Algorithms as A Vehicle to Reflective Equilibrium: Behavioral Economics 2.0

Jens Ludwig

Sendhil Mullainathan

Sophia L. Pink

Ashesh Rambachan[1]

October 6, 2025

Abstract

Behavioral economics has struggled to simultaneously accommodate two facts: (i) people make mistakes, even on very consequential decisions; and (ii) different people have different preferences. Doing nothing respects revealed preference, but does little to address mistakes. Nudging addresses mistakes but minimizes individual differences, requiring assumptions about what behavior is good for all people. We argue AI will reshape behavioral economics by providing a solution to this dilemma. Algorithms can serve as thought partners that help people get closer to "reflective equilibrium": what they would choose if they were to exercise their most considered judgment on all information available at the time. Behavioral Economics 2.0 ought to be the design of such algorithms that improve decisions while preserving agency.

[1] Ludwig: University of Chicago and NBER, jludwig@uchicago.edu; Mullainathan: MIT and NBER sendhil@mit.edu; Sophia Pink, University of Pennsylvania,: sopink@wharton.upenn.edu; Rambachan: MIT, asheshr@mit.edu. Thanks to Ajay Agrawal, Abhishek Nagaraj, Cass Sunstein and seminar participants at the University of Melbourne and the NBER conference on transformative AI at Stanford University for helpful comments. Any errors and all opinions are of course our own.

1. Introduction

People tend to avoid acknowledging a problem until there is a solution. Behavioral economics has a barely acknowledged problem,[2] for which there is now a solution. It is a criticism that was leveled by its earliest critics, very often very unfairly. Still, it contains more than a grain of truth.

Take the classic example from behavioral economics: retirement savings. People clearly save too little. Just as clearly, but surprisingly, a simple intervention can solve this perennial problem. Change the default – "check a box to sign up for a retirement savings plan" becomes "check a box *not* to sign up" – and savings rise. Beautifully dubbed "nudges," interventions such as these are appealing because they largely preserve personal freedom and yet are still effective in changing behavior, often in excess of more expensive interventions, e.g. subsidies for savings (Thaler and Sunstein 2008).[3]

The lingering problem is this: Yes, savings were too low. Yes, savings have gone up. But are people actually better off? The answer is more subtle than it appears. Consider where the increased savings came from – after all, everyone faces a budget constraint. Suppose people simply consume less. Suppose they put less money into a college savings account. Or suppose

[2] There are important exceptions; see for example Bernheim and Rangel (2007), Bernheim (2025), Sunstein (2024), and Sunstein (2025).

[3] Behavioral economics is obviously about much more than changing the choices of private individuals, but also, for example, about informing how we design and implement the law (Sunstein et al. 1998, Sunstein 2000) or design regulations (Sunstein 2013) or other public policies (Sunstein 2020).

they take on more high-interest credit card debt. Surely, our judgments of whether people are better off clearly differ in each of these three cases.

The problem is deeper than simply measurement, the tracking of dollars. Take the apparently easiest case: suppose people consume less. Even then, we would need to know what kind of consumption is being cut. What if someone is spending less on travel to go see their rapidly aging parents? Are we sure that higher retirement savings compensates for this loss? Even after we've measured all we can, we are left with something not in the data - how precious is this family time? That depends on preferences, values, their emotional makeup and even something as banal as how connected they can feel via (say) telephone?

Criticisms such as these can seem counter-productive since the implication in its extreme form can be "stop nudging." That cannot be the answer in a world where people frequently and predictably make economically meaningful mistakes. Should we just throw up our hands at low rates of retirement savings, even when data shows people wish they saved more?[4]

[4] A representative survey of Americans ages 60-79 found that over 50% wished they had saved more earlier in life (Börsch-Supan et al. 2023). Moreover, as Thaler and Sunstein (2008) point out, there is no neutral choice architecture – every choice is a nudge in some sense. That reinforces our argument: how are we to know which choice architecture is the right one?

What we need is a constructive path forward that acknowledges two truths: (1) people make mistakes, and (2) people have their own preferences. We need interventions that help people, but also respect their agency - that recognize people have private knowledge of their own wants.[5]

Algorithms allow us to do exactly that – or rather, algorithms *could* allow us to do that. Nudges move behavior in the direction we think will make people happier. Algorithms, on the other hand, could allow people to make better decisions about what direction will be right for them.

The best example of this distinction comes from another behavioral economics classic, Save More Tomorrow (Thaler and Benartzi 2004). Here, workers are given the chance to talk through their retirement goals with a financial advisor (which, revealingly, almost all workers chose to do). They settled on a savings rate and a way to save that was feasible. Then the person sets their own future defaults: automatically put some share of future pay raises into a retirement account.

Notice how Save More Tomorrow could be delivered in the future: by algorithm. An algorithm could not just make this more scalable but potentially be even better than the human advisor.

[5] Libertarian paternalism is a sophisticated attempt to navigate these two truths: nudges after all can easily be over-ridden. From our perspective, it is an incomplete solution. First, because it assumes there is a single direction of potential error. Yet people who are about to be evicted, or forgo a critical medical procedure for cost reasons, might need to save *less*, not more. Second, because it assumes a form of rationality. If it truly mattered to someone, they would over-ride the nudge. Yet, people may often comply with the nudge mindlessly or in a confused manner, revealing little about what's good for them.

Algorithms are tireless. They can constantly and continuously process every credit card and mortgage offer, monitor all of a person's bills and finances. They can answer all questions, however small, at any time and without getting frustrated.

Algorithms are actually better than humans are at providing certain kinds of advice. Key inputs into financial advice are exactly the kind of thing that predictive algorithms excel at predicting: What is my risk of job loss? How long will I live? What are my health risks?

