to their users. A recent research project by NYU researchers to collect machine behavior on ads presented to 6,500 volunteer FB users for studying political ad targeting was thwarted by Facebook, despite the users' consent[77]. The recent growth of industry-academic "Fellow programs" (Facebook Research Collaboration, Google Faculty Awards, Amazon Machine Learning Research Awards, etc.) unfortunately does not solve the access problem either, as most existing programs involve only funding, without granting access to either data or to platform manipulation/testing systems.

Another problem relates to how ethical reviews for studies are conducted. The ethical implications of academic data scientists' work is typically scrutinized more closely than in industry, usually by university-level ethics boards (e.g., IRB in the United States). In the IRB context, human subjects research is quite broadly defined, but one aspect of the definition particularly relevant to data science is "manipulations of the subject or the subject's environment that are performed for research purposes"[78]. This broad definition (and we believe it is safer to interpret it as broadly as possible in the aftermath of Facebook's emotional contagion experiment[51]) would seem to encompass most if not all personalization efforts of platforms, including A/B testing, uplift modeling, and even RL.[79] While the idea of respecting and protecting human subjects is critical, many ethics boards are tuned to evaluating biomedical and traditional behavioral studies, and are not yet familiar with BBD-type data science research. This and a lack of access to BMOD mechanisms conflicts with IRB rules requiring ethics board members to have the expertise to properly evaluate large-scale studies involving behavioral interventions on platforms.

Academic researchers also face the challenge of new and more demanding publication requirements. Scientific journals and conferences increasingly require researchers to not only propose new methods, but to also illustrate their impact in actual deployment. For example, the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining solicits research papers and applied data science papers. Applied data science papers further distinguish between *deployed* and *evidential* papers. Papers in the deployed category "must describe implementation of a system that. . . is (or was) in production use for an extended period of time." Without a collaboration with a digital platform, this requirement is impossible to meet for BBD-based empirical research that involves A/B testing or BMOD. Academics who do manage to collaborate with platforms and conduct research based on behavior-modified data face challenges of scientific reproducibility, where their findings cannot be reproduced by the scientific community.

Finally, recent events have raised concerns that platforms and corporations may try to restrict the publication of academic data science research and industry-academia collaborative studies critical of their methodology and ethics[80]. As a senior scientist at Google (who was subsequently fired) puts it, "If we are researching the appropriate thing given our expertise, and we are not permitted to publish that on grounds that are not in line with high-quality peer review, then we're getting into a serious problem of censorship"[81].

Implications

The lack of access to human and machine behavior data from BMOD-deploying platforms, and sufficient information on the deployed BMOD mechanisms, isolates academic researchers and poses multiple challenges not only to the researchers themselves, but also to scientific research, humans, society, and governance[82].

Academic isolation can potentially affect scientific research in multiple ways. Academic data scientists have voiced concerns about problems in data infrastructure and sharing between industry and academia, IRB's inexperience with BBD, and the difficulty of reproducible research due to BMOD aimed at shaping user behavior to achieve platform goals[8,26,83]. In addition to these points, we worry that the study of timely and crucial data science topics will dwindle in academia, resulting in fewer ethically-approved studies in peer-reviewed publications. Published research might be dominated by studies conducted by platform researchers, and only some will have passed internal company ethics committees (e.g. Facebook's internal "research ethics board"[84]). Although well-intentioned, "ethics washing"—as a convenient form of corporate self-regulation—can serve to normalize risky new technologies and applications of BMOD[85]. Research by platform data scientists is also now frequently published in non-peer-reviewed repositories such as arXiv and SSRN. Besides leading to findings of questionable validity due to a lack of peer-review and reproducibility, the absence of standardized (e.g., IRB) ethics review could damage the reputation of data science in the eyes of the general public, as the Cambridge Analytica scandal revealed. Without the broad trust of public stakeholders, academic data scientists may struggle to receive the funding and resources necessary to carry out important research, potentially motivating talented academics to trade their academic posts for positions in industry. A *catch-22* situation currently exists as academics lack the access needed to study the effects of BMOD at scale; yet this lack of evidence may be used as a reason to withhold research funds or stymie legislation that would enable such research.

