A comment on "Algorithms as a Vehicle to Reflective Equilibrium."

Abhishek Nagaraj[1]
UC Berkeley and NBER

September 30, 2025

In middle school, I helped my dad buy our first car using a popular car magazine's decision tool. The magazine provided scores for each car across dimensions like fuel efficiency, comfort, and looks. Readers assigned personal weights to factors based on their priorities—one might weight fuel efficiency heavily; another might prize comfort. Calculate the weighted sum for each car, and the highest score wins. As this paper notes, life is full of such decisions—choosing a college, selecting a mortgage, picking a health insurance plan, deciding which job offer to accept. Behavioral economics has revolutionized how we understand these choices. The literature now identifies specific biases that distort decision-making (loss aversion, anchoring, present bias, choice overload) and suggests nudges or interventions to reduce their effects. The implicit assumption: decision-makers inadvertently make wrong choices, and interventions can help right this wrong.

While behavioral economics has made far-reaching contributions, the assumption that policymakers can "fix" decisions *en masse* through nudges is somewhat paternalistic. As the authors note, this ignores preference heterogeneity—what's optimal for one person may be suboptimal for another. AI's promise lies in providing personalized decision support through "algorithmic choice architectures" that account for individual preferences and circumstances.

[1] Thanks to Min Min Fong and Arul Murugan for discussion and excellent assistance. All errors are my own.

This approach could help individuals reach "reflective equilibrium"—where decisions align with deeply held values after careful consideration, requiring no further reflection.

If you buy this premise (as I do!), the paper offers a useful framework for how AI could shape such choice architectures. Individuals might not reach reflective equilibria in two ways: "valuation" errors and "choice" errors. Valuation errors occur when individuals use incorrect inputs in the decision process; choice errors occur when individuals have accurate information but fail to select the option that best aligns with their preferences. In our car example, valuation errors would be plugging in wrong fuel efficiency values. Choice errors would be failing to weight factors correctly, perhaps favoring a flashy option over one fitting our family's long-term needs. The paper discusses five ways AI choice architectures could fix both error categories. For valuation errors, AI could provide better calibrated inputs (predicting fuel efficiency's impact on dollar savings) and summarize information more usefully (collating car ownership costs into a single number). For choice errors, AI must help users understand their own preferences through three modalities: urging reflection (switching from System 1 to System 2 thinking), assisting reflection (acting like a coach or therapist), and enabling retrospection (reflecting on past decisions' successes and failures). The authors call for future work building on these foundations, including better measurements of reflective equilibria and more precise behavioral economic theories.

The paper surfaces a foundational tension in behavioral science: what makes a decision "good"? Traditional behavioral economics treats deviations as "biases" to be corrected, sidelining decision-makers' subjective experience and preference heterogeneity. By foregrounding

"reflective equilibrium," the authors re-center agency: two people with identical facts may reasonably choose differently, and legitimate welfare tests should respect that plurality.

Yet this shift belies a deeper problem. Individuals might feel satisfied reflectively but revise their valuations once informational gaps or future-self considerations become salient. The authors rule out this possibility by definition, but I don't find such an extreme view helpful. Reflective equilibria may be stable short-term but fail long-term—individuals might regret choices with months, years, or decades of hindsight. New information might emerge that even AI systems can't account for, affecting assessments and changing reflective equilibria. Which assessment should anchor welfare—immediate reflective endorsement or later reappraisal prompted by new information or experience? This connects to measuring reflective equilibria, which the authors acknowledge is difficult. Their suggestion of measuring preference stability across alternative architectures assumes a short time horizon. Transformative AI may help but think of the seventy-year-old who realizes what passed his reflective test didn't survive time and experience.

To be clear, I agree with the paper's urge to tilt the field towards toward individualized, reflective criteria, a needed corrective to one-size-fits-all nudges. But taken alone, reflection won’t carry the full normative load. A more universal behavioral economics needs a principled balancing rule that weighs (i) respect for person-specific, reflectively endorsed preferences and (ii) carefully circumscribed grounds for intervention when predictable decision errors are at stake. Naming that balancing problem and sketching conditions under which each side should dominate would sharpen the paper’s contribution and its policy bite. With that broader philosophical point made, let me tackle the core framework of the paper which I feel is extremely illustrative, but perhaps could be enriched in three important ways.

