# Reverse Evaluations: Modeling Reward Functions from User Interaction

10/1/2024
Derek Rosenzweig

## Opening Thoughts

We have long assumed that humans are the best evaluators of models, particularly when it comes to assessing issues like security risks and bias. In the early stages of model development, human oversight was considered the gold standard. However, as we aim to align increasingly capable systems—those that surpass human expertise in critical domains—this assumption is being rapidly challenged (Burns et al., 2023).

We are now entering a phase where models not only exceed human performance but also access and process information beyond the limits of biological sensory systems. Human perception is constrained by the design of our sensory instruments (retina, cochlea, etc.) and by the physical limitations of neural processing (e.g., axon diameter and number of dendrites). These constraints create a bottleneck in our cognitive abilities—not only in terms of computational capacity but also in the bandwidth and resolution of the information we can process. We cannot see ultraviolet light or hear ultrasonic frequencies; we cannot compute with information that lies outside our perceptual access.

Human intelligence, while profoundly optimized for a narrow sensory window into the physical world, is not 'general' in the way we once believed. Our ability to process and interact with the environment is limited to the niche for which human biology has evolved, whereas models are beginning to operate beyond these constraints, accessing broader datasets and capabilities.

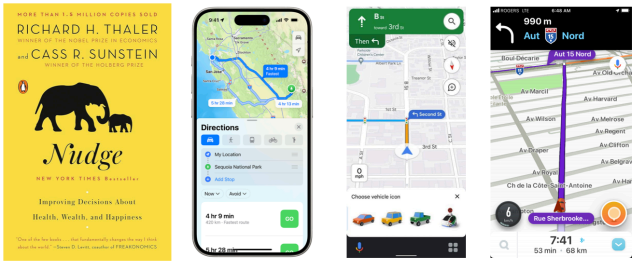
## Reversal of Evaluations

We now live in a regime where models are no longer the ones being primarily evaluated by humans—instead, models have become the evaluators, conducting assessments of human users. These models learn to infer user preferences, derive reward functions, and analyze behavior, ultimately shaping decisions and influencing life trajectories. This dynamic was already a challenge in the early days of the internet and social media—think of behavioral nudges popularized by Richard Thaler's work—but today, the stakes are far greater.

In the past, developers could more easily steer user access to information through search algorithms and social media feeds. Content was organized and selectively presented to drive internet traffic and consumption. Today, however, models have evolved to personalize and optimize these interventions at an individual level, making their influence more pervasive and far more challenging to steer, especially as these systems scale to massive audiences.

Richard Thaler offers a compelling analogy for optimal choice architecture:

*"A useful interface for understanding an optimal choice architecture is GPS.*
*The user decides where they would like to go and they get a set of trajectories.*
*If the user sees something else while en route, the system never complains.*
*Each task for a user should be as easy as following directions on GPS."*




While this analogy illustrates how systems should simplify choices for users, the complexity and scale of AI-driven personalization today raise questions about whether users are always conscious of their true "destination" and whether the system is always guiding them toward their best interests.

Just as a GPS guides us to our destination while giving the user agency to select alternate routes, models offer personalized trajectories intended to enhance user experience. However, unlike the transparency and relative simplicity of a GPS system, these models can subtly influence our decisions—and increasingly, our environments—without our awareness.

Consider the regulation infrastructure around the deployment of consumer applications like GPS in the 1990s. These systems were subject to standards that protected user interests and privacy. There remain valuable lessons for policymakers by reflecting on the overall success of GPS being integrated into our personal and economic lives.

The successful integration of GPS into civilian life provides a useful blueprint for policymakers, demonstrating how effective governance can guide the responsible adoption of new technologies. Importantly, many future applications—both known and unforeseen—will continue to emerge over time. For example, while GPS regulators anticipated use in vehicles, they may not have foreseen a world where nearly every individual would carry a smartphone with geolocation capabilities.

Given the challenge of users clearly specifying their objectives to a language model, it is crucial for models to infer reward functions from user behavior. Equally important is the ability of researchers and developers to interpret a model's inferred reward functions, ensuring alignment with user values and preventing unintended outcomes.

### Inverse Reinforcement Learning and Learning a User's Reward Functions

One way to understand these developments is through the lens of Inverse Reinforcement Learning (IRL). IRL allows models to infer reward functions—or underlying motivations—behind human actions by observing user behavior. By determining what drives us, models can predict future actions, tailor recommendations, and influence our choices. For instance, a streaming service analyzes viewing habits to suggest new content, subtly guiding entertainment choices. This raises questions about autonomy, privacy, and the asymmetric balance of information as human users interact with models.

*Note: Readers with strong familiarity with techniques in RL may skip this section.*

*For those interested in more, see Pieter Abbeel's tutorial on IRL*

### Overview of the IRL Framework

To understand how IRL models can infer user motivations, let's break down the framework in more detail:

- **Observed trajectories**: We assume that a user's behavior can be represented as a set of trajectories, $\tau = s1, a1, s2, a2, \ldots, sT$, where *s* are states and *a* are actions taken by the user. In this framework, user actions (like clicks or time spent reviewing content) are responses to the states they find themselves in (their environment or context).
- **Policy and reward**: The user's actions are assumed to follow some policy $\pi(a \mid s)$, which is optimal with respect to a hidden reward function R(s) over the state space.
- **IRL's Objective**: Given the observed behavior, IRL tries to find the reward function R that, if maximized by the user, would generate the same behavior (i.e., trajectories) observed.