Academics with access to only human BBD and even machine BBD (but not the BMOD mechanism) are effectively limited to studying *interventional* behavior on the basis of *observational* data. Besides increasing the risk of false and missed discoveries, answering causal questions becomes nearly impossible. Independent academic researchers are therefore left without the opportunity or funding for conducting scientifically meaningful research with social implications. The current high-pressure system of academic incentives based on peer-reviewed publication arguably diverts researchers' attention towards less-relevant—even marginal—problems based on easier-to-collect data and simpler-to-implement projects. Platform data scientists, unlike their academic counterparts, may have neither the time nor incentives to conduct long-term research without immediate business impact. We are thus worried about misdirected research efforts and opportunity costs. On top of this, an over-reliance on the few platforms with accessible BBD (e.g. Twitter[83]), and towards performance on a few heavily-reused, industry-focused public datasets (e.g., MovieLens) may threaten the generalizability of research. Academics now operate in a more complex legal environment when using data controlled by platforms (e.g., the second author's experience and the NYU Ad Observatory experience[77]). Academics must consider how their research may be subject to regulation (e.g., the GDPR) or lawsuits by platforms claiming violation of their constantly evolving terms of service.

Finally, growing researcher isolation threatens to create a chilling effect on the production of scientific knowledge regarding the consequences of large-scale, long-term BMOD on individuals and society. In terms of scientific training, academics will need to redesign traditional data science curricula to include human subjects research ethics, BMOD and persuasion techniques, and the skills and training needed to deploy large-scale online learning algorithms[63]. This includes a new focus on the kinds of performance measures favored by industry-centered (e.g., dwell time and engagement) as opposed to academic research (e.g., recall@top10). Table 3 summarizes the ten implications mentioned.

Recommendations

Academic data science research in the BMOD era faces a growing number of barriers affecting the interests of an array of actors, including governments, academic institutions, platforms, professional societies, individual data scientists and platform users. A *polycentric*[86] approach with rule-making authorities at multiple levels seems promising in addressing this complex situation, due to its focus on mutual monitoring and adaptive learning over time. Emerging frameworks such as *responsible innovation*[87] may also provide useful guidance in joining public sphere debate by citizens and academics with scientific responsibility for new and risky technologies such as BMOD. Below we consider specific proposals for informal governance (bottom-up) and formal regulatory (top-down) measures that could serve to reintegrate academics in data science research.

Platforms and Universities

Starting with bottom-up changes by platforms, several changes could reduce barriers to academic research. First, platforms could incentivize data science teams to share results of internal experiments and studies by pursuing peer-reviewed publications. Platforms could also trigger new research by offering data science challenges open to the public. The 2006 Netflix Prize contest is an example of how such sharing can lead to breakthrough scientific research.

Better transparency in the use and results of BMOD experiments could go a long way toward addressing some of the challenges and guesswork currently faced by academic data scientists doing both applied and methodological research. While access to the platform's code itself is likely to be ineffective and unfeasible, access to institutional information about the BMOD can be extremely valuable. Researchers would benefit from knowing the purpose of the BMOD, which users are affected, what types of manipulation or data logging policies are used, what types of algorithms, and linkage of BMOD with other functions,