And they could eventually be better at "understanding" people (or some aspects of people). A great advisor can develop interpersonal savviness from years of experience. A deployed AI system has many orders of magnitude more data to draw on. Developing algorithms that can effectively learn from that data is no small task, but is not out of reach.

Save More Tomorrow was ahead of its time. It was an idea tailor-made for algorithms, which happened to be delivered by people. The difference between defaults and Save More Tomorrow is the difference between behavioral economics and behavioral economics 2.0. One picks a choice to nudge people towards. The other helps people pick the right choice for them.[6] AI will allow many more such interventions, ones in which algorithms serve as *thought partners*.

---

[6] For example, of the people who met with the financial consultant in the earliest instantiation of Save More Tomorrow: 28% accepted an immediate increase in their savings contribution, 56% chose to start the savings increase when they received their next pay raise, and the remaining 16% chose not to increase their contributions.

What should the goal of a thought partner be? How to separate a bad algorithm from a good one? We draw on an insight by Rawls (1971) in a very different context. Focused on moral judgment, he argued people's choices should reflect their most "considered judgment," or what he called "*reflective equilibrium*." We extend that to a broader set of judgments and suggest a goal: choices should approximate what people would have chosen if they had carefully processed all the information available at the time. That is, the goal of algorithms should be to guide people towards reflective equilibrium as we mean the term.[7] Behavioral economics 2.0 should be the field that builds and tests such algorithms.[8]

Such algorithms will not only improve choices, they can change the logic of market equilibrium. Firms exploit consumer biases. Yet the current remedies are static (e.g., disclosure). This creates a cat and mouse game between firms and regulators, except the mouse is faster and more nimble: firms can constantly find new ways to frame things or obfuscate. Algorithms can keep up much faster than regulators: just as with a spam filter, it is easier to fool a person than an algorithm. We show how even existing algorithms appear capable of piercing through attempts to obfuscate.

[7] The act (and cognitive effort) of reflection itself is not necessary; only that the choice should approximate the most considered judgement. Given that algorithms could do some (or much!) of the reflection, it is a design choice as to how much reflection is efficient. There is a tradeoff between cognitive economy and the benefits of agency.

[8] Note this is fundamentally different from "precision nudging," which uses machine learning to target nudges. That answers the question "What treatment best nudges them in this direction?" Here, the problem is "what direction is even right for this person?"

Finally, the perspective of reflective equilibrium encourages us to think more broadly. Why focus on individual decisions? Why not ask if we can de-bias people, and improve their own thinking and decision-making capacities*?* Debiasing interventions have largely been ignored (though there are exceptions). We argue that is because we previously lacked the ability to deliver them in ways that work; and that in behavioral economics 2.0 we will be able to.

To fully realize this vision will require a number of scientific advances. For example, we need a more empirically grounded way to measure reflective equilibrium. We need behavioral theories that are more precise and so more "algorithm friendly." We need algorithms that can better listen to people, and can better understand and apply psychology. What these advances have in common is that they all require integrating computer science and behavioral economics.

These efforts, we argue, are worth it.

The radicalness of the change is seen in the most subtle of places: the grammar of the field. Interventions are done *to* people. People do things *with* tools. Nudges are interventions. Algorithms are tools. Behavioral economics will go from a field that designs interventions to one that builds tools.

2. Conceptual Framework

Consider a person who must make decision D – how much to invest in a 401(k) retirement plan, whether a judge should release or detain a defendant, whether a doctor should refer a patient for further testing, whether a teen wants to escalate an argument, etc. The utility from the decision is

$U(D|X)$, which depends on their choice and also on other features (e.g. their personal preferences, other circumstances). We write the utility maximizing choice as $D^*(X)$. This is the decision that people would make if they had access to all information feasibly accessible at the time, fully understood the context, and had thought through their choice carefully.

In contrast, we write $D(X)$ to be the choice people actually make. Since people can make mistakes, this produces an error term:

$$\varepsilon(X) = D(X) - D^*(X)$$

Different approaches to human decision-making boil down to different assumptions about $\varepsilon(X)$. Specifically:

- *Revealed preferences*: When we argue for full autonomy, we are often arguing that $\varepsilon(X) = 0$ or close enough that we should not intervene.
- *Nudging*: When we nudge, we believe that the sign($\varepsilon(X)$) is the same for everybody. That is, people do too little or too much of any one decision. A softer assumption would be that on average, across people, the sign of epsilon is known and large; we then live with the fact some people are worse off.
- *Recommender systems*: Traditional recommender systems give people the options that people like them would have chosen. That implicitly assumes that, once we account for features of the individual and the choice, the expected value of ε (X) = 0. People choose correctly, possibly noisily, and so we can predict what a person chooses $\widehat{D(X)}$.

None of these approaches accommodate both individual differences - that $D^*(X)$ varies between people and contexts; and that $\varepsilon(X)$ is not zero. When we allow for error, we either ignore individual preferences (e.g. nudges) or make the error quite trivial (e.g. the noise in recommender systems).

2.1. Reflective equilibrium

Let us assume that $\varepsilon(X)$ is non-zero and can have complex structure that varies across contexts and people. In this case, the goal is to produce decisions so that $\varepsilon(X)$ is as small as possible; or to get $D(X)$ towards $D^*(X)$. That ambition matches an old idea from Rawls (1971). In the context of moral judgment, he argued people's judgments might be prone to what he called "irregularities and distortions" (p. 48). That we should aspire to have people choose on the basis of their "considered judgment," leading to what he termed a "*reflective equilibrium*." A similar aspiration applies to everyday decisions (Sunstein 2022).

It is worth emphasizing what "reflective equilibrium" is and is not.

