Prospect Theory and Machine Learning-Based Goal Design in Flexible Labor Markets

Menish V. Sharma
Department of Computer Science, University of Houston, Houston, TX, USA.
manish1996@uh.edu

Bastian Ray
Department of Computer Science, University of North Texas, Denton, TX, USA.
bastian.ray@unt.edu

Blake Lyons
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
helloblake@unr.edu

Otis Lawson
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
otis1975@binghamton.edu

Abstract

Flexible labor markets, typified by ride-hailing, food delivery, and freelance platforms, grant workers unprecedented autonomy over when and how much to work. Yet this autonomy shifts the burden of self-regulation onto individuals, making the design of effective earning and effort goals a central challenge for both worker welfare and platform sustainability. This paper develops an interdisciplinary framework that integrates prospect theory with machine learning to personalize goal recommendations in these environments. We argue that classical expected utility models fail to capture the reference-dependent and loss-averse preferences that dominate gig workers' labor supply decisions. Drawing on the fourfold pattern of risk attitudes and the dynamics of reference point adaptation, we explore how behavioral goal design can improve decision quality, reduce churn, and align platform incentives with worker well-being. We then examine the machine learning architectures required to infer latent cognitive parameters and deliver personalized goals at scale, analyzing contextual bandits, meta-learning, and causal inference approaches. A central contribution is a system-level analysis of the trade-offs inherent in building such infrastructures: latency versus model sophistication, privacy preservation versus personalization depth, and fairness constraints versus efficiency gains. We discuss governance structures that must accompany algorithmic goal systems, including transparency mandates, auditability, and safeguards against exploitative framing. Through a deployment-oriented lens, we address sustainability challenges such as concept drift, engagement spirals, and the co-evolution of worker behavior and model recommendations. The paper concludes by outlining a research agenda that treats goal design as a socio-technical optimization problem, bridging behavioral economics, systems engineering, and policy design to create labor platforms that are simultaneously efficient, equitable, and human-centered.

Keywords

prospect theory, goal design, flexible labor markets, machine learning, personalization, platform governance.

1. Introduction

The rapid expansion of digitally mediated flexible labor markets has fundamentally reshaped the employer-worker relationship, dissolving the traditional boundaries of the salaried employment contract. Workers on platforms that offer on-demand services often face an unstructured work environment in which they must independently decide not only how many hours to supply but also how to pace their effort across days, weeks, and seasons. In the absence of external managerial oversight, the cognitive burden of self-motivation falls squarely on the individual. A growing body of evidence indicates that many gig workers set informal earning targets or activity goals to structure their labor supply, yet these self-imposed goals are frequently suboptimal, driven by present bias, reference point effects, and loss aversion rather than by a calculated maximization of long-run utility. Platforms, in turn, have begun to experiment with digitally nudged goals—such as "complete three more trips to earn a bonus" or "you are $15 away from your daily target"—in an effort to increase engagement and throughput. However, the design of these goal interventions rarely rests on a rigorous behavioral or computational foundation, raising concerns about worker manipulation, unfair distribution of incentives, and long-term burnout.

This paper proposes a unified framework that combines prospect theory with modern machine learning to systematically design and personalize goals in flexible labor markets. Prospect theory, originally formulated by Kahneman and Tversky [1], provides a descriptive model of decision-making under risk that accounts for reference dependence, loss aversion, and diminishing sensitivity. Subsequent developments in cumulative prospect theory [2] and third-generation prospect theory [3] have enriched the theoretical apparatus, enabling more nuanced predictions about behavior in dynamic, multi-period settings. In parallel, the field of machine learning has produced powerful tools for inferring latent preference parameters from observational data, optimizing interventions through reinforcement learning, and guaranteeing fairness through constrained optimization [4,5]. By coupling these two intellectual traditions, we can move beyond generic nudges toward goal designs that respect workers' heterogeneous risk attitudes, adapt to evolving reference points, and promote sustainable engagement without crossing into exploitative territory.