Table 3. Potential implications of academic isolation for scientific research

Implication	Description
1. Unethical research is conducted, but not published	Studies by platform data scientists may fail academic IRB review but still be conducted. Results are not shared with the scientific community.
2. More non-peer-reviewed publications	Posting on pre-print repositories (e.g. arXiv) or blogs benefits platform data scientists who do not have incentives for peer-reviewed publication. These posts tend to "go viral" via the popular media faster than peer-reviewed publications, due to breakthrough claims, platform reputation, and speed of publication.
3. Misaligned research topics and data science approaches	Isolation from platform BMOD mechanisms makes it difficult to identify relevant algorithmic issues and implications, study the behavior of existing systems, and develop and evaluate new data science algorithms. Instead, the focus is on issues that can be studied in isolation, often "overfitting" to one of the few publicly available datasets, using performance measures other than those used by industry.
4. Chilling effect on scientific knowledge and research	The lack of access to mechanisms diverts academic researcher efforts away from studying platforms that employ BMOD; platform data scientists do not have the time and incentives to conduct long-term research, which may have no business motivation. This combination results in lack of scientific knowledge and research in this crucial realm. Also, platforms may censor academic research critical of current methods and approaches.
5. Difficulty in supporting research claims	Opaque BBD mechanisms and strategies and in-access to machine BBD make it difficult for researchers to test their hypotheses, replicate findings, and evaluate interference of BMOD with effects of interest.
6. Challenges in training new researchers	Misalignment of industry practices (BMOD techniques) and academic teaching (non-BMOD data science techniques). Students in data science programs may lack exposure and experience in modifying *human* behavior; they may also be unfamiliar with the principles of human subjects research ethics typically taught in the biomedical and social sciences.
7. Wasted public research funds	Much academic work is funded by public sources, yet achieving generalizable knowledge is hampered by a lack of access to causal mechanisms behind BMOD. Government agencies distributing funds and IRB committees may not realize how lack of access limits the possibility of scientific research. This could lead to a negative feedback loop where data scientists then receive less funding, due to ungeneralizable findings.
8. Misdirected research efforts and insignificant publications	Academic promotion and tenure pressures may divert academic researchers' focus towards insignificant and practically-irrelevant research topics easier to implement and publish.
9. More observational-based research; research is slanted towards platforms with easier data access	Without access to BMOD mechanisms, academics resort to observational BBD. Selection bias and increased risk of false and missed discoveries are issues. Research also becomes slanted towards data from platforms with easier data access (e.g., Twitter). Generalizability is questionable. Obtaining observational data without platform consent runs risk of legal actions by users (e.g., GDPR), platforms (scraping data, e.g., *HiQ v. LinkedIn*) and governments (e.g., China).
10. Reputational harm to the field	If researchers lack access to platform BMOD mechanisms, we risk growing the divide between how the general public and media view data science (e.g., Netflix's *The Social Dilemma*) and how practitioners see themselves. With better access, academic data scientists could evaluate the social implications of behavioral modification; yet researchers may disagree about the ethical values guiding data science research. Without access to causal mechanisms, other disciplines may not trust results based purely on observational BBD.

such as prediction engines, etc. Intervention-level access would provide academic collaborators the ability to perform A/B tests and even BMOD on subsets of users—with ethics board approval, of course.

Universities and funding agencies can further help by creating incentives for platforms to collaborate with academics, such as reducing red tape and overhead charges. Academics can also learn important lessons. One is the importance of encouraging discussions on these matters by writing *Perspective* articles that shed light upon these issues. Academic researchers lacking access to BBD and platform BMOD mechanisms are encouraged to foster collaborations with research groups who do have access, and for the latter to be open to such collaborations (to the extent this does not violate agreements with platforms or any applicable laws), thereby expanding both research scope and access. International professional bodies that both academic and industry data scientists belong to, such as the ACM or IEEE, may play a role in facilitating these collaborations and in developing meaningful ethical guidelines for large-scale applications of BMOD[88,89]. Academic decision makers involved in editorial, promotion, and funding decisions should also realize that barriers to access constrain most academics, and that access decisions lie with platforms who tend to favor select institutions. Nevertheless, academic decision makers should encourage research aimed at fostering a better understanding of collective behavior on platforms, particularly BMOD interventions which help inform public policy and regulation[27].

Governments and Regulators

While the scope and social impact of BMOD on platforms make government regulation appear a natural solution to the problems outlined above, BMOD combines three features that make it difficult to borrow ready-made solutions from medical, pharmaceutical, or insurance industries, or for specific global hazards such as nuclear waste, cigarette smoking, or environmental pollution. Besides massively impacting society, academic BMOD research cannot be done without the cooperation of for-profit platforms. And further, unlike nuclear waste disposal or tax fraud, governments do not benefit from keeping BMOD research confidential for reasons of public safety or public interest. Even if novel regulatory measures are undertaken, BMOD is a global problem and—as with climate change—it is likely they will only have limited effect.