Reflective equilibrium is the best decision that a person can make in the moment given the information they could feasibly know at the time - that is the conceptual target at which behavioral interventions should be aiming. Reflective equilibrium is *not* omniscience; a decision made at a particular point in time is not compromised if new events change the pros and cons of some action, or if the person's preferences wind up changing based on life experiences.

Reflective equilibrium is about the outcome of a decision, not about the process through which the decision is reached. The goal is to reach the equivalent of considered judgment; it need not require extensive consideration. In fact, algorithms create a new ability to "offload" a great deal of mental computation to literal computation.

Why do we think that reflective equilibrium is the right goal? First, preferences differ across people. We have made this point before, but it is worth reiterating. Even in some seemingly obvious cases – "more retirement savings is good" – it is not obvious that for this person, at this time, the right choice is through the particular employer that is offering the savings plan. For example, they may have a spouse whose employer matches savings; they may already have a well-funded retirement account. While retirement savings programs could add questions about the situations above to a checklist, other behaviors are not as clear cut. Going to the gym more might be good on average but maybe not if it comes at the opportunity cost of other forms of exercise that create more social connections to people, or at the expense of helping a family member or friend in crisis. Less screen time is good on average but not if that screen time would involve making sure you are receiving all the means-tested transfers you're eligible for.

Second, even if we are clear about the overall outcome, it's not clear how to achieve it. Let's say that the best choice for a person is to increase savings. Should they start now? Should they pay off credit card debt first? Notice that in the first implementation of Save More Tomorrow, of the people who chose to meet with the financial consultant, 28% accepted an immediate increase in their savings contribution and 56% committed to starting at their next pay raise.

Third, decisions are interconnected. Whether a particular decision is right for a person or not depends on other decisions. So, even holding constant the actual choice, a person's understanding of that choice has material consequences. Take someone defaulted into a 401(k). The consequences depend on their awareness of that "decision." An unaware person will simply have less at the end of the month and could take actions such as borrowing more, not realizing their income will be lower every month. So, getting someone to save more only makes sense if they have thought through where the money has to come from. It is also worth noting that there may be other benefits. People may respect decisions more if they have put effort into them or have felt part of the process.

The key challenge now is one of operationalizing reflective equilibrium: how could we know if an intervention gets someone closer to their most considered judgment. This is not a new problem. It has always been there. We have simply ignored it. When we nudge, we still have this problem, we simply assert we know what the reflective equilibrium should be. We are making explicit (and confronting) this perennial problem. In Section 6, we argue that reflective equilibrium is operationalizable. For example, one could (in a small, random sample) provide users with other information and suggestions. Those at reflective equilibrium should be less prone to change their choices in response to these other pieces of information.

3. Changing Decisions: Algorithms as Thought Partners

Achieving reflective equilibria requires the ability to provide more help to the decision-maker - like a thought partner. That's exactly what algorithms can provide (Collins et al. 2024).

To see what sort of help would be most useful to people, we return to the non-algorithmic intervention of Save More Tomorrow (Thaler and Benartzi 2004). This intervention not only created a new behaviorally-informed savings vehicle, it gave advice tailored to the individual from an investment consultant. That advice was not always "save more." That consultant had financial planning software that most often calculated the maximum savings allowed according to plan rules. However, the consultant noted:

> *"When the average worker receives this recommendation from the computer program or the 'financial planner,' s/he shuts down and does nothing. So in all cases, after we reviewed their current plan but before I hit the 'Get Advice' button, I would discuss willingness to save with each participant. As you can imagine, the majority of workers live paycheck to paycheck and can barely make ends meet, and they tell you that immediately … If a participant indicated a willingness to immediately increase their deferral [savings] level by more than 5 percent, I hit the 'Get Advice' button. Otherwise, I would constrain the advice proposed to an increase of no more than 5 percent"* (Thaler and Benartzi, 2004, p. S172).

It is also instructive to see how eager the employees were for this thought partnership: out of a total of 315 employees approached, 286 (or 91%) were interested in the help.

Two places that Save More Tomorrow was helpful to people were before and after choosing. Before choosing, the person had help figuring out the consideration set for their savings decision. After choosing, Save More Tomorrow helped them implement that choice - that was the whole genius of getting people to dedicate *future* pay raises to savings.

Algorithms can obviously also help in both regards. Because algorithms can process large amounts of information, they can serve to curate choice sets for us. This can serve to eliminate many of the judgement errors that come from insufficient search or from something not being top of mind. After choosing, they can help implement the choice partly because they can increasingly act in agentic ways, for example by helping fill out forms, etc. A growing literature on AI agents explores these possibilities (see Hadfield and Koh in this volume).

But Save More Tomorrow also highlights how people can get help with the very process of choosing itself. Algorithms could in principle also help with that as well, even though we do relatively little of that right now. In what follows we highlight three ways algorithmic thought partners can help value options. The first two ways are things that behavioral economists already do, but now algorithms can do them real-time and in an adaptive way. The third is something we currently don't help people with much, but that algorithms can do extremely well.

3.1 Algorithms can explain what each option means.

Many poor choices come from confusion about what the choice is. For example, when people are required to appear in court for low-level offenses, they receive a "court summons" with the details. In one analysis of New York City data, fully 47% of the recipients of such forms failed to appear in court as the summons form told them they must (Fishbane et al. 2020). The implicit assumption of the criminal justice system is that people who skip court want to skip court, which is why people are threatened with losing their bail or even an arrest warrant if they don't show.