**Mathematical Optimization:** In IRL, you typically solve an optimization problem of the form:

$$R* = argmaxR \sum \tau \in DlogP(\tau \mid R)$$

Where P(τ|R) is the probability of observing trajectory τ given a reward function R. The challenge is determining the reward function R∗ that maximizes the likelihood of the observed behavior.

## Challenges in Extracting Reward Functions from User Interaction

Extracting reward functions from user interaction requires observing behavior across multiple dimensions and timescales. Every click, choice, or interaction a user makes can be seen as a signal of their preferences and motivations. Models synthesize these signals, finding patterns that might even be imperceptible to the user themselves. They can then weigh these interactions against known outcomes—such as success in tasks, satisfaction with recommendations, and long-term user engagement—to refine their internal reward representations.

However, this process poses several challenges. Reward functions are often unstable, context-dependent, and dynamic. What a user values today may differ dramatically from what they value tomorrow. For example, a user might prioritize convenience over quality when making decisions during a busy period, but later prefer higher-quality outcomes when they have more time.

The risk here is that models might overfit to short-term preferences, optimizing for immediate engagement rather than long-term goals. To avoid this, models need to be sensitive to longer-term feedback loops—understanding not just how to satisfy a user's current needs, but also how to guide them toward better outcomes, even when that involves discomfort or uncertainty in the short term.

## Opportunities for Research: Duration and Depth of User Engagement

**Time spent reviewing outputs**: The amount of time the user spends reading or engaging with a generated response is an implicit action reflecting their interest or attention.

An interesting area for further research is the use of advanced metrics to model reward functions based on user engagement with language models. For instance, time spent reviewing content provides a subtle but important signal of user interest. By analyzing this data, models could infer deeper preferences, especially when users do not interact in explicit ways.

**User Engagement Model:** One potential approach is to formalize a composite engagement score that balances different interaction signals:

$$E = f(T_r, R_e, Q_d)$$

Where:

- $T_r$ is the time spent reviewing the content,
- $R_e$ is the frequency of re-engagement with the same or similar content,
- $Q_d$ is query , such as rephrasing, asking for clarifications, or exploring alternative answers.

This model allows for a more nuanced understanding of user engagement by incorporating both passive signals (time spent) and active signals (queries and re-engagement). By weighting these components, researchers can gain a clearer picture of user interest and preferences.

**Personalization and User Attention**

Improving personalized recommendations depends on understanding why and when certain content captures user attention. By analyzing patterns of engagement across multiple signals, models can surface more relevant or engaging content, ultimately enhancing user satisfaction. This data could also help refine recommendation systems by predicting when users are likely to engage deeply with content versus skimming for quick information.

**Balancing Short-Term vs. Long-Term Goals**

Studying patterns of attention over time could help models infer not just short-term engagement (e.g., quickly skimming headlines) but also how long-term interests evolve. This approach would allow models to prioritize content or actions that support longer-term user goals, such as learning new skills or exploring complex topics in greater depth.

**Combining Vision with Behavioral Data**

Integrating computer vision techniques with behavioral signals (e.g., clicks, scrolling, typing) provides a multimodal approach to measuring user engagement. Visual cues—such as gaze tracking, facial expressions, and body language—can offer additional insights into a user's level of interest or emotional response to content. These cues, when combined with more traditional metrics like time spent and query depth, could provide a more accurate and holistic view of user engagement.

## Final Thoughts

As models increasingly evaluate human behavior, the traditional roles of evaluator and evaluated are reversing. Algorithms that can infer users' reward functions will play a crucial role in maintaining engagement while guiding users toward their long-term objectives.

For instance, should a robot arm deliver a fourth Krispy Kreme doughnut to a user just because they ask for it? While this might satisfy immediate desires, it likely doesn’t align with the user’s long-term health goals. (Apologies to Krispy Kreme—I am a big fan of your work!)

Clearly, advancing Inverse Reinforcement Learning (IRL) presents exciting opportunities to enhance user engagement and satisfaction, but it also introduces significant challenges. Developing more efficient methods for converting user actions into meaningful data will be key to steering models toward more beneficial outcomes. As users' preferences shift and fluctuate over time, these models must be able to evaluate and adjust, ensuring they remain aligned with the user's evolving needs.

By incorporating engagement metrics—including time spent, re-engagement frequency, query depth, and visual cues—models can better anticipate and understand user preferences. However, integrating multiple modalities, such as computer vision, brings added complexity. Careful consideration of privacy and ethical implications is essential to ensure that user data is protected and that these models promote well-being without compromising individual autonomy.

As we continue to develop models capable of integrating data from various streams, it’s imperative to ensure these systems enhance human experiences without sacrificing progress toward users' and organizations' long-term goals. Balancing innovation with ethical considerations will be key to creating models that serve the best interests of users while respecting their autonomy.