**Applying Behavioral Science to the Implementation of AI Ethics Frameworks**

U.S. leaders and the general public are relatively united in rooting for AI's potential to help solve for some of our most devastating crises, while recognizing the need for oversight to prevent societal hazards ranging from deepening inequalities, degradation of privacy, and a weakening of personal accountability.

Yet AI oversight is still largely being relegated to the goodwill of the companies creating and applying these tools. With the technology advancing so quickly and federal regulations still on the horizon, is the industry capable of policing itself?

On the contrary, we currently observe a kind of **asymmetric warfare regarding AI and human behavior**.

Tech platforms are using AI tools to rapidly *master our behavior*: predicting our preferences, prompting us to buy products we never considered. AI is helping tech companies master our attention – arguably our most precious behavioral resource.

But as tech companies exploit behavioral strategies to remarkable financial advantage – and towards a new kind of global behavioral dominance, these overseers of some of the most potent behavioral modification tools ever established have shown little interest in applying this expertise in behavioral science to their oft-stated goals of establishing ethical AI standards.

We can see this gulf in the ways tech companies apply *nudge theory* to customers, while ignoring this concept in their implementation of AI ethics.

Tech companies masterfully apply both *nudges* and *sludges* on consumers. The nudges are designed with the help of AI-powered systems to identify our preferences and craft our segmented online worlds – encouraging us to purchase in remarkably targeted ways or turn our attention to highly stratified content. Tech companies also exploit sludges: creating friction such that it's harder to click away, or to "unsubscribe" from a monthly purchase than it was to sign up.

Yet while Big Tech pays lip service to "aligning incentives" with ethical outcomes – and are staffing up AI ethics divisions – these same companies ignore strategies such as *nudges* and *sludges* in their ostensible efforts to develop ethical AI. Rather, in their race to bring products to market, these companies are rife with operational nudges and sludges which promote ethical hazards.

AI ethicists largely agree on where tech culture must head. They reliably emphasize the need for greater inclusivity among key decision makers to help root out bias in both data sets and models, the importance of documentation to allow for the transparency needed for oversight, and an attention to ethical considerations built-in at every phase of a company's and/or product's lifecycle.

Yet, despite such broad consensus, **there have been virtually no efforts to harness decades of behavioral science toward operationalizing and implementing these principles**. If we hope for ethical AI development to dominate, we must engage behavioral scientists and their research in shaping this project. Only then will we have a plausible chance of seeing an AI future that truly benefits us all. Until then, AI ethics implementation will be vastly outgunned by its converse, AI-consumer manipulation.

---

*Caroline Friedman Levy, PhD (she/her)*
*Policy Analyst*
*Center for AI and Digital Policy*

*917-363-8282*