Behavioral economics suggests an alternative hypothesis: perhaps people didn't understand the complicated form handed to them (see Table 1, column A for an example), so didn't know when they had to appear or what it meant to not appear. To test that hypothesis, Fishbane et al. (2020) redesigned the court summons forms so that the most important information (where and when to show up in court) was easily found near the top of the form. The result of simply simplifying the information for people was to reduce failures to appear in court by 7 percentage points, or about 15% of the base rate. Many people seem to skip court not because of a willful decision but because of confusing decision terms – and clarifying the form had life-changing benefits.

But notice that in an era of algorithmic tools, researchers may not be needed for this. Large Language Models (LLMs) can do this simplification automatically. We ran the original, confusing summons form through an LLM, asking it to explain clearly what this is, give specific instructions for what to do next, and what the consequences are (see Table 1, column B for results). In just a few seconds, an LLM was able to clearly summarize the main points in a way that is much easier to understand.

Table 1: New York City Court Summons

(A) Excerpts from court summons form	**(B) After interpretation by algorithm**
CRC-3206 (5/12) **Complaint/Information** **The People of the State of New York vs.** Name *(Last, First, MI)* Street Address / Apt. No. City / State / Zip Code ID/License Number / State / Type/Class / Expires *(mm/dd/yy)* / Sex Date of Birth *(mm/dd/yy)* / Ht / Wt / Eyes / Hair / Plate/Reg Reg State / Expires *(mm/dd/yy)* / Plate Type / Veh Type / Make / Year / Color **The Person Described Above is Charged as Follows:** Time 24 Hour *(hh:mm)* / Date of Offense *(mm/dd/yy)* / NYC Pink Copy / County Place of Occurrence / Precinct In Violation of Section / Subsection / VTL / Admin Code / Penal Law / Park Rules / Other 23456 **Title of Offense:** **Notice - If you do not follow these instructions a warrant may be ordered for your arrest.** To Plead Not Guilty: You must come to court at the place and time shown on the front of this summons. To Plead Guilty: You must come to court at the place and time shown on the front of this summons. However, if you are charged with **Public Consumption of Alcohol**, or **Public Urination**, you may plead guilty by ma If you plead guilty by mail you do not have to come to court. To Plead Guilty by Mail within 10 days of the date this summons was issued (To one of the above two charges only): * Complete and sign the **"Guilty Plea by Mail Form"** below. * Prepare a check or money order payable to NYC Criminal Court for **$25 for Public Consumption of Alcohol or $50 for Public Urination** * Write the summons number on the check or money order. **DO NOT SEND CASH.** * Mail the check or money order with this completed form to: **NYC CRIMINAL COURT** **P.O. BOX 555** **NEW YORK, N.Y. 10031-0555** **The court record of this case will show your guilty plea to an offense that is not a crime under New York law. This guilty plea will be a public record available at the courthouse. Consult with an attorney if you have questions about this.**	**1. What is this?** A NYC Criminal Court summons for a minor offense. You must either (a) appear in court on the date/time on the front **or** (b) *only if* the charge is **Public Drinking** or **Public Urination**, mail in a guilty plea with a payment instead of going to court. Do not ignore it. **2. What do I do next?** *(Pick ONE path)* **Mail a guilty plea (only for those two charges):** 1. Check the box for your charge in "Guilty Plea by Mail." 2. Sign and date. 3. Pay by check or money order to **"NYC Criminal Court"**: **$25 (Public Drinking)** or **$50 (Public Urination)**; write your summons number on it. 4. Mail form + payment to *NYC Criminal Court, P.O. BOX 555, New York, NY 10031-0555.* Keep a photo and mailing proof. **OR Go to court:** Show up on the appearance date/time (bring summons + ID; arrive early). Ask a lawyer if unsure or worried about job/immigration effects. **3. What deadlines and consequences should I know?** • **10 days:** If mailing plea, must be postmarked within 10 days. Miss it → go to court. • **Court date:** Mandatory if you don't (or can't) mail a proper plea; skipping it can lead to an **arrest warrant.** • **Mailing plea result:** Fixed fine ($25 or $50), no contest later, becomes public record. • **Ignoring everything:** High chance of a warrant and bigger problems.

3.2 Algorithms can fill in informational blind spots.

Another way algorithms can help people understand the value of the options in front of them is through adding in external information – particularly by making *predictions*. To make decisions, people take information about themselves, their choices, and their contexts, and then make predictions, which serves as key inputs to their choice process. When applying for a job, candidates forecast the probability they'll enjoy the job. When buying a house, people forecast the odds they'll get laid off and not be able to pay the mortgage. When doctors treat patients experiencing intense pain, they predict whether the patient would abuse opioid painkillers.

Behavioral scientists have shown that people are notoriously bad at forecasting. We underestimate how long it takes to complete a project (Buehler et al. 1994), we overestimate the odds of rare events (Lichtenstein et al. 1978), we are terrible at forecasting job performance based on job interviews (Dana et al. 2013), and because we overpredict the degree to which our tastes tomorrow will be like our tastes today ("projection bias"), we're too likely to buy convertibles or homes with pools on hot days and vice versa on cold days (Busse et al. 2012).

This is where algorithmic decision tools can add information. Even though people often have access to information about a specific prediction task that algorithms do not have, for decades we've known that even simple linear models often outperform humans on average on most prediction tasks (Dawes et al. 1989).

Modern machine learning algorithms further heighten that advantage. For example, a decision aid can help judges make far more accurate predictions of a defendant's risk of skipping court or getting re-arrested (Kleinberg et al. 2018, Rambachan 2024). This can lead to much better decisions of whether to release or detain the defendant pre-trial: algorithmic predictions create the potential to reduce crime by up to 25% without having to increase detention, or alternatively to reduce the detention rate by over 40% without any increase in crime. This isn't a cherry-picked example. Ludwig et al. (2024) show that for algorithmic decision aids it is common to have such large gains as measured by metrics such as the marginal value of public funds (Hendren and Sprung-Keyser 2020).