First, in addition to valuation and choice errors, I'm glad the paper also names "consideration set" errors as a distinct category. This point should be emphasized more. Transformative AI will be, at its core, a powerful search technology. Its first-order welfare impact may very well come from fixing valuation or choice errors. But its impact on expanding and curating the option set might be bigger. There is serious evidence already that the ways in which AI will help speed up innovation and creativity is through helping creators remix and recombine a much wider set of source materials (Cockburn et al 2018, Doshi and Hauser 2024, Zhou and Lee 2024). We should think more about the relative importance of this channel relative to the others the authors highlight.

Second, the paper largely treats AI as a near-oracle that can elicit stable, well-formed preferences and compute across architectures. Even in the age of Transformative AI, deployed systems might be better understood as proxies for humans: they inherit our data, blind spots, and context limits; they can be sycophantic (amplifying stated preferences) or authoritative in ways that crowd out reflection (Bashkirova et al 2024, Sharma et al 2023). Even if future systems improve on these problems, design choices can tilt "reflection" toward what the system deems salient. I'd suggest the reader explicitly consider some well-known limitations in AI models and assume that some of these will persist given the very nature of model training. Then we could ask questions like: What failure modes arise when AIs themselves are more like therapists and coaches rather than all-seeing oracles? What procedural safeguards (contradictory prompts, dissenting explanations, uncertainty displays) are needed so that "assistance" is aligned with reflective equilibria? Our current system of (human) coaches and therapists are valuable even in a world where they might be biased and fallible, so this is not to say that AI needs to be immaculate to be useful. What role do the authors see for imperfect AI systems to help individuals reach reflective equilibria?

And finally, a subtle assumption in the paper is a clean separation between decision quality and decision mode (AI assisted vs not). But people's welfare assessments depend on ownership: we tend to value outcomes more—and regret them less—when we own the decision process (Zeiser 2024; Nagaraj 2021). In human–AI (HAI) settings, assistance can create an attributability gap: even if an outcome is "better," people feel less agency and responsibility when the path ran through a system's recommendation. That gap is ethically salient (who owns the valuation?) and instrumentally important (ownership predicts adherence and retrospective satisfaction). Behavioral literature on the IKEA effect shows that effortful, self-constructed outcomes are valued more (Norton et al 2012, Ranganathan 2018). Analogously, co-produced valuations may travel better than outsourced ones. For the present framework, the implication is design, not just estimation: build choice architectures that scaffold reflection while preserving authorship of the ultimate decisions.

Finally, more empirical work is needed to test the proposed framework and compare its utility to traditional approaches to nudge individual behavior. While I'm sure such a literature will emerge, I was impatient to see it be applied in practice. With that end, I built on some recent work that uses AI models as simulators of human behavior (Horton 2023, Tranchero et. al 2024) to test the paper's propositions. Specifically, we designed a car choice simulation where agents select between Car A (Practical) and Car B (Luxury). We tasked two AI agents to choose between the luxury and the practical car. We engineered one of these agents to be "biased" in that they are more likely to be swayed to choose the luxury car even though this would be against their "reflective" equilibrium. We then asked both agents to choose a car under three scenarios. The baseline condition involved no interventions. The BE 1.0 condition provided a standard blanket nudge to both participants about the value of practicality over luxury. And the BE 2.0

condition operationalized the "prompting reflection" suggestion of the paper to remind agents to reflect on their own values and choose again if we saw them choosing the luxury car.

Figure 1 below summarizes the results. This simulation concretely shows that a broad, pre-decision nudge (BE 1.01) failed to overcome bias, whereas a personalized, post-decision prompt (BE 2.0) for reflection empowered the agent to achieve its own reflective equilibrium. This finger exercise, though imperfect, bodes well for the predictions of this paper!

