Chapter 2
The Role of AI in Detecting and Mitigating Behavioral Biases

Mustafa Kayyali
https://orcid.org/0000-0003-3300-262X
Maaref University of Applied Sciences, Syria

ABSTRACT

Behavioral biases distort human judgment, often steering financial decisions away from rational outcomes. As fields like neurofinance and behavioral economics shed light on the emotional and cognitive triggers behind these biases, artificial intelligence is emerging not just as a diagnostic tool—but as an active agent in correcting flawed reasoning. This chapter explores the nuanced role of AI in detecting and mitigating behavioral biases across financial domains. Drawing from machine learning, behavioral modeling, and cognitive computing, it examines how AI systems can identify patterns of irrationality and offer real-time interventions. But this partnership is not without friction; AI systems themselves may inherit or amplify biases, raising urgent questions of transparency, ethical design, and trust. Ultimately, this chapter advocates for a reflective, interdisciplinary approach—one that treats AI not as a cold corrective mechanism, but as a collaborator in our evolving understanding of human judgment and its limitations.

INTRODUCTION

There's something unsettlingly familiar about the errors we make when it comes to money. We chase sunk costs. We overreact to market noise. We follow the crowd when it feels safe, and panic when things get uncertain—ironically, precisely when

DOI: 10.4018/979-8-3373-1494-5.ch002

steadiness is most needed. For decades, traditional financial theory imagined a rational actor: an agent who weighs all available information, calculates optimal choices, and behaves predictably. But this theoretical construct has been steadily dismantled by wave after wave of psychological and neurological insight. We are not rational agents. We are storytelling creatures—emotional, biased, and prone to shortcuts in reasoning that make sense in daily life but often lead us astray in complex financial landscapes.

Enter behavioral finance, a field that reshaped the way we think about decision-making under uncertainty (Gabhane et al, 2023). Building on the foundations of psychology, it catalogued a wide array of cognitive biases—confirmation bias, loss aversion, overconfidence, anchoring, herding, and more—that influence our economic behavior in subtle but powerful ways. Yet understanding these biases is only part of the equation. Recognizing them is not the same as resisting them. Even seasoned investors and financial professionals, well-versed in the literature, fall prey to the same mental traps (Zheng et al, 2024). Awareness alone often isn't enough to overcome instinct.

Now, we find ourselves at the threshold of a new paradigm—one that merges behavioral science with artificial intelligence. This convergence opens a radical new possibility: that machines, trained on vast patterns of human behavior, might be able to detect our cognitive blind spots in real time—and perhaps even help us course-correct. The same algorithms that once powered ad targeting and automated trading are now being reoriented to understand us, not just as data points, but as flawed, feeling decision-makers.

Artificial intelligence—particularly machine learning—offers powerful tools for surfacing hidden patterns that signal bias (Hanna et al, 2024). By learning from historical behavior, AI systems can identify recurring anomalies: moments when decisions deviate from rational benchmarks in consistent, explainable ways. These systems don’t rely on human introspection or self-awareness; they observe, detect, and adapt, often before we realize we’ve made a misstep. More intriguingly, they can also be designed to intervene—subtly nudging users toward more reflective choices, flagging risky tendencies, or presenting counterfactuals to widen perspective. In a sense, AI becomes a kind of mirror: showing us not just what we’re doing, but how we think. But this mirror is not neutral. It reflects the assumptions of its makers. If AI is to play a role in correcting human bias, it must first be trained not to reproduce or amplify those same errors (Charkra et al, 2024). This introduces a deeply ethical dimension to our inquiry: Can a system designed by humans ever truly be unbiased? What happens when bias correction becomes automated? Are we outsourcing too much of our judgment to machines, and if so, what are we losing in the process?

This chapter sets out to explore these questions with intellectual honesty and interdisciplinary curiosity. It begins with a grounded explanation of the most persistent