The contribution of this work is threefold. First, we provide a comprehensive synthesis of the behavioral mechanisms through which prospect theory explains gig workers' responses to goal-setting interventions, extending the standard analysis of labor supply to incorporate dynamic reference point shifts and narrow framing. Second, we delineate a series of machine learning architectures—contextual bandits, inverse reinforcement learning, and meta-learning—that can operationalize prospect-theoretic goal personalization at scale, while explicitly addressing the systems-level challenges of latency, cold starts, and non-stationarity. Third, we embed the technical discussion within a broader governance and sustainability framework, mapping the design space of algorithmic goal systems onto regulatory principles of transparency, fairness, and accountability. Throughout, we maintain a system-centric perspective, emphasizing the architectural trade-offs, infrastructure requirements, and socio-technical feedback loops that determine the viability of such systems in production.

2. Theoretical Foundations: Prospect Theory and Flexible Labor Dynamics

Prospect theory departs from expected utility theory by positing that individuals evaluate outcomes relative to a reference point, exhibit loss aversion (losses loom larger than equivalent gains), and weight probabilities in a nonlinear fashion that overweights small probabilities and underweights medium to large probabilities [1]. These cognitive patterns have profound implications for goal design in flexible labor markets because they determine how workers perceive the attractiveness of earning targets, bonuses, and shortfalls. When a platform suggests a daily earning goal, that goal immediately becomes a candidate reference point. A worker who falls short of the target experiences the shortfall as a loss, triggering a strong motivational response due to loss aversion. Conversely, once the target is met, additional earnings are coded as gains in a region of diminishing sensitivity, meaning that the marginal utility of extra work declines rapidly. This asymmetry can generate a kinked labor supply curve that is difficult to reconcile with standard neoclassical models but aligns well with the fourfold pattern of risk attitudes captured in cumulative prospect theory [2].

The dynamic nature of flexible work adds further layers of complexity. Workers often make sequences of intertemporal labor supply decisions under uncertainty, facing stochastic wages, fluctuating demand, and variable non-work commitments. Reference points are not static; they adapt over time based on recent income, peer comparisons, and platform-generated benchmarks. When a worker repeatedly achieves a high goal, their reference point can ratchet upward, leading to hedonic adaptation and a need for ever-increasing targets to sustain the same perceived gain. This process, akin to the hedonic treadmill, carries the risk of escalating effort demands that can precipitate exhaustion. Conversely, if a worker consistently fails to reach a demanding target, learned helplessness may set in, reducing future motivation. The behavioral literature on goal setting has long recognized that difficult yet attainable goals enhance performance [6], but the precise calibration of difficulty must account for loss aversion, reference point stickiness, and the individual's degree of narrow framing—the tendency to evaluate each earning episode in isolation rather than as part of a broader portfolio of outcomes [7].

Flexible labor markets amplify the relevance of narrow framing because platform interfaces often present workers with discrete, session-based summaries— "you earned $34 this lunch shift"—which encourages piecewise mental accounting. When goals are presented within these micro-timeframes, the worker's decision utility becomes overwhelmingly driven by the immediate reference point rather than by a cumulative wealth perspective. Experimental evidence from field studies with gig workers confirms that setting personalized earning goals can significantly increase labor supply and earnings, but the effects are heterogeneous and depend on the specificity and reference level of the goal [8]. Workers who are more loss-averse respond more strongly to loss-framed goals that emphasize avoiding a shortfall, while those with lower loss aversion may respond better to gain-framed goals that highlight surplus. This heterogeneity underscores the necessity of personalization, as a one-size-fits-all goal strategy will inevitably be misaligned with a substantial fraction of the workforce.

A further theoretical consideration concerns the interaction between prospect theory and social preferences. Gig workers often compare their earnings not only to their own historical benchmarks but also to peers. Reference points can be socially constructed, and goals that carry implicit social comparisons can activate both upward- striving motivation and feelings of inequity. Systems that display leaderboards or average earnings of similar workers introduce a social rank-based reference point that can amplify the impact of goal framing. These social dynamics must be accounted for in models of goal response, because they can

shift the probability weighting function and reference point adaptation rate in ways that vary across cultural contexts and personality traits.