3.3 Algorithms can highlight psychological traps.

Even if someone fully understands their options and information, they can still make poor choices because of psychological fallacies. Algorithmic decision aids can be imbued with knowledge about decision heuristics and biases and then remind people of these biases when they arise. This is a functionality that behavioral economists should be involved in helping create, but it's something we've done little of so far.

For example: one common mistake is the "sunk costs fallacy" (Arkes and Blumer 1985), or the tendency to invest in a decision based on past costs that are unrecoverable, rather than based on future costs and benefits. In a classic study, researchers asked college students to imagine they are an airline company that has invested $10 million into an effort to build an undetectable plane, but after the project is 90% finished, they realized that another company had built a better and cheaper version. The question for subjects was: should you invest the last 10% of the research funds to finish your radar-blank plane?

The rational choice here is to *not* invest the extra money, as it would waste $1 million on a failing project. However, 85% of the Ohio and Oregon college students surveyed said that they'd invest. This is a decision error. We know this because when researchers asked students whether they would invest $1 million in the plane without being told they'd already sunk $9 million (same decision, different choice architecture), the majority of respondents chose not to invest.

What happens when we prompt GPT-4o, an off-the-shelf LLM, to help us think through this problem? The algorithm says: "*Ignore the money already spent. The $10 million is gone. It's a sunk cost. The only real decision now is: Is it worth spending the additional money (say, $1 million) to finish the project? Don't let the previous investment cloud your judgment — what matters is what you get for the next dollar you spend.*"

Note that the decision aid is not making the decision on behalf of the decision-maker. But it gives a decision-maker new insights and frameworks for thinking through their decision. Imagine if college students had access to a decision aid like this. Would 85% still choose to invest?

While the potential for algorithms to help us expand our consideration set and improve how we implement our choices is quite obvious, algorithms could also substantially change the way we make our choices as well.

4. Changing Markets: Algorithms as Adaptive Armor in Adversarial Environments

Algorithms don't just create new possibilities for moving people towards their own reflective equilibrium; the fact that algorithms are adaptive - as the world changes and the algorithm's training data changes, the algorithm itself changes - creates entirely new capacities for impact. A challenge with just trying to improve one static decision at a time is that market environments adapt - they're dynamic. But now we have dynamic algorithms that could act as a sort of "adaptive armor" to help people navigate changing choice environments.

Imagine for example that you're choosing between two credit cards. Table 2, Row 1 has what these terms typically look like. This can be confusing. How do you think about what "100,000 bonus points" means? What's "intro APR" vs. "post-intro APR"? Should you care? It's easy to fail to understand the terms of the cards and make a decision you would not have made at reflective equilibrium. It's also easy to miss that the second card has punishing fees for not paying on time, potentially putting a buyer in deep credit card debt.

These are the kind of choices that behavioral economists worry about (Gabaix and Laibson 2006). They involve shrouded attributes and hard comparisons. Attempts to regulate such choices run up against problems. A one-time regulatory nudge that changes some aspects of the credit card terms runs the risk of being undermined by some new intentionally-confusing language added somewhere else in the terms. For example, regulations that required credit card companies to disclose the minimum monthly payment customers had to make to pay off their balance in three years had minimal impacts on how much people actually paid – perhaps because companies hid this disclosure on monthly statements rather than making the information salient at the point people were paying their credit card bill (Keys and Wang 2019).

To test whether algorithms could help in this context, we ran the two example credit card offers through an LLM, prompting it to present the information more clearly, with most important information at the top, to someone who doesn't understand very much about credit cards. The outputs are in Table 2, row 2.

Table 2: Two example credit card offers. The first row is how credit cards are usually described.

	CashBack Silver Card	**MaxPoints Platinum Card**
Original offer	**Earn $200 cash back** after you spend $1,000 in the first 3 months. • **2% cash back** on all purchases, automatically applied as statement credit • No late fees, no penalty APR • Free access to your credit score and budgeting tools • Set custom spending alerts and autopay reminders • Freeze your card instantly from the app **Annual Fee:** $0 **Intro APR:** 0% for 12 months on purchases and balance transfers **Post-intro APR:** 18.24%–25.24% variable APR **Balance Transfer Fee:** 3% (minimum $5) **Foreign Transaction Fee:** None *No penalty APR.*	**Limited-Time Offer: 100,000 bonus points** after you spend $4,000 in the first 3 months. That's up to $1,000 toward travel, gift cards, or statement credits. • **5x points** on dining, entertainment, and rideshare • **3x points** at department stores and online shopping • **1x point** on everything else • Access exclusive VIP concert presales and lifestyle events • Complimentary DoorDash DashPass and Lyft Pink for 12 months **Annual Fee:** $0 the first year (then $99/year) **Intro APR:** 0% for 18 months on purchases **Post-intro APR:** 29.99% variable APR **Buy Now, Pay Later Option:** No interest if paid in full within 18 months on purchases over $500 (*but if not paid in full, interest is charged from the purchase date*) **Late Payment Fee:** Up to $40 **Penalty APR:** Up to 34.99% variable after a late or returned payment
After summarization	**Best for:** People who want simple, steady money back with no yearly fee. **Warning (lower risk)**: After the first 12 months, anything you still owe starts adding a moderate extra cost each month. Moving debt from another card costs 3%. Always pay on time even though there's no late fee or penalty jump. • **$200 back** after $1,000 spent in 3 months. • **2% back on every purchase** (automatically cuts your bill). • $0 annual fee. • Can move other card debt here (3% fee).	**Best for:** People who spend a lot on dining, entertainment, and rides—and are organized enough to always pay in full and track deadlines. **Bigger Warning (higher risk)**: After 18 months, any unpaid amount gets a *very high* extra cost (about 30% yearly). Miss a payment: fee (up to $40) *and* cost can jump even higher. "Pay later" deal on $500+ buys charges all that extra cost back to day 1 if not fully paid within 18 months. • **100,000 points** after $4,000 in 3 months (up to ~$1,000 value). • **5× points** dining/entertainment/rides; **3×** department & online; **1×** other. • **Year 1 $0; then $99/year** (only worth it if points + perks beat $99). • Extra perks: food & ride memberships, VIP presales (use them or lose value).

