THE INVERSION PROBLEM:
Why Algorithms Should Infer Mental State and Not Just Predict Behavior

October 6, 2023

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan and Manish Raghavan[1]

Abstract

More and more machine learning is applied to human behavior. Increasingly these algorithms suffer from a hidden - but serious - problem. It arises because they often predict one thing, while hoping for another. Take a recommender system: it predicts clicks but hopes to identify preferences. Or an algorithm that automates a radiologist: it predicts in-the-moment diagnoses while hoping to identify their reflective judgements. Psychology shows us the gaps between the objectives of such prediction tasks and the goals we hope to achieve: people can click mindlessly; experts can get tired and make systematic errors. We argue such situations are ubiquitous and call them "inversion problems": the real goal requires understanding a mental state that is not directly measured in behavioral data but must instead be inverted from the behavior. Identifying and solving these problems requires new tools that draw on both behavioral and computational science.

[1] Kleinberg, J., Ludwig, J., Mullainathan, S., & Raghavan, M. The Inversion Problem Why Algorithms Should Infer Mental State and Not Just Predict Behavior. Perspectives on Psychological Science (0, 0). Copyright [2023] (The Authors). DOI: [https://doi.org/10.1177/17456916231212138].
Thanks to Rafael Batista, Linda Chang, and Anuj Shah for very helpful discussions, to Emmanuel Roman and the Center for Applied Artificial Intelligence at the University of Chicago Booth School of Business for support, and to the editor Sudeep Bhatia and two anonymous referees for comments. Any errors and opinions are of course our own.
Kleinberg: Cornell University (kleinberg@cornell.edu); Ludwig: University of Chicago and National Bureau of Economic Research (jludwig@uchicago.edu); Mullainathan: University of Chicago and National Bureau of Economic Research (sendhil@chicagobooth.edu); Raghavan: MIT (mragh@mit.edu).

1. Introduction

There are two ways to analyze data about people, two ‘cultures’ of empirical work if you will (Snow, 1959, Breiman, 2001). One - call it the *psychology culture* - is now over a century old and entirely familiar. For behavioral scientists,[2] data is a means to an end, to be used to improve our theories about the human mind. The data serves to test competing theories and develop new ones. Ultimately the data is all about giving us a sense of what theories are correct and important.

The other culture - call it the *machine learning culture* - is newer but growing quickly. This culture is all about using large amounts of behavioral data to *predict* what people will do. The algorithms that result from this culture now operate at a vast scale in ever-expanding segments of society, including: curating content for people on social media; recommending products (books, movies, etc.); automating expert decisions.

The machine learning culture stands in stark contrast with the psychology culture. It touts as successful algorithms that successfully predict behavior largely without utilizing the theoretical insights of psychology. For many applications, theories are only a means to an end - and with enough data, the value of theory, so goes the argument, fades.

Our argument here is that the power of this approach for specific applications obscures a deeper set of limitations. The machine learning culture is already over-extended and is at risk of over-extending far more. It is particularly dangerous because it inadvertently hides its own failures. It lures us to focus on how well these algorithms predict behavior, leading us to overlook something comparably important.

Consider the following example (Kleinberg, Mullainathan and Raghavan 2022). Imagine a ‘smart’ pantry that offers an appealing proposition: by observing your eating patterns, the smart pantry promises to learn your preferences and keep your kitchen suitably stocked. (Assume smart pantries are sold like other appliances - so the only goal of the producer is to make a pantry you enjoy long-term, so that you will buy another one or tell your friends to get one.)

In one sense, the smart pantry proves a success. Your staples are ordered regularly so you never run out; unlike a human shopper, the smart pantry never forgets the milk. The smart pantry even delights you by ordering samples of food you didn’t know existed but turn out to like.

In another sense, the smart pantry is a dismal failure. Doritos are your Achilles’ heel. If you have them in front of you, you will eat them. Actually, “eat them” is too dignified a phrase for what you do to the bag. You have a self-control problem (see for example Thaler and Shefrin, 1981, Ainslie, 1992, Loewenstein and Prelec, 1992, Laibson, 1997, O’Donoghue and Rabin, 2000, Muraven and Slessareva, 2003, Fujita et al 2006 and Baumeister et al 2007, Milkman, Rogers and Bazerman, 2009, Inzlicht et al. 2021, Ward and Mann 2022). It is a war of selves – the one

[2] In what follows we use the term “psychology” and “behavioral science” interchangeably and broadly; many of our examples come from the biases-and-heuristics literature but there is nothing about our larger argument that is limited to that perspective, as we discuss further below. The use of “behavioral science” rather than just “psychology” is intended to recognize that much of the relevant work can also be done by behavioral economists, sociologists and computer scientists, among others.

that wants to eat Doritos and the one that does not want to invest in a whole new wardrobe because your old clothes are suddenly too snug.

Your solution was to broker a peace treaty. You do not buy large bags of Doritos but a life without Doritos is no life at all – so you keep around a few small bags. Your Doritos-loving-self gets some of what it wants but is forced to show some restraint.

Things were going just fine. And then the smart pantry showed up. Slowly small bags were replaced by large bags. And then there were more and more large bags. The result: you find yourself eating a lot more Doritos than you want to. So while there's a lot to love about your smart pantry, on net you're ambivalent. If some researchers asked you what you would pay for the smart pantry, you'd be somewhere between "take it or leave it" and "take it, please take it."