Despite these complications, academic data scientists still have an important role to play in BMOD governance. For example, researchers without financial conflicts of interest could participate in the independent auditing, compliance, and oversight of BMOD used by platforms. Various models of compliance monitoring and third-party assessment could be adapted from credit scoring, drug development, or the FTC's recent privacy compliance system used as part of a $5B settlement against Facebook[90]. Federal funds could be allocated to academics to study the effects of BMOD at scale, similar to the way academics are already developing and testing algorithms to detect deepfakes[91].

In the European context, EU-wide regulation might also be a way to remove barriers to academic data science on platforms. The EU's current General Data Protection Regulation (GDPR) offers legal tools such as data protection impact assessments (DPIAs) relevant to issues raised by large-scale BMOD. DPIAs are designed to achieve several goals including the systematic description of data processing, assessing the necessity and proportionality of data processing, evaluating risks to fundamental rights and freedoms, and mitigating risks through technological and organizational measures[92]. Academics thus could contribute not only to auditing and testing of large-scale systems implementing BMOD, but also to developing methodologies for DPIAs. A major caveat is that DPIA results are not required to be publicly disclosed[93]. For this reason, legal experts have suggested giving third parties with algorithmic expertise (i.e., academic data scientists) information rights to assess the results of DPIAs and convey the information to individual data subjects[94].

The European Commission's recently proposed 2021 Artificial Intelligence Act offers further opportunities for academic data scientists to research, monitor, and evaluate large-scale BMOD systems. In fitting with the EU's precautionary approach to new technology[95], systems posing "unacceptable" risk (e.g., social credit scoring and subliminal manipulation) will likely be prohibited[96]. Currently, however, it is not clear where BMOD fits in the regulation's three risk tiers: so-called "dark-patterns" based on fixed A/B testing might be lower risk than adaptive BMOD by RL-based systems and thus subject to lower standards of transparency and accountability. In any case, academics may act as experts tasked with monitoring "high-risk" systems—systems posing "significant risks to the health and safety or fundamental rights"— through "third-party conformity assessment," or by participation in regulatory sandboxes where innovative large-scale BMOD systems can be developed and tested[96].

The idea of using regulatory sandboxes to test BMOD systems before releasing them to market is an intriguing solution. Regulatory sandboxes were originally devised as a way to provide Fintech startups in the UK and US with an opportunity to compete against established banks and prove the safety and utility of novel financial products while being closely monitored by regulators[97]. Evidence gathered in the sandbox testing stage can lead to better assessments of a new technology's potential impact on the market and on consumers[98]. While a promising idea, the concept still requires considerable practical and legal clarification before becoming feasible. For one, it is not immediately obvious how regulatory sandboxes can promote greater collaboration between individual platform and academic data scientists. In any case, we see the EU's more top-down regulatory approach as a potential boon for academic data science research in the age of BMOD in the way it does justice to complex compromises between individual rights, novel yet potentially high-risk technologies, and regulatory innovation. Even for AI

systems “intended to distort human behavior," the Act permits academic research "if such research does not amount to use of the AI system in human-machine relations that exposes natural persons to harm and such research is carried out in accordance with recognised ethical standards for scientific research”[96].

Nevertheless, there is an inherent tradeoff between data privacy and data science research aimed at auditing BMOD platforms[99]. For example, the GDPR’s rights to delete or modify one’s personal data can interfere with research focused on understanding platform BMOD strategies over time. We realize there is no easy solution to the problems raised by platforms’ use of BMOD and encourage further development of data infrastructures and legal tools to promote fruitful industry-academic collaboration, such as the Social Science One project. These kinds of projects may be our best bet in easing barriers to socially-relevant data science research in the age of BMOD, but they will likely need substantial government buy-in and support to succeed. While certainly not perfect, university IRB review is designed to deal with ethical issues inherent to human subjects research, such as privacy, consent, and the fair distribution of benefits and burdens to research participants. As argued earlier, the IRB process should be updated and streamlined for BMOD-related research on platforms. But all of this is moot without access. Until things improve, profit-driven platforms, even with in-house ethics boards, cannot and should not be fully trusted to make ethical BMOD-related research decisions while insulated from public scrutiny.

