Explainable Artificial Intelligence for Financial Decision-Making: Modeling Investor Self-Beliefs through Natural Language and Behavioral Data

Mihir Mhanna
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
hellomihir@uc.edu

Darren Fox
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
darren.fox@oregonstate.edu

Jasser Olivor
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
jesseoliver@uab.edu

Abstract

The integration of artificial intelligence into financial decision-making has transformed the speed, scale, and complexity with which investment choices are made, yet opaque algorithmic rationales continue to erode trust and obscure latent psychological drivers of market behavior. This paper addresses the dual challenge of rendering AI-based financial systems interpretable while incorporating investor self-beliefs as measurable constructs derived from natural language and behavioral data. We propose a system-level framework that unites explainable artificial intelligence (XAI) with computational modeling of self-schemata, capturing how an individual's narrative of competence, control, and identity shapes risk perception, portfolio allocation, and reaction to market volatility. Drawing on interdisciplinary foundations in behavioral finance, natural language processing, and socio-technical infrastructure, the paper analyzes the architectural trade-offs involved in fusing heterogeneous data streams—ranging from social media discourse and earnings call transcripts to biometric and interaction logs—within a transparent inference pipeline. Emphasis is placed on the governance structures required to manage data provenance, model drift, and multi-stakeholder accountability, as well as on the fairness implications of encoding psychologically sensitive attributes. Through a comparative discussion of post-hoc explanation methods and inherently interpretable architectures, we highlight how different design choices impact robustness, latency, regulatory compliance, and the capacity to surface self-belief dynamics over time. The analysis extends to deployment considerations at scale, examining sustainability in live trading environments, the challenges of continuous recalibration against evolving investor narratives, and the policy frameworks needed to ensure that self-belief modeling does not become a vector for manipulation or exclusion. By reframing financial XAI as a socio-cognitive infrastructure, the paper provides a roadmap for building systems that not only explain decisions in post-hoc terms but also reveal the recursive loop through which an investor's self-concept is both input to and reshaped by artificial intelligence.

Keywords

Explainable artificial intelligence; financial decision-making; investor self-beliefs; natural language processing; behavioral data; system architecture 1 Introduction The accelerating adoption of machine learning in asset management, retail investing, and institutional trading has generated a pressing need for explanatory mechanisms that transcend mere performance metrics and address the cognitive and social dimensions of financial reasoning. Contemporary AI models often rely on high-dimensional representations of market signals, textual sentiment, and behavioral traces, producing outputs that are inscrutable to end-users and regulators alike. The XAI research community has responded with a diverse toolkit of interpretability methods, yet the dominant emphasis on model transparency has frequently overlooked a deeper challenge: the fact that financial decisions are not only shaped by external information but are heavily mediated by an investor's self-beliefs, including perceived competence, overconfidence, self-efficacy, and identity narratives [1,2]. When trading algorithms and robo-advisory systems deliver recommendations without rendering visible the psychological constructs embedded in their input features or inferred by their internal representations, they risk alienating users and undermining the reflective capacity that is essential for long-term financial well-being. This paper advances the thesis that explainable AI for financial decision-making must be reconceived as a socio-technical infrastructure that actively models and communicates the role of investor self-beliefs, drawing on natural language and behavioral data streams to make the self-referential loop between machine inference and human identity both observable and governable. Financial markets have long been understood as ecosystems saturated with narrative, affect, and identity performance, and the behavioral finance literature has systematically documented the influence of overconfidence, self-attribution bias, and mental accounting on trading frequency, portfolio diversification, and risk-taking [2,4]. Concurrently, the proliferation of digital platforms has yielded vast quantities of textual and behavioral data that carry rich signals about how individuals construct and narrate their financial selves. Social media posts, earnings call commentaries, search queries, and interaction patterns on trading applications all provide windows into the linguistic framing of agency, control, and vulnerability. Computational linguistics and natural language processing techniques now make it feasible to extract fine-grained indicators of self-belief from large-scale text corpora, moving beyond generic sentiment toward the measurement of modal verbs, first-person pronoun usage, causal attribution patterns, and epistemic stance [3,6]. At the same time, behavioral data such as login frequency, order revision behavior, post-trade regret expressions, and biometric responses captured through wearable devices offer complementary channels for triangulating an investor's self-concept under conditions of uncertainty. The integration of these heterogeneous data streams into a unified XAI framework introduces substantial architectural and governance challenges that demand a system-level perspective. Decisions about where explanation is generated—at the feature extraction layer, within a latent representation, or via a post-hoc surrogate—carry consequences for latency, auditability, data sovereignty, and the potential for explanation to be weaponized as a tool of persuasion rather than empowerment. Furthermore, modeling self-beliefs at scale raises novel fairness concerns, as psychological profiles may be correlated with protected attributes such as age, gender, or socioeconomic background, and the deployment of such models in credit assessment, insurance underwriting, or targeted product recommendation could amplify existing inequalities if not subjected to rigorous oversight [7,8]. The present paper addresses these structural trade-offs through an extended conceptual analysis that connects advances in XAI, behavioral finance, and computational social science, and that proposes design principles for future systems capable of rendering self-belief dynamics transparent, contestable, and aligned with investor welfare. 2 Theoretical Foundations and Behavioral