This smart pantry example is fictional, but the concern it illustrates isn't. An algorithm that recommends Tweets to people is trained using data on their past behavior looking at and engaging with Tweets. Unfortunately many Tweets are like temptation goods - the digital equivalent of a giant bowl of Doritos. To see why, consider how emotionally charged so many Twitter threads are. Research in psychology has contrasted choices made in 'hot' states with what we want in 'cool' states (see for example Loewenstein, 1996, Metcalfe and Mischel, 1999, Read and Van Leeuwen, 1998). The Twitter algorithm is giving us ever more and larger digital bags of Doritos in our Twitter feed. The data reveal the consequence of no one having recognized this fact sooner: Randomized experiments show that people who get disconnected from social media wind up being happier (Tromholt, 2016; Allcott, Gentzkow and Song, 2022).

The applicability of the machine learning culture depends on the deeper goal of the task at hand; specifically, in what we seek to glean from the data. The smart pantry promises to learn your preferences,[3] but it merely learns what you will eat. If our goal were to increase eating, this would be a (smart?) pantry. But if our goal were to improve well-being, we must confront a problem. The behavioral data (what is eaten) does not perfectly reflect the mental state we care about: people recognize and admit that they sometimes eat food that does not promote their well-being.

We call such problems - where the goal involves mental states not directly measured in behavioral data - as *inversion problems.* To make good use of behavioral data, we must rely on psychological insights about how mental state translates into behavior so that we can then invert mental state from the behaviors we observe in the data.

Inversion problems do not succumb to the implicit argument behind the machine learning culture, that more data alone is a solution. More data alone is *not* a solution. More data may help us predict behavior better but better behavioral predictions alone do not translate into clearer inferences of mental state. We need some understanding of the psychological complexity by which mental states translate into behavior.

[3] By "preferences" here we mean what people would choose if they chose more deliberately and consciously, what some psychologists call 'system 2' and others call 'the self.' We recognize that we cannot ever hope to measure *true* preferences (what that might even mean).

We argue that inversion problems are ubiquitous (Section 2). They likely arise whenever machine learning algorithms are trained on data generated by human behavior. In very few applications are we interested in the measured behavior alone; behavior is often merely a proxy for mental state. (See Samuelson, 1938, for an early articulation within economics of why behavior should be the focus - revealed preference - and see Thaler and Sunstein 2008 and Kahneman 2011 for excellent overviews of the vast literature within psychology about how and why behavior need not reflect mental states.) Inversion problems extend well past the self control example discussed here. For example, algorithms trained to automate expertise merely predict expert judgments. But a rich psychology tells us that expert judgments (the behavior we predict) do not accurately reflect the expertise (the mental state we seek to infer).

We argue that naively treating inversion problems as pure prediction problems has and will continue to prove problematic (Section 3). For example, it provides one explanation for why social media (in particular, algorithms that curate content) seems to leave so many users unsatisfied. Finally, we will argue that creating techniques for combining machine learning with psychology is a goal that both disciplines can contribute to and should work collaboratively towards (Section 4). While psychologists are often involved in helping think about how humans interact with algorithms on the back end once the algorithm is constructed (Card et al., 1983, Shneiderman 1986, Preece et al., 1994, Carroll 1997, Johnson 2020, Dix 2003, Hassenzahl and Tractinsky 2006, Hassenzahl et al 2010, Hartson and Pyla 2012, Helander 2014, Lazar 2017, Card 2018), they are far less frequently involved in helping think about the implications of the human behavior captured by the training data for the construction of a given algorithm. We do not mean to imply that there is *never* any attention to the underlying psychology of the human beings whose behavior generates the training data for algorithms; see for example Green and Daniels (2014), Joachims et al. (2017), Chan, Critch and Dragan (2021), Trueblood et al. (2021) and the studies reviewed by Bhatia and Aka (2022). But this sort of careful attention to how behavioral science insights about people affect computer science decisions around algorithm construction remains far too rare, and the conceptualization of the problem as one of inversion rather than pure prediction problems is, we believe, new.

2. Prediction vs. Inversion

2.1 Sonar

Imagine a submarine commander, encased in metal, submerged in the deepest, darkest part of the ocean, with neither windows nor lights. The commander ideally wants what a video camera could provide: a picture of the ocean depths surrounding her. Instead all she has is a string of data that by themselves are meaningless: how long it takes sound waves to return to the submarine.

To help the submarine commander use these sonar pings to "see," we rely on our knowledge of physics. The speed of sound waves in water turns out to depend on (among other things) the water's temperature, pressure and salinity. By using a structural model from physics, sonar converts data on sound wave return times to a measure of distance to the nearest object. By modeling these physical processes, we can *invert* something useful (e.g. is there an enemy sub nearby) from a string of not-directly-helpful sonar-ping data. These sorts of inversion problems are rampant in natural sciences extending far beyond sonar; see for example Tarantola (2005).

Often, behavioral data is like sonar data, except less transparently so. Making sense of sonar data obviously requires physics. Yet, behavioral data *seems* like it can be directly interpreted. We argue that in many applications that sense of direct interpretability is an illusion.

An understanding of behavioral science quickly reveals that when it comes to analyzing human behavior, we are much closer to submarine commanders; our data on people is much closer to sonar pings than to videos. That is because, in many applications, we are not interested in the data that are observed, like behavior, but rather in unmeasured mental states, such as preferences or knowledge. Like inferring surroundings from sonar pings, we must infer these mental states from behavioral data.

We illustrate these inversion challenges with two canonical algorithmic applications: Curation; and automating expertise. These are not meant to be exhaustive but merely illustrative of the breadth of such problems.

2.2 Curation

What we call curation algorithms are ubiquitous in modern life. In a world with so many options, we need some way to sift and rank them for customers. Retailers have many items; entertainment companies have oceans of content; social media companies have many posts; and so on.