3. Machine Learning Architectures for Goal Personalization

Translating the insights of prospect theory into operational goal design requires a computational layer that can infer latent behavioral parameters from trace data, predict how a given worker will respond to alternative goal frames, and continuously optimize recommendations in real time. The learning problem is inherently a contextual bandit task: the platform observes a context vector for each worker at a decision point (features such as recent earnings trajectory, time of day, weather, historical goal attainment, and session characteristics), selects a goal recommendation from a set of feasible interventions, and receives a reward signal that captures both platform objectives and worker well-being. Multi-armed bandit algorithms and their contextual extensions, such as linear upper confidence bound or Thompson sampling approaches, have been widely used in recommendation systems [4] and can serve as the backbone for goal personalization. However, standard contextual bandits typically assume a stationary reward function and do not explicitly model the reference point dynamics that are central to prospect theory. To address this, the bandit formulation can be modified to incorporate a state-dependent utility transformation that applies a value function shaped by loss aversion and diminishing sensitivity to the observed reward. This allows the system to optimize goal recommendations with respect to the worker's prospect-theoretic utility rather than raw earnings.

A more ambitious approach leverages inverse reinforcement learning to recover a worker's underlying reward function from observed sequences of decisions, implicitly capturing their risk attitudes and reference points. Recent work in behavioral inverse reinforcement learning has shown promise in inferring prospect theory parameters from demonstrated choice patterns [9]. Once a personalized model of the worker's decision process is learned, forward optimization can generate goal recommendations that maximize predicted experienced utility, subject to platform-level constraints on aggregate supply or fairness. This method faces significant challenges related to identifiability, data sparsity, and the curse of dimensionality. Each worker's behavioral history may consist of only a few hundred decision points, making it difficult to jointly estimate prospect theory parameters, reference point dynamics, and environmental influences. Transfer learning and meta-learning paradigms offer a solution: by training a population-level model on data from thousands of workers and then adapting it to an individual with few-shot learning, the system can rapidly personalize goals even for new entrants. Model-agnostic meta-learning has been applied to personalization tasks and could be extended to the behavioral setting by structuring the inner-loop adaptation to update prospect theory parameters based on the worker's early interactions [10].

Causal inference must be embedded into the learning architecture to disentangle the effect of goal interventions from confounding factors. When a platform observes that workers who receive higher goals work longer, this association may reflect a selection bias wherein intrinsically motivated or opportunity-rich workers are assigned such goals. Randomized controlled trials provide the gold standard for estimating causal effects, but continuous A/B testing is resource-intensive and raises ethical concerns when the intervention space includes potentially detrimental goal frames. Off-policy evaluation techniques, such as doubly robust estimation or importance sampling with learned behavior policies, enable the use of observational data to estimate treatment effects under counterfactual goal policies [11]. These methods can be combined with fairness constraints to ensure that goal personalization does

not systematically disadvantage protected subgroups, an issue we revisit in the governance section.

The machine learning architecture must also contend with non-stationarity arising from the co-evolution of worker behavior and platform algorithms. As workers adapt to personalized goals, their reference points shift, changing the mapping from goal to response. This feedback loop can be modeled as a coupled dynamical system in which the platform's policy and the worker's behavioral parameters co-evolve over time. Control-theoretic approaches, such as model predictive control with online re-estimation of prospect theory parameters, offer a principled way to manage these dynamics while maintaining stability. However, the computational complexity of repeatedly solving a personalized, prospect-theoretic optimal control problem at scale pushes against real-time latency constraints, leading to architectural trade-offs between model sophistication and responsiveness.

4. System Design and Infrastructural Trade-offs

Deploying a prospect-theoretic goal personalization system in a production environment with millions of active workers introduces a cascade of infrastructural decisions that shape the performance, reliability, and cost profile of the service. The high-level architecture typically consists of a data ingestion layer that streams real-time worker activity logs, a feature engineering pipeline that computes behavioral summary statistics over multiple time horizons, a model serving layer that hosts the personalized goal optimization engines, and an intervention delivery module that injects goal prompts into the worker-facing application via push notifications or in-app messages. Each of these components involves trade-offs that are rarely explored in academic literature but are decisive for operational viability.