Data Ecologies Understanding how self-beliefs influence financial decisions requires an interdisciplinary synthesis that bridges cognitive psychology, linguistic anthropology, and machine learning. Classic behavioral models characterize overconfidence as a miscalibration between subjective probability assessments and objective accuracy, manifesting in excessive trading, under-diversification, and the illusion of control. Self-efficacy theory further distinguishes between generalized confidence and domain-specific beliefs about one's capability to execute courses of action, which in turn shape how investors interpret gains and losses, seek information, and react to feedback from automated advisory systems. These constructs are not static; they evolve through social comparison, market experience, and narrative self-construction, which means that any AI system that purports to explain financial decisions must capture the dynamic interplay between self-belief and market behavior rather than treating psychological variables as fixed traits [4,6]. Natural language data offer a particularly powerful medium for modeling self-beliefs because language is the primary vehicle through which individuals articulate, negotiate, and revise their self-concepts. Linguistic markers such as pronoun shifts, hedging, causal connectives, and expressions of certainty or doubt have been linked empirically to psychological states including self-esteem, locus of control, and cognitive dissonance. In financial contexts, the language used by executives during earnings calls, by analysts in research reports, and by retail investors on discussion forums carries layered signals about collective and individual self-belief. Recent work in computational linguistics has demonstrated that models trained on large corpora can recover representations of self-schemata from interactional data, revealing how the way individuals talk about themselves covaries with decision-making patterns [6]. These insights suggest that XAI systems designed for finance should incorporate explicit modules for self-belief inference that operate over continually updated textual and behavioral streams, providing interpretable outputs that contextualize a model's recommendation in terms of the inferred psychological state of the user. The behavioral data ecology extends beyond text to encompass fine-grained digital traces that record how investors configure interfaces, set alerts, execute and cancel trades, and respond to notifications. Such data capture pre-reflective aspects of self-belief that may not surface in explicit language, including hesitation, impulsivity, and risk avoidance. The fusion of linguistic and behavioral signals within a single modeling framework entails significant data engineering and alignment burdens, requiring robust pipelines for temporal synchronization, noise reduction, and missing data handling. Moreover, the very act of modeling self-belief creates a recursive loop: when an AI system communicates an inferred psychological profile to the user, that feedback can alter the user's subsequent behavior and self-narrative, potentially inducing self-fulfilling prophecies or defensive reactions. System designers must therefore build mechanisms that not only explain financial decisions with reference to self-beliefs but also allow users to interrogate, challenge, and update the system's psychological model of them, thereby preserving agency and preventing the ossification of outdated or inaccurate self-representations. 3 System Architecture for Multimodal Self-Belief Inference Designing an XAI infrastructure that operationalizes self-belief modeling demands careful attention to the layering of data ingestion, feature engineering, inference, and explanation components so that transparency is not an afterthought but a constitutive property of the architecture. At the data layer, structured and unstructured inputs must be ingested through secure, consent-based pipelines that preserve provenance metadata, an essential requirement for any explainability regime. Textual sources such as social media excerpts, chat logs with robo-advisors, and journal entries require natural language preprocessing modules capable of extracting psychologically relevant lexico-grammatical features while protecting user privacy through on-device