Though such algorithms can differ dramatically in how they are implemented, their construction typically has a shared foundation. Take the case of content on social media, which of course has itself been the subject of a great deal of research in behavioral science (see for example Vogel et al., 2014, Neubaum and Kramer 2017, Hunt et al. 2018, Bayer 2020, Allcott, Gentzkow and Song, 2022). When you log on, the content could in principle just be ordered by (say) when it was posted and then it is left up to you to scroll through the hundreds or even thousands of posts since you last logged on. There’s a reason few social media platforms work like that: it’s very cumbersome and not very helpful.

A curation algorithm attempts to rank your social media feed to put the posts you would like the most nearest to the top. While the mechanics of the algorithm can vary dramatically across companies, in essence they all have a shared foundation. They take data on you (e.g. which posts have you previously liked or interacted with) and use that data to build a predictor of what kind of post you are likely to interact with in the future. That prediction of how you will engage with a post is how the algorithm then ranks the ocean of content.

Looking for inversion problems makes one question transparent: what is the goal of a curation algorithm? If the deeper goal is simply to maximize engagement then there is no problem. But if the goal were to maximize user *satisfaction*, then this raises a behavioral (inversion) question: in what ways does someone’s observed engagement level with a post not fully reveal whether they like that post?

Of course, a large literature in psychology shows numerous ways in which behaviors deviate from preferences. For example, we have self-control problems (Thaler and Shefrin 1981; Ainslie, 1992; Loewenstein and Prelec, 1992; Laibson, 1997; O'Donoghue and Rabin 2000; Muraven and Slessareva 2003; Fujita et al. 2006; Baumeister et al 2007; Milkman, Rogers and Bazerman 2009; Inzlicht et al. 2021; Ward & Mann 2022); our aspirations can differ from our momentary impulses (our "shoulds" are not our "wants"). That means our actions can deviate from our intentions.

As another example, we might make our choices without much conscious deliberation, almost automatically (see for instance Trueblood et al. 2018). Those automatic behaviors are generally of enormous value to us. They help us deal with common situations we encounter in our daily lives over and over again in a way that is usually adaptive (see for example Nisbett and Wilson 1977; Gilbert 1991; Jacoby 1991; Bargh 1994; Kahneman and Frederick 2005; Chaiken and Trope 1999; Haidt 2001; Wilson 2002; Kahneman 2011; Gawronski and Creighton 2013). In situations where the environment is fairly predictable and people have a chance to learn situational regularities through repeated exposure and practice, automatic responses can be better - sometimes even substantially better - than more deliberate ones (see for example Kahneman and Klein 2009 and Gigerenzer and Gaissmaier 2011).

But as a large literature in psychology has documented, automatic responses can sometimes also lead to problems - to lead us to act in ways that our more deliberate selves do not necessarily want. For example, we may stereotype; in-group bias tends to be more likely when we are - roughly speaking - acting automatically (e.g., Devine, 1989, Gilbert & Hixon, 1991, Greenwald, McGhee & Schwartz, 1998, Cunningham et al., 2004). Our body language can be inadvertently more open towards people like us (in-group members) than towards people not like us (out-group members). So our past choices might include subconscious bias that our deliberate selves wish wasn't there, which can then lead to algorithmic bias (see for example Dwork et al. 2012; Barocas and Selbst 2016; Chouldechova 2016; White House Report on Big Data 2016; Lum 2017; Kleinberg, Mullainathan and Raghavan 2017; Obermeyer et al. 2019; and Gebru 2020).

There are, in other words, countless inversion problems here that we are neglecting.

2.3 Automating Expertise

Another category of algorithms is those that attempt to automate human experts. Grading student essays. College admissions. Resume screening. Judge decisions. Medical diagnosis.

A very commonly discussed example from medicine will make the problem clear: automating the reading of an x-ray. One might think such algorithms are built to read x-rays using biomedical ground truths. In fact, in almost all cases, they are instead built using data on the behavior of the expert. Physiological ground truth is rarely available: what is available is typically a diagnosis provided by an expert. (For example, even something that seems like a physical ground truth, like the results of a biopsy, is often actually an expert judgment - in this case, that of the pathologist reading the biopsy data.) So when we automate, we are automating that *human* expertise.

To build such an algorithm, typically we would collect data on (say) x-rays, along with radiologist judgments of the patient's conditions for those x-rays. We would then use those data to build a predictor of what a radiologist would say about any given x-ray. Using data never before seen by the algorithm, we could then validate the algorithm's predictions; that is, how do its predictions on new x-rays compare with radiologist's statements about these x-rays?

Thinking about inversion here again forces us to think about our goal in building this algorithm. Is our goal to predict what a radiologist will say? On first blush, of course. But on second blush we can see the answer is of course not. Our goal is to understand what the radiologist's expert opinion would be on this x-ray. Those two things – what a radiologist says on any given x-ray, versus their expert judgment of that x-ray – seem deceptively similar. Those untrained in psychology may not even see the distinction.

Yet psychologists have long recognized that in any given instance, an expert may behave in ways that are at variance with their expertise; see for example Shanteau (1992), Payne et al. (1993) and Berthet (2022). (There are also separate literatures on how actuarial models outperform experts, e.g. Dawes, Faust and Meehl, 1989 and Kleinberg et al. 2018, and on how far automation may go and the implications for the future workforce, e.g. Brynjolfsson and Mitchell 2017).