The collective scientific and social costs of academic isolation in the era of platform-based BMOD are too great to ignore. Innovative regulation and changes in universities and professional bodies can help remove some, but not all, barriers. Creative solutions will likely arise through open-minded and transparent deliberation by stakeholders in academia, industry, and government. But we should not underestimate the efforts of determined and outspoken individuals[100]—from both industry and academia—who will be indispensable in advancing meaningful and responsible platform-related data science research in the new BMOD era.

References

1. Shmueli, G. Research dilemmas with behavioral big data. *Big data* **5**, 98–119 (2017).
2. Helmond, A. The platformization of the web: Making web data platform ready. *Soc. Media+ Soc.* **1**, 2056305115603080 (2015).
3. Srnicek, N. *Platform capitalism* (John Wiley & Sons, 2017).
4. Zuboff, S. *The age of surveillance capitalism: The fight for a human future at the new frontier of power* (Profile Books, 2019).
5. Gauci, J. *et al.* Horizon: Facebook’s open source applied reinforcement learning platform. *arXiv preprint arXiv:1811.00260* (2018).
6. Kosinski, M., Stillwell, D. & Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proc. national academy sciences* **110**, 5802–5805 (2013).
7. De Cnudde, S. *et al.* What does your facebook profile reveal about your creditworthiness? using alternative data for microfinance. *J. Oper. Res. Soc.* **70**, 353–363 (2019).
8. Olteanu, A., Castillo, C., Diaz, F. & Kıcıman, E. Social data: Biases, methodological pitfalls, and ethical boundaries. *Front. Big Data* **2**, 13 (2019).
9. Wu, A. X. & Taneja, H. Platform enclosure of human behavior and its measurement: Using behavioral trace data against platform episteme. *New Media & Soc.* 1461444820933547 (2020).
10. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).
11. Michie, S. *et al.* The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals behavioral medicine* **46**, 81–95 (2013).
12. Fogg, B. J. *Persuasive technology: using computers to change what we think and do* (Morgan Kaufmann, 2002).
13. Schneider, C., Weinmann, M. & Vom Brocke, J. Digital nudging: guiding online user choices through interface design. *Commun. ACM* **61**, 67–73 (2018).
14. Gorwa, R., Binns, R. & Katzenbach, C. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Soc.* **7**, 2053951719897945 (2020).
15. Mathur, A. *et al.* Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proc. ACM on Human-Computer Interact.* **3**, 1–32 (2019).
16. Eyal, N. *Hooked: How to build habit-forming products* (Penguin, 2014).

17. Menczer, F. 4 reasons why social media make us vulnerable to manipulation. In *Fourteenth ACM Conference on Recommender Systems*, 1–1 (2020).

18. Pariser, E. *The filter bubble: How the new personalized web is changing what we read and how we think* (Penguin, 2011).

19. Beam, M. A., Hutchens, M. J. & Hmielowski, J. D. Facebook news and (de) polarization: reinforcing spirals in the 2016 US election. *Information, Commun. & Soc.* **21**, 940–958 (2018).

20. Cao, Y., Chen, X., Yao, L., Wang, X. & Zhang, W. E. Adversarial attacks and detection on reinforcement learning-based interactive recommender systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1669–1672 (2020).

21. Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A. & Menczer, F. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2725–2732 (2020).

22. Dutt, R., Deb, A. & Ferrara, E. "senator, we sell ads": Analysis of the 2016 Russian Facebook ads campaign. In *International conference on intelligent information technologies*, 151–168 (Springer, 2018).

23. Liberini, F., Redoano, M., Russo, A., Cuevas, A. & Cuevas, R. Politics in the facebook era. evidence from the 2016 us presidential elections. https://warwick.ac.uk/fac/soc/economics/research/centres/cage/manage/publications/389-2018_redoano.pdf (December 9, 2020). Online Working Paper Series No. 389, Centre for Competitive Advantage in the Global Economy, Online, accessed December 18, 2020.

24. Clark, L. S. Participants on the margins:# BlackLivesMatter and the role that shared artifacts of engagement played among minoritized political newcomers on Snapchat, Facebook, and Twitter. *Int. J. Commun.* **10**, 235–253 (2016).