Latency is a paramount concern. A goal recommendation that arrives five minutes after the optimal window for a lunch-time driving shift has lost much of its effectiveness. To achieve sub-second latency, the system must either compress the feature computation and model inference into tight time budgets or adopt a pre-computation strategy in which personalized goal policies are updated offline during low-traffic periods and cached for rapid retrieval. The caching approach introduces staleness, because the worker's context vector may change significantly between the update and the intervention moment. Engineers must therefore decide on the granularity of personalization updates, balancing freshness against computational cost. A hybrid design that combines batch processing for deep model retraining with a lightweight online layer for real-time adjustments—such as shifting the goal frame from gain to loss based on the latest session earnings—can reconcile these demands.

Data privacy and sovereignty represent another critical axis of trade-off. The richness of prospect-theoretic personalization grows with the availability of fine-grained behavioral data, including geolocation, temporal patterns of app usage, and earnings sensitivity. However, regulatory frameworks such as the General Data Protection Regulation and the California Consumer Privacy Act impose strict limits on data collection and cross-context profiling. Federated learning offers an architectural pathway to personalization without centralizing raw data, allowing a global model to be trained across distributed worker devices while keeping personal logs local [12]. Yet federated learning introduces communication overhead, heterogeneous hardware capabilities, and challenges in aggregating model updates from non-identically distributed data. The decision to adopt a federated architecture thus depends on the platform's risk posture regarding regulatory compliance and the technical capability to manage a distributed training infrastructure.

Robustness and fault tolerance are equally essential. A goal personalization system that systematically malfunctions—for example, recommending unattainably high goals to a subset of workers due to a feature pipeline error—can cause real financial and psychological harm. Therefore, the design must incorporate guardrails, such as hard limits on goal magnitude, anomaly detection on model outputs, and gradual rollout mechanisms with automatic rollback triggers. Canary deployments and staged exposure, in which new goal policies are first applied to a small, consenting cohort before wider release, mitigate the blast radius of errors. Additionally, the system must be resilient to adversarial behavior, such as workers attempting to game the goal-setting algorithm by strategically manipulating their activity patterns to receive easier goals. Multi-armed bandit algorithms with adversarial robustness guarantees can be incorporated to limit the impact of such gaming.

The scalability of the inference infrastructure hinges on the complexity of the underlying behavioral model. A simple logistic regression that maps a handful of features to a goal magnitude can be served at millions of queries per second with minimal hardware. In contrast, a full prospect-theoretic inverse reinforcement learning model that requires Monte Carlo sampling and optimization of a non-convex value function may demand dedicated GPU clusters and still fail to meet tight latency windows. System architects must therefore decide how much behavioral fidelity to sacrifice for the sake of throughput and cost-efficiency. Approximate Bayesian inference methods, such as variational autoencoders that learn a compressed latent representation of worker types, can strike a middle ground, enabling rich personalization at manageable computational expense [13].

5. Fairness, Governance, and Policy Implications

The introduction of algorithmic goal design into flexible labor markets raises profound questions of fairness, power asymmetry, and accountability. When a platform leverages prospect theory to optimize worker engagement, it is explicitly targeting cognitive biases that shape decision-making. While such interventions can be welfare-improving if they help workers overcome present bias and achieve savings goals, they also risk sliding into manipulation, especially when the platform's performance metrics are misaligned with worker well-being. The field experiment reported in [14] demonstrates that personalized earning goals can significantly increase gig workers' productivity, but the study also highlights the need to carefully calibrate goal specificity and difficulty to avoid unintended negative effects. This calibration problem is not purely technical; it is a governance challenge that requires transparent principles, external oversight, and mechanisms for worker voice.

Fairness in goal personalization comprises both distributive and procedural dimensions. Distributive fairness demands that the benefits and burdens of goal interventions be equitably shared across worker subgroups, preventing a scenario where loss-framed goals are disproportionately assigned to more loss-averse workers who may then work longer under stress, while gain-framed, less demanding goals are offered to others. Algorithmic fairness constraints, such as demographic parity or equality of opportunity, can be encoded into the optimization objective, but their application in the presence of prospect-theoretic utilities is nontrivial. The reference points and loss aversion parameters that drive goal response may be correlated with socioeconomic vulnerability, creating a tension between statistical accuracy and fairness [15]. Regularization techniques that bound the disparity in predicted labor supply response across protected groups can mitigate this, yet they may come at a cost in overall engagement lift that the platform's business model relies upon.