For example, fatigue affects judgments. Studies suggest that radiologist errors may be as much as 25% more common at the end of a doctor's shift than at the beginning (Taylor-Phillips and Stinton, 2019). Fatigue is so intuitive that it seems almost obvious - painfully obvious. Yet to our knowledge none of the algorithms that automate medical judgments account for fatigue. In fact, as far as we know they typically don't even collect the key piece of data we would need to address fatigue (something like "how many hours had the radiologist worked before seeing this x-ray?"). The difference between specific judgments and expertise is not limited to fatigue. In coming up with a diagnosis the radiologist may ignore the base rate of different conditions. Or radiologists may be prone to the so-called 'gambler's fallacy': After several negative X-rays in a row, the radiologist will be inclined to think that a positive X-ray is 'due' or vice versa (Chen et al. 2016).

An algorithm that simply predicts the radiologist's behavior will automate all of these biases alongside the variable we really want, which is the radiologist's true expertise. When we automate what radiologists say - or teachers, college admissions officers, resume screeners, or judges - we are neglecting important inversion problems.

3. The problem of treating inversion problems as pure prediction problems

Confusing inversion for pure prediction problems can lead algorithms to generate social consequences that are quite different from what users really want.

3.1 Curation

To see the real-world consequences of treating inversion problems like pure prediction problems, consider an audit study of one of the world's most widely-used social media platforms: Facebook (Agan, Davenport, Ludwig and Mullainathan, 2022). Two of the most widely-used curation algorithms on Facebook are People You May Know (PYMK), which ranks potential new 'friends' by a prediction of your likelihood of connecting with them, and Newsfeed, which ranks posts by people you've already friended, based on the algorithm's prediction of whether you would engage with a given post.

The Facebook audit reveals a striking fact: NewsFeed recommendations are subject to out-group bias. That is, NewsFeed systematically "hides" (downranks) posts by friends that are of a different ethnicity or race: In the US, for a white user, their Black friends' posts are ranked systematically lower (and vice versa). A second audit in India finds the same pattern: for a Hindu user, posts by Muslim friends are much lower in the feed. Since users often only read the first few posts this means that even once a user has befriended someone from a different group, the algorithm implicitly works to weaken the tie to the 'out-group' friend.

Importantly, these patterns hold true even controlling for how much the user is interested in the content that their friends have posted. In the Agan et al. study, the authors also were able to survey study subjects about their explicit, deliberate preferences about how much they'd like to see different Newsfeed posts. People's deliberate, stated preferences seem to value the posts of Black versus white friends (or Muslim versus Hindu) friends equally well. But somehow the algorithm, built using people's past choices, is giving them something different.

The psychological model of automaticity gives us an explanation for why Newsfeed shows significant ingroup bias. How do we know its automaticity that is contributing to this bias? Partly because we don't see the same pattern with the PYMK algorithm, where the data show us users are acting much more deliberately when making their choices, and partly because in a laboratory experiment that presents users with a stylized recommender system, adding time pressure to make subjects behavior more automatically seems to increase bias.

Ignoring the role of automaticity in people's past NewsFeed choices - that is, treating the construction of the Newsfeed algorithm as a pure prediction problem rather than an inversion problem – leads to a large-scale algorithm that may impede social interactions across lines of race or religion in two of the largest social media markets in the world.

This problem is of course not limited to Facebook; it is endemic to curation algorithms.

Consider just one other example: The algorithms that help people surf the web try to learn our preferences from our past choices, assuming those are one and the same. But that ignores the psychological insight that all of us tend to under-invest in exploring new things; our past choices will under-state the degree to which we like variety rather than just familiarity.[4] This provides an explanation for the following blogger's post, which captures something surely most of us feel to some degree: "Obviously this isn't objectively true, the internet is always adding content … but it doesn't really give me that open, untamed frontier feeling that it gave me when I was younger. These days, it feels like 99% of my internet usage is confined to a handful of the same websites

[4] See Guo and Yu (2020) for a study of how people under-explore in bandit problems.

that I never really venture away from."[5] The algorithm that mistakenly thinks we really only crave the familiar by looking at our past choices winds up making the Internet feel small.

3.2 Automating expertise

While we do not have the same sort of direct evidence of the real-world consequences of confusing inversion problems for pure prediction problems in the case of automating expertise, there is every reason to believe the consequences here may be substantial as well.

When we try to automate expert judgment, from the expert's behavior the algorithm will learn not just their expertise but also their biases. One study asked radiologists to review a set of x-rays and then on the last x-ray, inserted a superimposed image of a gorilla, an homage to a classic study in psychology showing how attention gets allocated automatically and the limits of the human attentional spotlight (see Drew, Vo and Wolfe, 2013). A shockingly large share of radiologists did not see the gorilla at all. Those are the types of choices that are getting incorporated and learned by the algorithm behind the automated radiologist.

A different hypothetical example makes the same point more vividly (and visually): An algorithm trained to automatically call balls and strikes by watching baseball umpires. The result is shown in Figure 1 (taken from Walsh 2010). The actual strike zone is the rectangle. Yet umpires—and hence a hypothetical robot umpire that was trained on their calls rather than on the formal rectangular boundaries—end up with a called strike zone that is shorter and wider, an irregular oval rather than the rectangle that the rulebook says it should be.

One empirical example that we do have for the real-world consequences of confusing inversion for pure prediction problems comes from medicine. An algorithm trained to detect knee pain is usually trained not to predict knee pain, but to predict instead a clinician's judgment of knee problems. A psychologist would look at this and ask, "Don't doctors - like all people - sometimes suffer from implicit biases?" Those implicit biases will in turn get baked into an algorithm that is built using data from the clinician's past choices. Pierson et al. (2021) show the consequences - clinicians are more likely to fail to recognize physical problems in the knees of Black and low-income patients.

By treating inversion problems as if they are pure prediction problems, we are inadvertently building algorithms that do not achieve the goals we set out for them - and sometimes accomplish the opposite of what we really want.