25. Elish, M. C. Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Sci. Technol. Soc.* **5**, 40–60 (2019).

26. Lazer, D. M. *et al.* Computational social science: Obstacles and opportunities. *Science* **369**, 1060–1062 (2020).

27. Bak-Coleman, J. B. *et al.* Stewardship of global collective behavior. *Proc. Natl. Acad. Sci. United States Am. (PNAS)* (2021).

28. Saar-Tsechansky, M., Melville, P. & Provost, F. Active feature-value acquisition. *Manag. Sci.* **55**, 664–684 (2009).

29. Saar-Tsechansky, M. & Provost, F. Handling missing values when applying classification models. *J. machine learning research* **8**, 1623–1657 (2007).

30. Yahav, I., Shmueli, G. & Mani, D. A tree-based approach for addressing self-selection in impact studies with big data. *MIS Q.* **40**, 819–848 (2016).

31. Athey, S. & Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci.* **113**, 7353–7360 (2016).

32. Martens, D., Provost, F., Clark, J. & de Fortuny, E. J. Mining massive fine-grained behavior data to improve predictive analytics. *MIS quarterly* **40**, 869–888 (2016).

33. Ramon, Y., Martens, D., Provost, F. & Evgeniou, T. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, lime-c and shap-c. *Adv. Data Analysis Classif.* 1–19 (2020).

34. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).

35. Walker, D. & Muchnik, L. Design of randomized experiments in networks. *Proc. IEEE* **102**, 1940–1951 (2014).

36. Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S. & Athey, S. Confidence intervals for policy evaluation in adaptive experiments. *Proc. Natl. Acad. Sci.* **118** (2021).

37. Wachter, S., Mittelstadt, B. & Russell, C. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Comput. Law & Secur. Rev.* ((forthcoming)).

38. Hill, S., Provost, F., Volinsky, C. *et al.* Network-based marketing: Identifying likely adopters via consumer networks. *Stat. Sci.* **21**, 256–276 (2006).

39. Tobback, E., Bellotti, T., Moeyersoms, J., Stankova, M. & Martens, D. Bankruptcy prediction for smes using relational data. *Decis. Support. Syst.* **102**, 69–81 (2017).

40. Stephens-Davidowitz, S. & Pabon, A. *Everybody lies: Big data, new data, and what the internet can tell us about who we really are* (HarperCollins New York, 2017).

41. Robertson, R. E., Olteanu, A., Diaz, F., Shokouhi, M. & Bailey, P. "i can't reply with that": Characterizing problematic email reply suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–18 (2021).

42. Buolamwini, J. & Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91 (2018).

43. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Adv. neural information processing systems* **29**, 4349–4357 (2016).

44. Praet, S., Van Aelst, P., Martens, D. *et al.* I like, therefore i am. predictive modeling to gain insights in political preference in a multi-party system. *Res. paper, Univ. Antwerp, Fac. Bus. Econ.* 1–34 (2018).

45. Bapna, R., Ramaprasad, J., Shmueli, G. & Umyarov, A. One-way mirrors in online dating: A randomized field experiment. *Manag. Sci.* **62**, 3100–3122 (2016).

46. Pentland, A. *Social physics: How good ideas spread-the lessons from a new science* (Penguin, 2014).

47. Matz, S. C. & Netzer, O. Using big data as a window into consumers' psychology. *Curr. Opin. Behav. Sci.* **18**, 7 – 12 (2017). Big data in the behavioural sciences.

48. Rahwan, I. *et al.* Machine behaviour. *Nature* **568**, 477–486 (2019).

49. King, G. & Persily, N. A new model for industry–academic partnerships. *PS: Polit. Sci. & Polit.* **53**, 703–709 (2020).

50. Verbeke, W., Martens, D. & Baesens, B. Social network analysis for customer churn prediction. *Appl. Soft Comput.* **14**, 431–446 (2014).

51. Kramer, A. D., Guillory, J. E. & Hancock, J. T. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci.* **111**, 8788–8790 (2014).