Procedural fairness concerns the transparency and contestability of goal-setting algorithms. Workers should be able to understand, at a high level, why a particular goal has been suggested and how it relates to their past behavior. Explainable artificial intelligence methods, such as Shapley value decompositions, can attribute the driving factors behind a goal recommendation [16], though communicating prospect-theoretic concepts like reference point adaptation in plain language remains a design challenge. Auditability is equally important; independent auditors must be able to reconstruct the decision pipeline and verify that the system complies with fairness commitments. This necessitates immutable logging of model versions, feature inputs, and recommendations, as well as the ability to replay historical queries.

The regulatory landscape is evolving rapidly. The European Union's Platform Work Directive introduces provisions on algorithmic management, requiring platforms to inform workers about automated decision-making systems and to conduct regular assessments of their impact on working conditions [17]. In the United States, regulatory attention has focused on worker classification and minimum earnings guarantees, but algorithmic goal-setting is likely to attract scrutiny under consumer protection frameworks that prohibit unfair or deceptive acts. Platform operators must proactively design governance mechanisms that exceed minimum compliance requirements, such as establishing independent ethics boards with worker representation and conducting algorithmic impact assessments that explicitly evaluate prospect-theoretic nudges for coercive potential. These governance structures should be complemented by worker-controlled settings that allow individuals to adjust the aggressiveness of goal framing, opt into or out of personalization, or select among different framing styles, thus restoring a degree of autonomy that is central to the promise of flexible work.

6. Deployment and Sustainability Considerations

Sustaining a prospect-theoretic goal personalization system over years of operation requires active management of model drift, engagement dynamics, and long-term worker outcomes. Concept drift occurs when the statistical relationships between worker features and goal responses change due to shifts in the platform's demand patterns, macroeconomic conditions, or the worker population composition. A model that was carefully tuned during a period of labor surplus may fail when the market tightens and workers' outside options improve, altering their reference points and loss sensitivity. Continuous monitoring pipelines that compare predicted versus observed labor supply response, combined with automated retraining triggers, are essential to maintain performance. However, retraining cycles must be designed with caution, because frequent model updates can introduce volatility in goal recommendations that confuse and frustrate workers, undermining trust.

The most pressing sustainability challenge is the potential for algorithmic goal systems to create engagement spirals that harm worker well-being in the long run. A myopic optimization that maximizes short-term platform throughput may progressively escalate goal difficulty as workers adapt, pushing them to exhaustion without detecting the deterioration in subjective well-being. Longitudinal indicators of burnout, churn, and self-reported satisfaction must be integrated into the reward function, transforming the problem from a single-objective bandit into a multi-objective optimization that balances productivity with retention and health [18]. Platforms that deploy goal personalization must invest in longitudinal studies that track cohorts of workers over multiple years, comparing outcomes under different goal design philosophies. Such studies can reveal whether prospect-theoretic personalization generates

genuine win-win scenarios or merely extracts short-term surplus at the expense of long-term human capital.

Sustainability also encompasses the environmental footprint of the machine learning infrastructure. The computational demands of continuously retraining personalized models across millions of users can be substantial, contributing to energy consumption and hardware depreciation. Techniques such as model distillation, sparse architectures, and event-triggered inference can reduce the carbon intensity of the system, aligning with broader corporate sustainability goals. From a systems engineering perspective, these optimizations must be balanced against the accuracy losses that accompany model compression, creating a new trade-off frontier between ecological sustainability and behavioral fidelity.

The co-evolution of worker behavior and the goal personalization algorithm creates a complex adaptive system that can exhibit emergent phenomena such as polarization of work patterns or the entrenchment of inequality. If the system learns that workers from higher-income neighborhoods are more responsive to certain goal frames and allocates more aggressive targets to them, it may inadvertently widen earnings disparities. Causal feedback loops of this nature require the integration of system dynamics tools into the model monitoring framework, enabling the detection of runaway feedback and the deployment of countermeasures such as fairness-aware exploration or deliberate injection of pro-equity goals [19]. The design of feedback control laws that stabilize the joint worker-platform system around an equitable equilibrium is a frontier challenge that bridges control theory, behavioral economics, and machine learning.