4. Building a Science of Inversion

[5] While computer scientists have recognized the general problem that algorithms can inadvertently show people a narrower set of options than they'd normally choose, and so try to ensure the recommendations are as diverse as people's actual choices (known as 'calibrated recommendations' see Steck, 2018), even that technical fix misses the psychological insight that the user's own initial choices under-value diversity. A recommender system would need to override user preferences and give the user something they appear not to want (in behavior). See, e.g., Garcia-Gathright et al. (2018) for an example of how platforms might help users discover new content through a holistic understanding of user satisfaction.

Solving inversion problems will require combining the best of both cultures - psychological models (of a sort that are even more specific and explicit than is currently common within psychology) and machine learning to ‘personalize’ the inversion specific to each person or case. Development of these new methods won’t be possible without having psychology and psychologists at the center of the effort.

We do not have - to our knowledge - formal techniques that combine the best of psychology and the best of machine learning to solve inversion problems. There is reason, however, to believe these two methods can be combined, and combined in a way that yields practical implications.

In what follows we offer some insights or principles about inversion that might help guide the development of these new methods.

4.1 Having multiple behavioral measures helps triangulate

How would we ever invert in practice? In our simple examples so far we have assumed that there is just a single behavioral measure in the data. But in reality there are often different types of behavioral measures available in the data. That is a boon for inversion since psychological theory is quick to point out that not all behaviors are equally automatic.

Consider curation algorithms. Most social media platforms have a variety of engagement measures - clicking may be more automatic, while reading an article all the way through, or even commenting on it, may be more deliberate. Moreover psychology has even produced validated measures of automaticity that can be used to quantitatively measure the relative degree of automaticity of different engagement measures that might be captured in available data.[6] These two kinds of content (more versus less automatic) will behave differently on the different types of behavioral measures we have available. That is, for the sort of content more prone to automaticity, we will see greater divergence between behaviors that tend to be more versus less automatic, while for other content we will see less divergence. That degree of divergence can be a roadmap to the regions where inversion is needed more than pure prediction.

It’s very possible that solving the inversion problems relevant to those segments of the data may require formal structural models of psychological phenomena of a sort that currently don’t exist. While we could in principle simply put more weight on those behavioral engagement measures that are relatively less automatic, that leaves us short of the target: “less prone to bias” is better than “more prone to bias” but not equal to “unbiased.” More sophisticated variants are also possible of this basic idea. For example, the more automatic measures are likely to be the ones that are available in higher frequency and volume than the less automatic measures (since things people do relatively more of are, all else equal, likely to be things they do relatively more quickly). Some weighting scheme could be used to optimally weight together the higher-frequency biased engagement measures with the lower-frequency, less-biased but noisier engagement measures. The key point is that a formal psychological model is needed to extrapolate beyond the range of automaticity captured in the data.

[6] Milli, Belli and Hardt (2021) describe computational procedures to use once we have identified which behavioral measures contain more information about actual user preference (‘value’ in their terminology). The behavioral science helps us do this kind of identification.

That is, the variation in automaticity of the engagement measures captured in the data can be used to learn the relationship between level of automaticity and out-group bias in the support of the data. But data alone can't let us say anything beyond the support of the data; we need some additional information or assumptions to do that - i.e., a model. Specifically we need some explicit mathematical model of automaticity that makes some assumptions about the nature of automaticity and out-group bias based on psychological theory. The key question is one of functional form: Does the gradient between automaticity and out-group bias flatten or steepen as we move beyond the support of the data towards lower and lower levels of automaticity?

This need for quantification in the solving of inversion problems is what requires an extra degree of precision and formality that seems to remain fairly rare with psychological models. The field currently works largely in directional terms: "Automaticity makes people relatively more prone to rely on stereotypes." But for solving inversion problems we need to understand something about *magnitudes*, not just directional effects, to be incorporated into the design of training algorithms. That requires formalization of psychological ideas in terms of mathematical models, akin to what we have started to see as part of the development of behavioral economics. For example, time inconsistency had been a phenomenon that psychologists had studied and documented for decades (I would prefer $5 today over $10 tomorrow, but I would prefer $10 next Sunday rather than get $5 the day before on Saturday). It was the development of formal models of quasi-hyperbolic discount functions (see for example Laibson, 1997) that allowed economists (and in principle psychologists, for that matter) to empirically estimate the value for the key beta and delta parameters of the hyperbolic discount function. Those parameter values help behavioral economists determine, for example, which specific commitment devices or tax policies or financial market innovations make consumers better versus worse off. Much more work along those lines remains to be done, though, since behavioral economics has incorporated and formalized just a vanishingly small share of all the psychological insights that are likely to be important for the development of trainer algorithms.

4.2 Inverting on small datasets can help build algorithms on larger datasets

The main challenge of treating inversion problems for pure prediction problems is that the algorithm builder has no way to judge whether any given candidate algorithm is doing what we want, because the outcome being examined, behavior, does not correspond to the thing we really want to know, mental state. Sometimes inversion may involve costly solutions - like the collection of additional data - to learn the mapping between behavior and mental state. The good news is that learning that mapping for a small dataset can sometimes be useful for constructing industrial-scale algorithms on much larger datasets.

On social media, for instance, one way to learn what people really want may be to ask them in ways that try to elicit the preferences of their more deliberate selves. Such explicit data collection cannot reach the scale of passive collection - surveys are more expensive than simply tracking clicks. Still, they can be done: social media companies regularly survey user satisfaction. Such survey data can be immensely helpful for learning the structure of inversion. Behavioral science can help us craft what to survey; and these data can be used (albeit in a small sample) to learn the mapping between behavior and desired mental state; in essence, it can help us learn the best way to aggregate large-scale data to best proxy for mental state.