processing or federated learning configurations. Behavioral telemetry, including interaction rhythms, trading frequency, and biometric arousal indicators, must be captured in a temporally granular format that allows alignment with textual narratives and market events. The architecture must support the coexistence of these diverse modalities without forcing premature simplification, which often introduces distortions that undermine both predictive accuracy and the fidelity of subsequent explanations. The inference layer confronts a fundamental trade-off between model fidelity and interpretability that is particularly acute when modeling nuanced psychological constructs. Black-box deep learning architectures, such as transformer-based language models fine-tuned on financial discourse, can capture subtle contextual dependencies in self-belief expressions, yet their opacity complicates the task of tracing how a given linguistic pattern contributed to a specific portfolio recommendation. In contrast, inherently interpretable models such as generalized additive models with pairwise interactions or rule-based systems built on psycholinguistic feature sets provide direct traceability but may under-represent the complex, hierarchical structure of self-belief dynamics. Hybrid architectures that embed psychologically grounded latent variables within a broader neural framework, and then expose those variables through concept-based explanation interfaces, offer a promising middle ground, enabling both predictive power and the surfacing of self-relevant dimensions like perceived risk competence or narrative coherence. The design of such architectures requires close collaboration between computational linguists, behavioral scientists, and HCI specialists to define psychologically valid constructs that can be operationalized without losing their multidimensional character. Explanation generation must be integrated at multiple levels of the system: local explanations that clarify a single recommendation in terms of recent linguistic and behavioral cues, global explanations that reveal long-term patterns in how the investor's self-belief correlates with financial outcomes, and counterfactual explanations that illustrate how a different self-narrative might have led to an alternative decision path. The computational cost of maintaining multiple explanation modalities introduces latency concerns, particularly for applications in high-frequency trading environments where decisions occur in sub-second intervals. System architects must therefore implement a caching and incremental update strategy that pre-computes explanation scaffolds during low-activity periods and adapts them in real time as new data arrive. The infrastructure must also be resilient to concept drift, both in market dynamics and in the investor's evolving self-concept, which requires continual monitoring of explanation fidelity and the deployment of fallback interpretability mechanisms when primary models exhibit degraded calibration. This orchestration of data, inference, and explanation components constitutes a complex socio-technical system whose dependability hinges on observability, versioning of psychological models, and robust rollback capabilities. 4 Explainability Mechanisms and the Transparency-Performance Frontier The decision to adopt a particular class of explainability mechanisms in a financial self-belief modeling system is far from trivial, as it implicates competing definitions of what counts as an adequate explanation for different stakeholders. Post-hoc explanation methods, such as LIME, SHAP, and integrated gradients, can be retrofitted onto high-performing black-box models and provide localized feature attribution scores that illuminate which textual or behavioral signals most influenced a given prediction. While these methods are computationally convenient and broadly applicable, they suffer from several well-documented limitations: instability across similar inputs, sensitivity to the choice of perturbation kernel or baseline, and an inability to capture higher-order interactions among self-belief dimensions that jointly drive decision behavior [1]. Moreover, when applied to language data, feature attribution scores often highlight individual tokens without conveying how those tokens combine into coherent