We venture that anyone choosing based on the LLM summary of the terms, compared to the provider-written terms, has a better understanding of the late fees and the benefits.

More importantly, this type of intervention is harder to game with adversarial strategies. If the credit card company switched to another tactic – e.g. using large fonts for the benefits and small print for the downsides – an LLM would still be able to surface key information and explain it to the user. To be helpful in that sort of dynamic adversarial environment, any behavioral tool also needs to be dynamic and adaptive to act as "armor" against adversarial actors.

5. Changing People: Algorithms as Trainers

Behavioral economics and psychology have largely given up on the idea of de-biasing people; for example, only one of the 35 chapters of the "Handbook of Judgement and Decision-Making," (Keren and Wu 2015) is devoted to the topic. The field's focus on changing choice architecture rather than changing the chooser is not based on a lack of interest in the latter so much as a learned pessimism about the prospects for that based on past experience. However, we argue that people *can* be de-biased and algorithms can act as trainers to help with that.

Consider one successful non-algorithmic example: the Becoming a Man (BAM) program. Young men from Chicago participated weekly to learn about biases and discuss them with a group and facilitator. In two separate randomized controlled trials (RCTs), BAM was shown to reduce violent-crime arrests by 45-50% (Heller et al. 2017). These are remarkable effect sizes, particularly for an intervention implemented in difficult circumstances.

Why was BAM successful? Much of BAM involved discussions, where students talked through their own lives, reviewing and rehearsing examples of psychological biases in their own life. In other words, the key to BAM was opportunities to *practice*.

That shouldn't be a surprise; just think of how we learn, say, algebra. A teacher first explains some concept, then illustrates it with several examples, then gives students lots and lots of problems to give them practice applying the concept. The vast majority of time spent in class and on homework assignments involves doing practice problems again and again.

However, when it comes to decision-making, we typically struggle to get many units of real-world practice. Young men participated in BAM for just an hour a week not necessarily for reasons of optimal pedagogy, but for reasons of cost. We argue that algorithms can provide another way to deliver these practice units.

Some evidence that this is indeed possible comes from Dube et al. (2025), who developed and delivered a behavioral science training program to officers in the Chicago Police Department, a sort of "BAM for cops."

The training included an algorithmically-driven "force simulator" that automates the same sort of practice repetitions that BAM delivers with human providers. The force simulator puts them in ambiguous situations where officers have to construe what's going on and decide what they should do, then gives them feedback that highlights decision-making errors. But the algorithm now makes it possible to greatly increase the intensity of the intervention: the force simulator (once the fixed costs are incurred) lets us deliver countless more practice “reps” at a marginal cost that’s as close to zero as we ever get in public policy.

What good does this semi-automated “BAM for cops” do? In an RCT with 2,070 active Chicago police officers, this training reduces officer’s non-lethal use of force by 23%, reduces by 23% discretionary arrests for minor offenses of the sort that previous research suggests has limited public safety value (Harcourt and Ludwig 2006, Agan et al. 2023), and also reduces racial disparities in such arrests. This encouraging initial proof-of-concept raises the possibility of more wins like this in other applications.

While the training program for cops simulated practice, in other contexts algorithms can turn real life experiences into opportunities for practice. For example, a new product called TeachFX records classroom audio and gives teachers feedback on their class – including total minutes of silence, time spent talking over students, and general alignment of stated goals with what happens in the classroom. This feedback aims to give teachers the opportunity to practice, receive feedback, and update their actions. That’s like “BAM-on-steroids” for teachers.

Algorithms can enable practice and debriefing on demand, at scale, potentially breaking through the “we can’t debias” barrier.

6. What We Need to Build Out

Fully realizing the potential of behavioral economics 2.0 will require scientific advances along many fronts. In this section, we articulate capacities that both economists and computer scientists will need to develop to help make this vision a reality.

6.1. We need to make reflective equilibrium more implementable.

There's an important question we've ignored so far: How do we know when we're getting closer to versus further away from reflective equilibrium? If algorithmic thought partners impact choices, are these choices really closer to the choices that the decision-maker would make at reflective equilibrium? Without some way to answer that question - to "know reflective equilibrium when we see it" - we will be trapped in the same box as behavioral economics 1.0, changing behavior but not knowing whether we are improving welfare (Bernheim and Rangel 2007, Bernheim 2025).

One (imperfect) way to address this question is by measuring decision stability. That is, we assume that people are at reflective equilibrium if their decision does not change when given new framings, persuasive tactics, etc. In a small, random, sample, we could provide users of an algorithmic decision aid with other suggestions and ways of framing the decision. If a user's decisions don't change in the face of these additional perturbations to the choice architecture, we might in some circumstances be willing to assume that's because they are making the decision that they would make under reflective equilibrium.

In other cases, we could use existing behavioral models to make a decent forecast as to whether someone is in reflective equilibrium. For example, our knowledge of the gambler's fallacy might make us confident concluding that the doctor's diagnosis of a patient being seen after the doctor has seen a mix of positive and negative cases is closer to reflective equilibrium than a diagnosis made on the heels of seeing a dozen positive (or negative) cases in a row.