Notice why it is not necessary to collect preference or satisfaction data for literally every user, site, post and product that will be ‘touched’ by the at-scale algorithm: Because algorithms are essentially all about grouping things together by observable characteristics. So long as we have ‘enough’ preference or satisfaction data for the relevant user-site-post-product ‘cells,’ defined by observable characteristics of users, sites, posts and products, we can extrapolate to other users-sites-posts-products with similar characteristics (subject to the usual assumption that the data generating process out in the world that relates user preferences to those characteristics is stable). Learning from a sample of data that 50-year-old smart pantry users deeply regret eating more than a handful of Doritos a week while 18-year-old high-school athletes have less ambiguous feelings is a useful if-then rule that can be applied more broadly to guide the construction of larger-scale algorithms.

That same logic can highlight where small amounts of additional data collected at large scale can be of disproportionate utility. For example in the Facebook case, automaticity manifests itself in bias against Newsfeed posts by outgroup friends. We know that because an audit study collected costly user preference data at small scale (as in Agan et al., 2022). But with that information at hand we now know what the key variable is to make sure we collect at large scale as part of any attempt to debias the Newsfeed algorithm: the race (or religiosity, etc.) of the friend posting.

4.3 One need not be able to invert perfectly

Perhaps the only thing we are willing to confidently predict about the new science of inversion, whatever that winds up looking like, is that it will not be perfect. There will surely be many applications where it will not be possible to confidently invert the mental state of interest from the behavior observed in data. Does that render the idea of inversion irrelevant? Our answer is: So what. Inversion can still be enormously useful.

There are past success stories that illustrate this principle for online content. For example, an early insight in the ranking of search results was the notion of "position bias" -- that people are more likely to click on search results that are higher up in a ranking, even if other results lower down might be better responses to their query. Position bias arises naturally from psychological insights into how people browse for information --- that they may engage in satisficing behavior and stop early --- and it was measured through eye-tracking studies (Joachims et al 2017), leading to improved ranking algorithms that take into account the possibility that highly-ranked items might receive more clicks even when they're worse. In this case, an inversion strategy created important improvements even though we are still far from inverting all aspects of a searcher's mental state.

Considering other applications prospectively, think again for example of curation algorithms for social media. Some content (posts, tweets, etc.) is likely to lead to more automatic behavior than others. Some content causes us to pause and be more considered; other content simply provokes automatic responses. For many current applications in content curation, each piece of content is essentially treated the same from a psychological perspective - as equally reflective of the user’s true preferences. But learning that some type of content might be engaging to people perhaps because they really like it, but perhaps because it’s a temptation good, is enormously valuable. The smart pantry that gives you all the fruits and vegetables you want but says “Hmm, let’s be open to multiple possible motivations here with the Doritos” has just done you a great favor.

Put differently, data scientists and behavioral scientists alike are used to living with the idea of a statistical uncertainty interval around our estimates. Inversion, even when it cannot ‘point identify’ our mental states, can still be useful by giving us something analogous to a psychological uncertainty interval.

4.4 Inversion is not the same as having multiple objectives

Data scientists have long been used to solving problems with multiple objectives: We want an algorithm that for example recommends the shortest driving route, but also tries to minimize fuel consumption as well to help address climate change. That might sound like inversion. It’s not.

Inversion problems are those where the thing we really want to know - whether that’s a single objective, or multiple objectives - are not represented in our data, and there’s no transparent way to go from here to there using the available behavioral data alone. Machine learning engineering tools to design algorithms that balance competing goals against one another can’t solve the problem that the goals that we really care about aren’t measured in data. Multiple-objective optimization algorithms are no substitute for the development of new models that combine psychological insights with machine learning.

The one common thread across all of these objections is that algorithms - and everyone who relies on algorithms (which is to say, everyone) - *needs* psychologists to be centrally involved in the solving of inversion problems.

5. Conclusion

Our core argument is that many algorithms built using data on human behavior treat inversion problems as if they were pure prediction problems. Ignoring mental state and focusing solely on behavior can lead us astray. For example, it can result in algorithms that are intended to make people better off but often may inadvertently make people worse off instead.

Inversion problems are ubiquitous. The sheer volume of cognitive biases and heuristics identified in the psychology literature raises the question of how often behavior actually transparently represents mental state. In coming up with examples for this essay we struggled to find such cases. In contrast, it was trivially easy to come up with countless examples where some psychological insight made clear that behavior is very clearly *not* what we’d really want to predict. It may not be an exaggeration to think that inversion problems may be more common than pure prediction problems - perhaps much more common.

Solving inversion problems will require new tools. The standard playbook of building training algorithms using data on people’s past behaviors and choices is by itself not sufficient. We must augment it with a new set of methods, most not yet developed, that requires psychological insights (of which we luckily have no shortage), formalized mathematical models based on those insights (mostly lacking), and some way of incorporating those psychological models into machine learning to solve inversion problems on a personalized basis (entirely lacking and so would need to be developed).

The payoff to psychologists being involved in this activity is not just the real-world impact that comes from building much more socially helpful algorithms. The payoff also comes from the ability to expand psychological theory itself, since algorithms can themselves be used for scientific discovery and theory development (Ludwig and Mullainathan, forthcoming; Mullainathan and Rambachan, 2023). In the same way that behavioral economics has transformed the fields of both economics and psychology, a new field of 'behavioral computation' could transform both psychology and computer science.

References

Agan, A., Davenport, D., Ludwig, J., & Mullainathan, S. (2022). Automating automaticity: How the context of human choice affects the extent of algorithmic bias. Cambridge, MA: National Bureau of Economic Research Working Paper.