narratives about the self, thereby providing a fragmented and psychologically impoverished explanation that may mislead users into over-relying on isolated cues. Inherently interpretable architectures, including attention-based recurrent networks with structured attention heads that align with predefined psychological categories, or prototype-based models that ground explanations in actual exemplars drawn from the user's own behavioral history, offer a complementary path toward transparency. Such architectures can be designed to output not only a decision but also a structured representation of the investor's self-belief state, mapping it onto dimensions such as perceived internal versus external locus of control, stability of self-confidence, and narrative coherence. The challenge lies in constraining the model sufficiently to ensure that these dimensions remain meaningful without sacrificing the expressive capacity needed to capture idiosyncratic self-representations that vary across cultural and linguistic contexts. A further complication arises from the fact that self-belief is inherently a latent construct, and any architectural commitment to a particular factor structure may inadvertently impose a normative psychological framework that marginalizes non-normative self-concepts, introducing a subtle but profound form of representational bias. The transparency-performance frontier in self-belief-aware financial XAI is therefore not a fixed technical boundary but a design space shaped by regulatory expectations, user trust, and the acceptable cost of explanation errors in high-stakes financial contexts. For systems deployed in consumer-facing robo-advisory platforms, where the primary goal is to foster long-term financial literacy and reflective decision-making, sacrificing a modest amount of predictive accuracy for greater explanatory richness and user contestability may be justified. In contrast, institutional algorithmic trading systems that incorporate self-belief sentiment from executive communications may prioritize predictive performance, relegating explainability to periodic model audits rather than real-time transparency. Navigating this frontier requires governance frameworks that explicitly define the minimal acceptable standards of explanation for different use cases, incorporating insights from human factors research on how investors with varying levels of financial sophistication interpret and act upon self-referential explanations [5,9]. Multi-stakeholder deliberation, involving regulators, consumer advocates, and behavioral ethicists, is essential to ensure that technical optimization does not occur in a normative vacuum. 5 Governance, Fairness, and the Societal Implications of Self-Belief Modeling The encoding of investor self-beliefs within operational AI systems raises a constellation of governance challenges that extend far beyond model accuracy and into the domains of data rights, non-discrimination, and psychological autonomy. Self-belief data inferred from language and behavior are not merely sensitive in the privacy sense; they touch upon the core of personal identity, and their misuse could enable forms of manipulation that exploit cognitive biases with unprecedented precision. A robo-advisor that detects a user's declining self-efficacy could, in principle, calibrate its recommendations to nudge the user toward riskier products under the guise of empowerment, or conversely, could steer a momentarily overconfident user toward excessively conservative allocations, thereby undermining the very autonomy that financial advisory systems claim to support. Governance structures must therefore encode ethical boundaries around the uses of inferred psychological states, perhaps through mandatory disclosure of self-belief model outputs to users and the establishment of a right to opt out of psychological profiling without losing access to core financial services [7,10]. Fairness considerations multiply when self-belief models are trained on datasets that reflect historical inequalities in financial participation and linguistic expression. Language models pre-trained on large web corpora have been shown to absorb societal stereotypes related to gender, race, and socioeconomic status, and these biases can propagate into downstream self-belief inference when the model associates certain patterns of

self-reference with particular demographic groups. An XAI system that attributes cautious language to low financial competence may disproportionately disadvantage women or minority investors whose communication styles have been shaped by systemic marginalization rather than by underlying inability. Addressing these biases demands a layered approach: debiasing at the representation level through adversarial training or counterfactual data augmentation, routine fairness auditing across intersectional subgroups, and the involvement of diverse annotator pools during the development of psychologically labeled datasets. Importantly, fairness cannot be reduced to a static metric; it must be monitored longitudinally as self-belief models co-evolve with market culture and linguistic norms, requiring governance infrastructures that include ongoing participatory feedback loops from affected communities. The policy dimension of self-belief modeling intersects with existing regulatory frameworks such as the European Union's Artificial Intelligence Act and the evolving guidance from financial conduct authorities on the use of AI in consumer credit and investment services. These frameworks increasingly demand that AI systems operating in high-risk domains provide meaningful transparency and allow for human intervention, yet they offer limited specific guidance on the treatment of inferred psychological constructs. Developing industry standards and regulatory sandboxes that specifically address the unique risks of self-belief modeling is an urgent priority. Such standards could define stratified explanation requirements: a base layer of interpretability that discloses the existence and general logic of self-belief inference, an intermediate layer that provides aggregated insights suitable for consumer-facing dashboards, and a deep audit layer accessible to regulators and independent researchers. This stratified approach balances the confidentiality of proprietary models with the societal need for accountability, while also creating a pathway for cross-jurisdictional harmonization of norms around psychologically informed AI in finance [8,11]. 6 Deployment, Sustainability, and Long-Term Evolution Transitioning a self-belief-aware XAI system from a controlled research prototype to a resilient, production-grade financial service entails confronting a host of engineering and operational challenges that are seldom addressed in the academic literature on explainability. One of the most significant is the management of temporal dynamics: investor self-beliefs are not stationary properties but evolve in response to life events, market shocks, and the very feedback provided by the AI system itself. A deployment architecture must therefore incorporate continual learning loops that periodically update self-belief representations while guarding against catastrophic forgetting and destabilizing feedback cycles that could amplify short-term psychological fluctuations. Federated and decentralized learning paradigms offer a promising avenue for enabling personalization without centralizing sensitive psychological data, though they introduce additional complexity in ensuring that explanation modules remain consistent across asynchronously updated model instances. The choice of edge versus cloud computation for self-belief inference directly impacts latency and energy consumption, with implications for the environmental sustainability of deploying such systems at the scale of millions of retail investors [5,12]. Robustness to adversarial manipulation constitutes a further critical dimension of system sustainability. Financial markets incentivize strategic behavior, and adversaries might attempt to game self-belief models by injecting crafted language into public forums or by altering behavioral patterns to influence algorithmic recommendations that, in turn, move market prices. Defense-in-depth strategies are required, including anomaly detection at the data ingestion layer, ensemble-based robustness checks at the inference layer, and explanation consistency monitoring that flags when a user's self-belief profile shifts in ways that are incongruent with their long-term behavioral history. These robustness measures must be complemented by business continuity planning that specifies fallback procedures