Surely there are other ways, perhaps much better ways, to solve this measurement problem as well. This would have great value for not just behavioral economics 2.0 but any form of behavioral intervention.

6.2. We need algorithms that understand people.

We will also need to develop algorithms that can understand human psychology and behavior. Standard algorithm design focuses on predicting people's behavior as accurately as possible, on the assumption that more accurate prediction aligns with improved outcomes. But if people's behaviors reflect mistakes, predicting people's behavioral mistakes more accurately may simply lead algorithms to more strongly amplify those mistakes.

What we need are algorithms that can tell what we really want from our behavior, recognizing that our behavior doesn't always reflect what we really want. That is, we need some way for algorithms to "invert" people's true preferences - their reflective equilibrium - from their behavior (Kleinberg et al. 2024).

The only way to do that is with behavioral models, and the only way we can incorporate those behavioral models into algorithm design is if they become more precise than they currently are. The most important insights we have from behavioral economics - people are subject to the gambler's fallacy, defaults matter, losses are treated asymmetrically from gains, etc. - are currently all directional and qualitative. But inverting someone's reflective equilibrium from their behavior requires far more specificity; for example, with the gambler's fallacy, what is the shape of the functional form between the number of previous negative (positive) cases in a row the decision-maker has seen and the odds that they classify the next case as negative (positive)? In other words, we need behavioral models that are much more *computable*.[9]

As we have noted above, rather than automatically inverting our preferences from our behavior, an algorithm could instead help us reason our own way to reflective equilibrium. But that would require substantial advances in the way algorithms and humans communicate. For example, consider a simple algorithmic decision aid – for example, one that just communicates a single forecast to a human decision-maker. While the algorithm predicts more accurately than the human on average, the human often has their own source of comparative advantage - private information the algorithm does not have - that might enable them to be more accurate than the

[9] One example is Rabin's (2013) idea of "portable extensions of existing models" (PEEMs). A second-best alternative to formalizing psychological theories is to train algorithms on less computable theories from psychology, using textual data. For example, we could make sure an LLM learns about the gambler's fallacy by inputting academic papers and textbooks as training data. It's easy to think that LLMs already know this, but past work has shown that while LLM might give the illusion of understanding, they do not already have understanding of behavioral economics and psychology (Mancoridis et al. 2025).

algorithm in at least a subset of cases (Ludwig and Mullainathan 2021, Mullainathan 2025). How can the algorithm help the person learn their own comparative advantage relative to the algorithm and vice versa?

Things become even more complicated in less structured, more complicated decision settings. Imagine, for instance, helping a mom select a school for her child. Helping her isn't as simple as "here's a prediction of a single outcome from a supervised learning model." She has a complicated set of objectives. She might not always fully understand her own preferences herself, and so may need help to reason them through. Even if she does know her own preferences she may have trouble articulating that, perhaps leaving some important things unsaid. Or she may *think* she knows her own preferences but is wrong about that.

In other words, the potential for algorithms to serve as thought partners will only be as good as the algorithm's ability to understand the other side of the thought partnership - the humans.

7. Conclusion

Richard Thaler has famously said that behavioral economics should stop existing soon and the name should become obsolete. His reasoning: *all* of economics should become behavioral economics. Much the same could be said for algorithmic behavioral economics. When all is said and done, one hopes that what we are calling algorithmic behavioral economics or behavioral economics 2.0 simply becomes…economics.

References

Agan, Amanda Y, Jennifer L Doleac, and Anna Harvey. "Misdemeanor Prosecution." *The Quarterly Journal of Economics* 138, no. 3 (2023): 1453.

Arkes, Hal R, and Catherine Blumer. "The Psychology of Sunk Cost." *Organizational Behavior and Human Decision Processes* 35, no. 1 (1985): 124–40. https://doi.org/10.1016/0749-5978(85)90049-4.

Bernheim, B. Douglas. "What Is a Mistake? Guardrails for Behavioral Public Policy." *Social Research: An International Quarterly* 92, no. 1 (2025): 35–74. https://doi.org/10.1353/sor.2025.a956286.

Bernheim, B. Douglas, and Antonio Rangel. "Toward Choice-Theoretic Foundations for Behavioral Welfare Economics." *American Economic Review* 97, no. 2 (2007): 464–70. https://doi.org/10.1257/aer.97.2.464.

Börsch-Supan, Axel, Tabea Bucher-Koenen, Michael D. Hurd, and Susann Rohwedder. "Saving Regret and Procrastination." *Journal of Economic Psychology* 94 (January 2023): 102577. https://doi.org/10.1016/j.joep.2022.102577.

Buehler, Roger, Dale Griffin, and Michael Ross. "Exploring the 'Planning Fallacy': Why People Underestimate Their Task Completion Times." *Journal of Personality and Social Psychology* 67, no. 3 (1994): 366–81. https://doi.org/10.1037/0022-3514.67.3.366.

Busse, Meghan, Devin Pope, Jaren Pope, and Jorge Silva-Risso. *Projection Bias in the Car and Housing Markets*. Working Paper No. w18212. National Bureau of Economic Research, 2012. https://doi.org/10.3386/w18212.

Collins, Katherine M., Ilia Sucholutsky, Umang Bhatt, et al. "Building Machines That Learn and Think with People." *Nature Human Behaviour* 8, no. 10 (2024): 1851–63. https://doi.org/10.1038/s41562-024-01991-9.

Dana, Jason, Robyn Dawes, and Nathanial Peterson. “Belief in the Unstructured Interview: The Persistence of an Illusion.” *Judgment and Decision Making* 8, no. 5 (2013): 512–20. https://doi.org/10.1017/S1930297500003612.