Ainslie GW. (1992) Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person. Cambridge University Press; Cambridge: 1992.

Allcott, H., Gentzkow, M., & Song, L. (2022). Digital addiction. *American Economic Review*, *112*(7), 2424-63.

Bargh, John A. "The four horsemen of automaticity: Intention, awareness, efficiency, and control as separate issues." (1994).

Barocas, Solon and Andrew D. Selbst (2016) “Big data’s disparate impact.” *California Law Review*. 671.

Batia S., and Aka A. (2022) “Cognitive modeling with representations from large-scale digital data.” *Current Directions in Psychological Science*. 31(3): 207-214.

Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The Strength Model of Self-Control. *Current Directions in Psychological Science*, 16(5), 351–355. https://doi.org/10.1111/j.1467-8721.2007.00534.x

Bayer, J. B., Triệu, P., & Ellison, N. B. (2020). Social media elements, ecologies, and effects. *Annual Review of Psychology*, *71*, 471-497.

Berthet, Vincent (2022) The impact of cognitive biases on professionals’ decision-making: A review of four occupational areas. *Frontiers in Psychology*. 12 https://www.frontiersin.org/articles/10.3389/fpsyg.2021.802439

Breiman, Leo (2001) Statistical modeling: The two cultures. *Statistical Science*. 16(3): 199-231.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530-1534.

Card, Stuart K., Thomas P. Moran, and Allen Newell. *The psychology of human-computer interaction*. Crc Press, 1983.

Card, Stuart K. *The psychology of human-computer interaction*. Crc Press, 2018.

Carroll, John M. "Human-computer interaction: psychology as a science of design." *Annual review of psychology* 48, no. 1 (1997): 61-83.

Chaiken, Shelly, and Yaacov Trope, eds. *Dual-process theories in social psychology*. Guilford Press, 1999.

Chan, L., Critch, A., & Dragan, A. (2021). Human irrationality: both bad and good for reward inference. *arXiv preprint arXiv:2111.06956*.

Chen, D. L., Moskowitz, T. J., & Shue, K. (2016). Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *Quarterly Journal of Economics*, 131(3), 1181–1242. https://doi.org/10.1093/qje/qjw017

Chouldechova, Alexandra (2016) "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." FATML 2016 conference paper.

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science*, 15(12), 806–813.https://doi.org/10.1111/j.0956-7976.2004.00760.x

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668-1674.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psycholog*y, 56(1), 5–18. https://doi.org/10.1037/0022-3514.56.1.5

Dix, Alan, Janet Finlay, Gregory D. Abowd, and Russell Beale. *Human-computer interaction*. Pearson Education, 2003.

Drew, Trafton, Melissa LH Vo, and Jeremy M Wolfe (2013) "The invisible gorilla strikes again: Sustained inattentional blindness in expert observers." *Psychological Science*. 24(9): 1848-1853.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226)

The Executive Office of the President. (2016, May). Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

Fujita, K., Trope, Y., Liberman, N., & Levin-Sagi, M. (2006). Construal levels and self-control. *Journal of Personality and Social Psychology*, 90(3), 351–367.

Garcia-Gathright, J., St. Thomas, B., Hosey, C., Nazari, Z., & Diaz, F. (2018, June). Understanding and evaluating user satisfaction with music discovery. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 55-64).

Gawronski, B., & Creighton, L. A. (2013). Dual process theories. In D. E. Carlston (Ed.), *The Oxford handbook of social cognition* (pp. 282–312). Oxford University Press.

Gebru, T. (2020). Race and gender. In M. D. Dubber (Ed.), *The Oxford handbook of ethics of AI* (pp. 251-269). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.16

Gigerenzer, Gerd and Wolfgang Gaissmaier (2011) "Heuristic decision making." *Annual Review of Psychology*. 62: 451-482.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, 60(4), 509–517. https://doi.org/10.1037/0022-3514.60.4.509

Green, E., & Daniels, D. P. (2014). What does it take to call a strike? Three biases in umpire decision making. In *2014 MIT Sloan Sports Analytics Conference Proceedings,* Boston, MA

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Guo, D., & Yu, A. J. (2019). Human learning and decision-making in the bandit task: Three wrongs make a right. In *Conference on cognitive computational neuroscience.*

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Hartson, Rex, and Pardha S. Pyla. *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier, 2012.

Hassenzahl, Marc, and Noam Tractinsky. "User experience-a research agenda." *Behaviour & information technology* 25, no. 2 (2006): 91-97.

Hassenzahl, Marc, Sarah Diefenbach, and Anja Göritz. "Needs, affect, and interactive products–Facets of user experience." *Interacting with computers* 22, no. 5 (2010): 353-362.

Helander, Martin G., ed. *Handbook of human-computer interaction*. Elsevier, 2014.

Hunt, M. G., Marx, R., Lipson, C., & Young, J. (2018). No more FOMO: Limiting social media decreases loneliness and depression. *Journal of Social and Clinical Psychology*, *37*(10), 751-768

Inzlicht, M., Werner, K. M., Briskin, J. L., & Roberts, B. W. (2021). Integrating models of self-regulation. *Annual review of psychology*, 72, 319-345.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Joachims, T., Swaminathan, A., & Schnabel, T. (2017). Unbiased Learning-to-Rank with Biased Feedback. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17). Association for Computing Machinery, New York, NY, USA, 781–789. https://doi.org/10.1145/3018661.3018699

Johnson, Jeff. *Designing with the mind in mind: simple guide to understanding user interface design guidelines*. Morgan Kaufmann, 2020.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., & Frederick, S. (2005). A Model of Heuristic Judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). Cambridge University Press.