7. Conclusion

This paper has articulated a vision for integrating prospect theory and machine learning to personalize goal recommendations in flexible labor markets, moving from an analysis of cognitive foundations through system architecture to governance and deployment. We have argued that the behavioral richness of prospect theory—reference dependence, loss aversion, probability weighting, and dynamic reference point adaptation—provides an essential lens for understanding and predicting how gig workers respond to goal-setting interventions. Machine learning offers the computational machinery to infer these latent behavioral parameters and optimize goal framing at an individual level, transforming goal design from a blunt instrument into a precision tool. Yet the transition from theoretical possibility to operational reality is fraught with systems-level trade-offs. The choice of learning paradigm—contextual bandit, inverse reinforcement learning, or meta-learning—must be weighed against latency constraints, privacy requirements, and scalability demands. Fairness and governance cannot be afterthoughts; they must be engineered into the optimization objectives and auditing infrastructure from the outset. Deployment sustainability depends on managing concept drift, avoiding engagement spirals, and accounting for the co-evolution of worker and algorithm.

The path forward requires interdisciplinary collaboration among behavioral economists, systems engineers, machine learning researchers, legal scholars, and platform operators. A promising direction for future research is the development of benchmark environments and digital twins that simulate the dynamics of a flexible labor market under prospect-theoretic goal personalization, enabling rigorous testing of alternative architectures and governance protocols before real-world deployment [20]. Another critical area is the design of worker-centric data trusts that give individuals collective bargaining power over the behavioral data that fuels personalization, thereby rebalancing the power asymmetry between platform and worker. As algorithmic management becomes increasingly pervasive, the academic

community carries a responsibility to ensure that the sophisticated tools of behavioral science and artificial intelligence are deployed not merely to optimize platform metrics, but to enhance the dignity, autonomy, and economic security of the people who power the flexible labor economy.

References

1. Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263–291.

2. Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5(4), 297–323.

3. Schmidt, U., Starmer, C., & Sugden, R. (2008). Third-generation prospect theory. Journal of Risk and Uncertainty, 36(3), 203–223.

4. Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. Proceedings of the 19th International Conference on World Wide Web, 661–670.

5. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press.

6. Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. American Psychologist, 57(9), 705–717.

7. Read, D., Loewenstein, G., & Rabin, M. (1999). Choice bracketing. Journal of Risk and Uncertainty, 19(1-3), 171–197.

8. Camerer, C. F., Babcock, L., Loewenstein, G., & Thaler, R. H. (1997). Labor supply of New York City cabdrivers: One day at a time. The Quarterly Journal of Economics, 112(2), 407–441.

9. Harman, J. L., & Marley, A. A. J. (2022). A Bayesian approach to inverse reinforcement learning for prospect theory. Decision, 9(2), 121–136.

10. Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. Proceedings of the 34th International Conference on Machine Learning, 1126–1135.

11. Swaminathan, A., & Joachims, T. (2015). Batch learning from logged bandit feedback through counterfactual risk minimization. Journal of Machine Learning Research, 16, 1731–1755.

12. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1–210.

13. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. Proceedings of the 2nd International Conference on Learning Representations.

14. Min, X., Chi, W., Hu, X., & Ye, Q. (2024). Set a goal for yourself? A model and field experiment with gig workers. Production and Operations Management, 33(1), 205-224.

15. 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.

16. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems, 4768–4777.

17. European Commission. (2024). Directive on improving working conditions in platform work. Official Journal of the European Union, L 167.

18. Agrawal, A., Gans, J., & Goldfarb, A. (2019). Prediction machines: The simple economics of artificial intelligence. Harvard Business Review Press.

19. Doroudi, S., Thomas, P. S., & Brunskill, E. (2017). Importance sampling for fair policy selection. Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence.

20. Reinecke, P., & Bernstein, A. (2023). Digital twins for socio-technical systems: A research agenda. IEEE Software, 40(4), 42–50.