when the self-belief inference pipeline is compromised or when regulatory orders mandate its temporary suspension. The operational discipline required to maintain such systems underscores the need for interdisciplinary operations teams that blend expertise in machine learning, cognitive psychology, and financial risk management. The long-term sustainability of self-belief modeling in finance also depends on the cultivation of public trust and the demonstration of tangible benefits that extend beyond incremental improvements in portfolio returns. Research should investigate how transparent self-belief feedback influences investor well-being, decision quality, and engagement with financial planning, ideally through longitudinal field studies conducted in partnership with financial institutions. The findings of such studies would inform iterative refinements to both the technical architecture and the governance frameworks, ensuring that the system evolves in a direction that is technically robust, ethically sound, and socially valued. By treating the self-belief modeling infrastructure as a living socio-technical system rather than a one-time deployment, developers can align the system's trajectory with the broader goals of inclusive and psychologically informed financial innovation [13,14]. 7 Conclusion This paper has argued that explainable AI for financial decision-making must move beyond narrow technical conceptions of transparency to engage with the deeply human phenomenon of investor self-belief. By synthesizing insights from behavioral finance, natural language processing, and systems engineering, we have outlined a vision in which linguistic and behavioral data streams are fused within a transparent architecture to model and communicate the role of self-schemata in shaping financial behavior. The analysis has highlighted the architectural trade-offs between black-box performance and interpretable psychological representations, the governance imperatives required to prevent the weaponization of self-belief inference, and the fairness challenges that arise when models absorb societal biases through language. Deployment considerations have further underscored the importance of temporal robustness, adversarial resilience, and participatory design in ensuring that such systems remain sustainable and trustworthy over the long term. The intersection of XAI and self-belief modeling is not merely a technical frontier but a site of profound normative choice. Systems that render visible the reciprocal influence between machine inference and investor identity have the potential to deepen financial self-awareness and promote reflective decision-making, yet the same capabilities, if left ungoverned, could erode autonomy and entrench inequality. The path forward requires a multidisciplinary coalition of researchers, regulators, and industry practitioners committed to building infrastructures that honor both the complexity of human psychology and the demands of technical transparency. As the financial industry continues its inexorable integration of AI, the conceptual and structural frameworks presented here offer a foundation for ensuring that explainability serves not only the instrumental goals of compliance and risk management but also the more ambitious aspiration of fostering a financial ecosystem in which technology amplifies, rather than diminishes, human understanding of the self.

References

1. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.

2. Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. The Quarterly Journal of Economics, 116(1), 261-292.

3. Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research, 54(4), 1187-1230.

4. Gervais, S., & Odean, T. (2001). Learning to be overconfident. The Review of Financial Studies, 14(1), 1-27.

5. Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: deep portfolios. Applied Stochastic Models in Business and Industry, 33(1), 3-12.

6. Solanki, D., Hsu, H. M., Zhao, O., Zhang, R., Bi, W., & Kannan, R. (2020, July). The way we think about ourselves. In International Conference on Human-Computer Interaction (pp. 276-285). Cham: Springer International Publishing.

7. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

8. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707.

9. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

10. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

11. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM/2021/206 final.

12. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.

13. Susskind, R., & Susskind, D. (2015). The future of the professions: How technology will transform the work of human experts. Oxford University Press.

14. Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. Science, Technology, & Human Values, 41(1), 118-132.

15. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

16. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59-68). ACM.

17. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.

18. Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. Journal of Economic Perspectives, 21(2), 129-152.

19. Hirshleifer, D. (2015). Behavioral finance. Annual Review of Financial Economics, 7, 133-159.

20. McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text. In Proceedings of the 7th ACM Conference on Recommender Systems (pp. 165-172). ACM.