52. Li, L., Chu, W., Langford, J. & Wang, X. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 297–306 (2011).

53. Weller, K. & Kinder-Kurlanda, K. E. A manifesto for data sharing in social media research. In *Proceedings of the 8th ACM Conference on Web Science*, 166–172 (2016).

54. Bastos, M. & Walker, S. T. Facebook's data lockdown is a disaster for academic researchers. https://theconversation.com/facebooks-data-lockdown-is-a-disaster-for-academic-researchers-94533 (April 11, 2018). The Conversation, Online, accessed May 1, 2020.

55. Graph api. https://developers.facebook.com/docs/graph-api (December 9, 2020). Facebook, Online, accessed September 24, 2020.

56. Schroepfer, M. An update on our plans to restrict data access on Facebook. https://about.fb.com/news/2018/04/restricting-data-access (December 19, 2020). Facebook, Online, accessed September 26, 2020.

57. Mattu, S., Yin, L., Waller, A. & Keegan, J. How we built a facebook inspector. https://themarkup.org/citizen-browser/2021/01/05/how-we-built-a-facebook-inspector (January 5, 2020). The Markup, Online, accessed January 9, 2021.

58. Messing, S. *et al.* Dataverse. https://socialscience.one/facebook-dataverse (June 5, 2020). Social Science One, Online, accessed June 20, 2021.

59. den Hengst, F., Grua, E. M., el Hassouni, A. & Hoogendoorn, M. Reinforcement learning for personalization: A systematic literature review. *Data Sci.* 1–41 (2020).

60. Zhou, S. *et al.* Interactive recommender system via knowledge graph-enhanced reinforcement learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 179–188 (2020).

61. Chen, M. *et al.* Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 456–464 (2019).

62. Tang, L., Rosales, R., Singh, A. & Agarwal, D. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 1587–1594 (2013).

63. Li, L., Chu, W., Langford, J. & Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, 661–670 (2010).

64. Dezfouli, A., Nock, R. & Dayan, P. Adversarial vulnerabilities of human decision-making. *Proc. Natl. Acad. Sci.* **117**, 29221–29228 (2020).

65. Whittlestone, J., Arulkumaran, K. & Crosby, M. The societal implications of deep reinforcement learning. *J. Artif. Intell. Res.* **70**, 1003–1030 (2021).

66. Schnabel, T., Swaminathan, A., Singh, A., Chandak, N. & Joachims, T. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, 1670–1679 (PMLR, 2016).

67. Lee, D., Hosanagar, K. & Nair, H. S. Advertising content and consumer engagement on social media: Evidence from facebook. *Manag. Sci.* **64**, 5105–5131 (2018).

68. Verma, S., Dickerson, J. & Hines, K. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).

69. Puiutta, E. & Veith, E. M. Explainable reinforcement learning: A survey. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 77–95 (Springer, 2020).

70. Lardinois, F. Microsoft finally starts doing something with linkedin by integrating it into office 365. https://techcrunch.com/2017/09/25/microsoft-finally-starts-doing-something-with-linkedin-by-integrating-it-into-office-365/ (September 25, 2017). Tech Crunch, Online, accessed June 25, 2021.

71. de Myttenaere, A., Le Grand, B., Golden, B. & Rossi, F. Reducing offline evaluation bias in recommendation systems. In *23rd annual Belgian-Dutch Conference on Machine Learning (Benelearn 2014)*, 55–62 (2014).

72. Summary judgment opinion. https://www.aclu.org/legal-document/summary-judgment-opinion-0 (March 27, 2020). ACLU, Online, accessed June 22, 2021.

73. Gorwa, R. What is platform governance? *Information, Commun. & Soc.* **22**, 854–871 (2019).

74. McGuigan, L. This tool lets you confuse google's ad network, and a test shows it works. https://www.technologyreview.com/2021/01/06/1015784/adsense-google-surveillance-adnauseam-obfuscation/ (January 6, 2021). MIT Technology Review, Online, accessed June 28, 2021.

75. Russell, S. *Human compatible: Artificial intelligence and the problem of control* (Penguin, 2019).

76. Yao, S. *et al.* Measuring recommender system effects with simulated users. *arXiv preprint arXiv:2101.04526* (2021).

