Measurement, Assumptions, and Inference in Deployed Vision Systems

David Woollard

Overview

My research focuses on understanding what can be validly inferred about human behavior from computer vision systems deployed in real-world environments. Rather than treating vision models as abstract predictors evaluated solely through benchmark performance, I approach deployed sensing systems as *scientific instruments*: they make certain phenomena observable under specific constraints, while rendering others latent or fundamentally inaccessible. The central question that motivates my work is not simply how accurately models perform, but how the assumptions embedded in sensing, labeling, and representation constrain the claims that can be supported.

This perspective has emerged through sustained engagement with vision systems that must operate continuously, degrade gracefully, and remain interpretable under real operational conditions, including my current work at Standard AI. Building systems that must function over long time horizons has shaped my confidence in these views: deployment acts as a forcing function that exposes hidden assumptions, brittle representations, and overconfident claims in ways that offline experimentation often does not.

Deployed Vision Systems as Instruments

In contrast to controlled laboratory datasets, deployed vision systems operate under persistent and irreducible constraints: fixed camera geometries, occlusions, missing data, privacy requirements, and long-term operational drift. These conditions shape not only model performance, but also what aspects of human behavior are observable in the first place.

Many commonly reported outputs—tracks, poses, dwell times, or behavioral segments—are best understood as *proxy measurements*. They are neither direct observations of intent nor complete descriptions of action, but partial, instrument-dependent representations. As Bowker and Star argue in their study of classification systems, such representations embed assumptions that become invisible over time, even as they exert material consequences on downstream analysis and decision-making [1]. Treating proxy measurements as ground truth risks overstating the scope of inference, particularly when evaluation is decoupled from the conditions under which the data were produced.

Behavioral Representation and Structure

A second thread of my research concerns how human behavior is represented once it has been rendered observable. Continuous trajectories, discrete events, and learned behavioral tokens each impose different inductive biases and support different classes of questions.

Early models of pedestrian dynamics, such as the social force formulation, introduced a now-standard abstraction of pedestrian behavior as continuous trajectories shaped by interactions and environmental constraints [2]. My interest is not in the specific mechanics of such models, but in how representational choices—whether physical, statistical, or learned—mediate what aspects of behavior become salient and what claims they can support. Rather than advocating for a single representation, I study how these choices interact with downstream analyses, influencing interpretability, robustness, and the propagation of uncertainty.

Prediction, Evaluation, and Limits of Inference

Prediction in real-world behavioral systems raises distinct challenges from those encountered in benchmark settings. Predictive models are often evaluated on short horizons, homogeneous contexts, or artificially balanced datasets, obscuring the operational meaning of prediction under deployment conditions.

Following classical results in statistical learning theory, generalization must be understood as contingent on the alignment between model class, data-generating process, and evaluation regime [3]. In deployed systems, non-stationarity, partial observability, and class imbalance are the norm rather than the exception. My work examines prediction as a conditional statement: given a particular sensing configuration, representation, and operational context, what forms of future behavior can be anticipated with calibrated uncertainty—and where prediction should explicitly abstain.

Broader Implications

While retail environments provide a practical and scalable setting for this research, the implications extend beyond any single domain. Structured, repeatable physical spaces with sustained human activity offer a unique substrate for studying behavior at scale, informing simulation, robotics, and broader efforts toward physical world modeling.

At the same time, this work argues for epistemic restraint. Richer data and more powerful models do not eliminate the need to reason carefully about what is observable, what is inferred, and what remains unknowable. Progress in physical AI will depend not only on model capacity, but on clarity about the assumptions that link data to claims.

Relationship to My Work

My published research should be read as a series of empirical probes into these questions, each grounded in deployed systems and motivated by specific measurement challenges. Collectively, they reflect an effort to align technical innovation with principled inference, ensuring that advances in perception translate into knowledge that is both useful and defensible.

This document is intended as a research perspective rather than a comprehensive survey or project proposal. Updated versions may reflect evolving empirical findings, but the emphasis on measurement, assumptions, and inference is intended to remain stable over time.

References

[1] G. C. Bowker and S. L. Star. *Sorting Things Out: Classification and Its Consequences*. MIT Press, 1999.

[2] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, 1995.

[3] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.