Kahneman, Daniel and Gary Klein (2009) "Conditions for intuitive expertise: A failure to disagree." *American Psychologist*. 64(6): 515-526.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237-293.

Kleinberg, Jon, Sendhil Mullainathan and Manish Raghavan (2017) "Inherent trade-offs in the fair determination of risk scores." *Proceedings of Innovations in Theoretical Computer Science*.

Kleinberg, J. M., Mullainathan, S., & Raghavan, M. (2022). The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *Proceedings of the 23rd ACM Conference on Economics and Computation*.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2), 443–478. https://doi.org/10.1162/003355397555253

Lazar, Jonathan, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, 65(3), 272–292. https://doi.org/10.1006/obhd.1996.0028

Loewenstein G, Prelec D. (1992) Anomalies in intertemporal choice: Evidence and an interpretation. *Quarterly Journal of Economics*. 107:573–597.

Ludwig, J., & Mullainathan, S. (2022). Machine Learning as a Tool for Hypothesis Generation. [National Bureau of Economic Research Working Paper].

Lum, Kristian. (2017) Limitations of mitigating judicial bias with machine learning. *Nature Human Behavior,* 1, 0141.

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, 106(1), 3–19. https://doi.org/10.1037/0033-295X.106.1.3

Milkman, K. L., Rogers, T., & Bazerman, M. H. (2009). Highbrow films gather dust: Time-inconsistent preferences and online DVD rentals. *Management Science*, 55(6), 1047–1059. https://doi.org/10.1287/mnsc.1080.0994

Milli, Smitha, Luca Belli, and Moritz Hardt. (2021) From optimizing engagement to measuring value. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Mullainathan, Sendhil and Rambachan, Ashesh, From Predictive Algorithms to Automatic Generation of Anomalies (May 9, 2023). http://dx.doi.org/10.2139/ssrn.4443738

Muraven, M., & Slessareva, E. (2003). Mechanisms of selfcControl failure: Motivation and limited resources. *Personality and Social Psychology Bulletin*, 29(7), 894–906. https://doi.org/10.1177/0146167203029007008

Neubaum, G., & Krämer, N. C. (2017). Monitoring the opinion of the crowd: Psychological mechanisms underlying public opinion perceptions on social media. *Media psychology*, *20*(3), 502-531.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*(4), 250–256.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.

O’Donoghue, T., & Rabin, M. (2000). The economics of immediate gratification. *Journal of Behavioral Decision Making*, 13(2), 233–250. https://doi.org/10.1002/(SICI)1099-0771(200004/06)13:2<233::AID-BDM325>3.0.CO;2-U

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). The adaptive decision maker. Cambridge University Press. https://doi.org/10.1017/CBO9781139173933

Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S., & Obermeyer, Z. (2021). An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1), 136-140.

Preece, Jenny, Yvonne Rogers, Helen Sharp, David Benyon, Simon Holland, and Tom Carey. *Human-computer interaction*. Addison-Wesley Longman Ltd., 1994.

Read, D., & van Leeuwen, B. (1998). Predicting hunger: The effects of appetite and delay on choice. *Organizational Behavior and Human Decision Processes*, 76(2), 189–205. https://doi.org/10.1006/obhd.1998.2803

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica, 5*(17), 61-71.

Shanteau, J. (1992). The psychology of experts an alternative view. In G. Wright & F. Bolger (Eds.), *Expertise and decision support* (pp. 11-23). Springer. https://doi.org/10.1007/b102410

Shneiderman, B (1986). *Designing the user interface: Strategies for effective human-computer interaction*. Addison-Wesley.

Snow, CP (1959) *The Two Cultures and the Scientific Revolution*. Cambridge University Press.

Steck, H. (2018). Calibrated recommendations. *Proceedings of the 12th ACM Conference on Recommender Systems*, 154–162.

Tarantola, Albert (2005) Inverse Problem Theory and Methods for Model Parameter Estimation. Society for Industrial and Applied Mathematics.

Taylor-Phillips, Sian and Chris Stinton (2019) “Fatigue in radiology: A fertile area for future research.” *British Journal of Radiology*. 92(1099).

Thaler, R. H., & Shefrin, H. (1981). An economic theory of self-control. *Journal of Political Economy*, 82(2), 392–406. https://doi.org/10.1086/260971

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Tromholt, Morten (2016) “The Facebook experiment: Quitting Facebook leads to higher levels of well-being.” *Cyberpsychology, behavior, and social networking*. 19(11): 661-666.

Trueblood, J.S., Holmes, W.R., Seegmiller, A.C. *et al.* The impact of speed and bias on the cognitive processes of experts and novices in medical image decision-making. *Cogn. Research* 3, 28 (2018). https://doi.org/10.1186/s41235-018-0119-2

Trueblood, J.S. Eichbaum Q., Seegmiller A.C., Stratton C., O’Daniels P., and Holmes W.R. (2021) “Disentangling prevalence induced biases in medical image decision-making.” *Cognition*. 212: 104713.

Vogel, E. A., Rose, J. P., Roberts, L. R., & Eckles, K. (2014). Social comparison, social media, and self-esteem. *Psychology of popular media culture*, *3*(4), 206

Walsh, John (2010) “The compassionate umpire.” *Hardball Times*. https://tht.fangraphs.com/the-compassionate-umpire/

Ward, A., & Mann, T. (2022). Control yourself: Broad implications of narrowed attention. *Perspectives on Psychological Science*, 17456916221077092. https://doi.org/10.1177/17456916221077093

Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious.* Belknap Press/Harvard University Press.

Figure Captions

Figure 1: Umpire-called balls and strikes (shown by heat map), relative to official strike zone according to the rule book (white rectangle). Source: Walsh (2010)