77. Horwitz, J. Facebook seeks shutdown of nyu research project into political ad targeting. https://www.wsj.com/articles/facebook-seeks-shutdown-of-nyu-research-project-into-political-ad-targeting-11603488533 (October 23, 2020). Wall Street Journal, Online, accessed January 15, 2021.

78. Activities that require IRB review. https://research.uci.edu/compliance/human-research-protections/researchers/activities-irb-review.html. UCI, Online, accessed December 18, 2020.

79. Shmueli, G. "Improving" prediction of human behavior using behavior modification. *arXiv preprint arXiv:2008.12138* (2020).

80. Fried, I. Scoop: Google CEO pledges to investigate exit of top AI ethicist. https://www.axios.com/sundar-pichai-memo-timnit-gebru-exit-18b0efb0-5bc3-41e6-ac28-2956732ed78b.html (December 9, 2020). Axios, Online, accessed December 18, 2020.

81. Dave, P. & Dastin, J. Google told its scientists to 'strike a positive tone' in ai research - documents. https://www.reuters.com/article/us-alphabet-google-research-focus-idUSKBN28X1CB (December 23, 2020). Reuters, Online, accessed December 23, 2020.

82. Kitchin, R. Thinking critically about and researching algorithms. *Information, Commun. & Soc.* **20**, 14–29 (2017).

83. Tufekci, Z. Big questions for social media big data: representativeness, validity and other methodological pitfalls. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8 (2014).

84. Boka, Z. Facebook's research ethics board needs to stay far away from facebook. https://www.wired.com/2016/06/facebooks-research-ethics-board-needs-stay-far-away-facebook/ (June 23, 2016). Wired Magazine, Online, accessed September 26, 2020.

85. Bietti, E. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 210–219 (2020).

86. Ostrom, E. Polycentric systems for coping with collective action and global environmental change. *Glob. environmental change* **20**, 550–557 (2010).

87. Stilgoe, J., Owen, R. & Macnaghten, P. Developing a framework for responsible innovation. *Res. Policy* **42**, 1568–1580 (2013).

88. Delacroix, S. & Wagner, B. Constructing a mutually supportive interface between ethics and regulation. *Comput. Law & Secur. Rev.* **40**, 105520 (2021).

89. IEEE. The IEEE global initiative on ethics of autonomous and intelligent systems. ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2. ieee.org/develop/indconn/ec/autonomous_systems.html (2017).

90. FTC. Ftc imposes $5 billion penalty and sweeping new privacy restrictions on facebook. www.ftc.gov/news-events/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions (June 12, 2020). FTC, Online, accessed June 16, 2021.

91. Adee, S. World's first deepfake audit counts videos and tools on the open web. https://spectrum.ieee.org/tech-talk/computing/software/the-worlds-first-audit-of-deepfake-videos-and-tools-on-the-open-web (October 7, 2019). Technology Review, Online, accessed June 20, 2020.

92. Official Journal of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Tech. Rep., European Parliament and Council (2016).

93. Kaminski, M. E. & Malgieri, G. Multi-layered explanations from algorithmic impact assessments in the gdpr. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 68–79 (2020).

94. Veale, M., Binns, R. & Ausloos, J. When data protection by design and data subject rights clash. *Int. Data Priv. Law* **8**, 105–123 (2018).

95. Tosun, J. How the eu handles uncertain risks: Understanding the role of the precautionary principle. *J. Eur. Public Policy* **20**, 1517–1528 (2013).

96. Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence (2021). European Commission, Online, accessed June 22, 2021.

97. Allen, H. J. Regulatory sandboxes. *Geo. Wash. L. Rev.* **87**, 579 (2019).

98. Jeník, I. & Duff, S. How to build a regulatory sandbox: A practical guide for policy makers. https://www.cgap.org/sites/default/files/publications/2020_09_Technical_Guide_How_To_Build_Regulatory_Sandbox.pdf (2020).

99. Aral, S. *The hype machine* (New York: Currency, 2020).

100. Sadowski, J., Viljoen, S. & Whittaker, M. Everyone should decide how their digital data are used–not just tech companies. *Nature* **595**, 169–171 (2021).