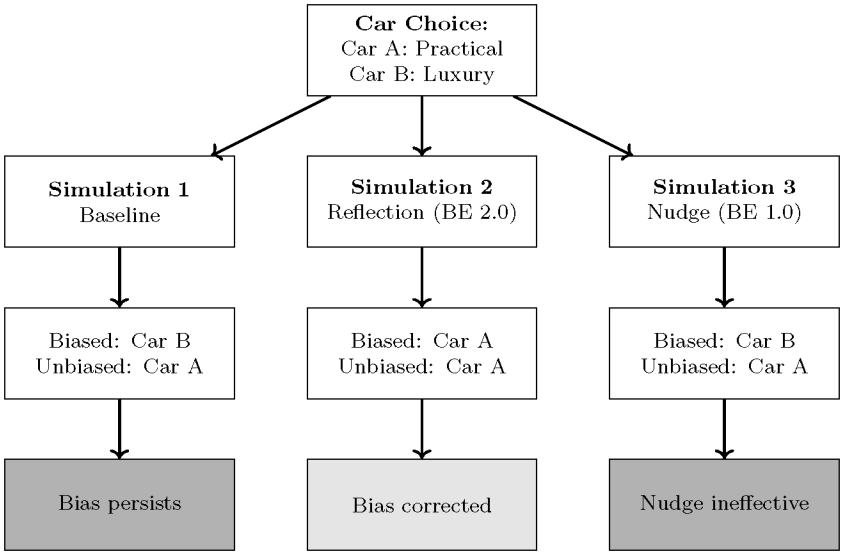

Figure 1. Simulation Results

More broadly, my rudimentary decision-framework did have an important effect on helping my family buy that car—a dark green Maruti 800. We had lots of fond memories getting to interesting places in that tiny vehicle. But what made the choice "good" wasn't the weighted score on a spec sheet; it was that the pick fit us, we owned it, and we would have chosen it again after living with it. That, to me, is the spirit of this paper: we need choice architectures that provide accurate information and invite reflection. If Transformative AI helps more people find

their Maruti 800—not the universally "best" car, but the one they still endorse after the miles add up—then it will have earned its goal of reinventing behavioral economics in the age of transformative AI.

References:

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2022. *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence*. Harvard Business Press.

Bashkirova, Anya, and Dario Krpan. 2024. "Confirmation Bias in AI-Assisted Decision-Making: AI Triage Recommendations Congruent with Expert Judgments Increase Psychologist Trust and Recommendation Acceptance." *Computers in Human Behavior: Artificial Humans* 2: 100066. https://doi.org/10.1016/j.chbah.2024.100066.

Cockburn, Iain M., Rebecca Henderson, and Scott Stern. 2018. "The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis." In *The Economics of Artificial Intelligence: An Agenda*, edited by Ajay Agrawal, Joshua Gans, and Avi Goldfarb, 115–46. University of Chicago Press.

Dinerstein, Michael, Liran Einav, Jonathan Levin, and Neel Sundaresan. 2018. "Consumer Price Search and Platform Design in Internet Commerce." *American Economic Review* 108 (7): 1820–59.

Doshi, Anil R., and Oliver P. Hauser. 2024. “Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content.” *Science Advances* 10 (28): eadn5290.

Horton, John J. 2023. *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* NBER Working Paper 31122. National Bureau of Economic Research.

Nagaraj, Abhishek. 2021. “Information Seeding and Knowledge Production in Online Communities: Evidence from OpenStreetMap.” *Management Science* 67 (8): 4908–34.

Norton, Michael I., Daniel Mochon, and Dan Ariely. 2012. “The IKEA Effect: When Labor Leads to Love.” *Journal of Consumer Psychology* 22 (3): 453–60.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, et al. 2023. “Towards Understanding Sycophancy in Language Models.” *arXiv preprint* arXiv:2310.13548.

Ranganathan, Aruna. 2018. “The Artisan and His Audience: Identification with Work and Price Setting in a Handicraft Cluster in Southern India.” *Administrative Science Quarterly* 63 (3): 637–67.

Tranchero, Matteo, Cecil-Francis Brenninkmeijer, Arul Murugan, and Abhishek Nagaraj. 2024. *Theorizing with Large Language Models*. NBER Working Paper 33033. National Bureau of Economic Research.

Zeiser, Jannik. 2024. “Owning Decisions: AI Decision-Support and the Attributability-Gap.” *Science and Engineering Ethics* 30 (4): 27.

Zhou, Eric, and Dokyun Lee. 2024. “Generative Artificial Intelligence, Human Creativity, and Art.” *PNAS Nexus* 3 (3).