Dawes, Robyn M., David Faust, and Paul E. Meehl. “Clinical Versus Actuarial Judgment.” *Science* 243, no. 4899 (1989): 1668–74. https://doi.org/10.1126/science.2648573.

Dube, Oeindrila, Sandy Jo MacArthur, and Anuj K Shah. “A Cognitive View of Policing.” *The Quarterly Journal of Economics* 140, no. 1 (2025): 745–91. https://doi.org/10.1093/qje/qjae039.

Fishbane, Alissa, Aurelie Ouss, and Anuj K. Shah. “Behavioral Nudges Reduce Failure to Appear for Court.” *Science*370, no. 6517 (2020): eabb6591. https://doi.org/10.1126/science.abb6591.

Gabaix, X., and D. Laibson. “Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets.” *The Quarterly Journal of Economics* 121, no. 2 (2006): 505–40. https://doi.org/10.1162/qjec.2006.121.2.505.

Harcourt, Bernard E., and Jens Ludwig. “Broken Windows: New Evidence from New York and a Five-City Social Experiment.” *University of Chicago Law Review* 73, no. 1 (2006). https://chicagounbound.uchicago.edu/uclrev/vol73/iss1/14.

Heller, Sara B., Anuj K. Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A. Pollack. “Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago.” *The Quarterly Journal of Economics* 132, no. 1 (2017): 1–54. https://doi.org/10.1093/qje/qjw033.

Hendren, Nathaniel, and Ben Sprung-Keyser. “A Unified Welfare Analysis of Government Policies*.” *The Quarterly Journal of Economics* 135, no. 3 (2020): 1209–318. https://doi.org/10.1093/qje/qjaa006.

Keren, Gideon, and George Wu, eds. *The Wiley-Blackwell Handbook of Judgment and Decision Making*. Wiley Online Library. Wiley-Blackwell, 2015. https://doi.org/10.1002/9781118468333.

Keys, Benjamin J., and Jialan Wang. "Minimum Payments and Debt Paydown in Consumer Credit Cards." *Journal of Financial Economics* 131, no. 3 (2019): 528–48. https://doi.org/10.1016/j.jfineco.2018.09.009.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133, no. 1 (2018): 237–93. https://doi.org/10.1093/qje/qjx032.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Manish Raghavan. "The Inversion Problem: Why Algorithms Should Infer Mental State and Not Just Predict Behavior." *Perspectives on Psychological Science* 19, no. 5 (2024): 827–38. https://doi.org/10.1177/17456916231212138.

Lichtenstein, Sarah, Paul Slovic, Baruch Fischhoff, Mark Layman, and Barbara Combs. "Judged Frequency of Lethal Events." *Journal of Experimental Psychology: Human Learning and Memory* 4, no. 6 (1978): 551–78. https://doi.org/10.1037/0278-7393.4.6.551.

Ludwig, Jens, and Sendhil Mullainathan. "Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System." *Journal of Economic Perspectives* 35, no. 4 (2021): 71–96. https://doi.org/10.1257/jep.35.4.71.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. "The Unreasonable Effectiveness of Algorithms." *AEA Papers and Proceedings* 114 (May 2024): 623–27. https://doi.org/10.1257/pandp.20241072.

Mancoridis, Marina, Bec Weeks, Keyon Vafa, and Sendhil Mullainathan. "Potemkin Understanding in Large Language Models." Version 2. Preprint, arXiv, 2025. https://doi.org/10.48550/ARXIV.2506.21521.

Mullainathan, Sendhil. "Economics in the Age of Algorithms." *AEA Papers and Proceedings* 115 (May 2025): 1–23. https://doi.org/10.1257/pandp.20251118.

Rabin, Matthew. "An Approach to Incorporating Psychology into Economics." *American Economic Review* 103, no. 3 (2013): 617–22. https://doi.org/10.1257/aer.103.3.617.

Rambachan, Ashesh. "Identifying Prediction Mistakes in Observational Data." *The Quarterly Journal of Economics*139, no. 3 (2024): 1665–711. https://doi.org/10.1093/qje/qjae013.

Rawls, John. *A Theory of Justice*. Belknap Press of Harvard Univ. Press, 1971.

Sunstein, Cass R., ed. *Behavioral Law and Economics*. 1st ed. Cambridge University Press, 2000. https://doi.org/10.1017/CBO9781139175197.

Sunstein, Cass R. *Behavioral Science and Public Policy*. 1st ed. Cambridge University Press, 2020. https://doi.org/10.1017/9781108973144.

Sunstein, Cass R. "Choice engines and paternalistic AI." H*umanities and Social Sciences Communications* 11, no. 1 (2024): 1-4.

Sunstein, Cass R. "Brave new world? Human welfare and paternalistic AI." T*heoretical Inquiries in Law* 26, no. 1 (2025): 1-23.

Sunstein, Cass R. “Fixed Points in Constitutional Theory.” Harvard Public Law Working Paper No. 22-23, May 30, 2022. Available at SSRN: https://ssrn.com/abstract=4123343.

Sunstein, Cass R. “Nudges.Gov: Behavioral Economics and Regulation.” *SSRN Electronic Journal*, ahead of print, 2013. https://doi.org/10.2139/ssrn.2220022.

Sunstein, Cass R., Christine Jolls, and Richard H. Thaler. “A Behavioral Approach to Law and Economics.” *Stanford Law Review* 50 (1998): 1471.

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

Thaler, Richard H., and Shlomo Benartzi. “Save More Tomorrow™: Using Behavioral Economics to Increase Employee Saving.” *Journal of Political Economy* 112, no. S1 (2004): S164–87. https://doi.org/10.1086/380085.