

Literature Review: An Abstract View of Behavioral Predictions in Artificial Intelligence

by Christopher Cook (May 1, 2020)

INTRODUCTION

How do we define freewill? Do we assume that an actor has the ability to make their own decisions outside of external influences, or are the current limits of predictability an illusion of free choice? Contemporary machine learning algorithms inadvertently attempt to fortify the latter hypothesis by creating predictive models that can determine an actor's choices. These algorithms cannot only answer philosophical questions, but they have real world applications. Advertisers can use these for targeted marketing, mutated cell growth in cancer patients can be easily predicted for targeted radiation treatment, and human behavior in mentally ill individuals can be better understood and acknowledged.

CURRENT RESEARCH

The Two System Model

Reinforcement Learning (RL) and Inverse Reinforcement Learning (IRL) algorithms are two such algorithms that offer potential insight into cause and effect in an actor's behavior [1]. Alexander Peysakhovich [1] notes that RL looks at an actor's end goals to predict the actor's choices, whereas IRL looks at the actor's choices to determine their end goals. While both are important to building predictive models in behavioral analytics, Peysakhovich focuses heavily on IRL and there is a well-justified reason for this.

Intuitively, it is easier to predict the choices of an actor given their end goal, but it is harder to figure out the end goal given certain pre-defined choices, especially when we add an undefined temporal dimension. Peysakhovich [1] shows this concept to be the caveat to the 2S model. The 2S model, or two system model, categorizes the actor's behavior into the impulsive self for immediate reward and the patient self for long-term reward. He uses an analogy with donuts and kale to help the readers understand this issue. Essentially, an actor has the option to eat a donut and gain an immediate reward, or eat kale and gain a future reward [1]. However, this limits each of the two selves to mutually exclusive rationalities.

Rationality of the Impulsive and Patient Selves

Peysakhovich's [1] approach alludes to the impulsive self as irrational and the patient self as rational, thereby creating a barrier between the two systems. Drew Fudenberg and David Levine [2] offer a more in-depth

look into why these two systems are not mutually exclusive. They mention this specifically in their example on health club memberships. If the patient self sees the long-term benefit of going to the club, it directly influences the actor to not cancel the membership, despite the final outcome. This means the actor may never go to the health club, yet continue to pay for the membership [2]. This is an irrationality, because if the predictive model sees that the actor is paying for the membership, but not using the membership, the desired end goal may not match the actual end goal, leaving room for higher error rates in IRL models. These can especially be troublesome if we do not consider the uncertainty level associated with the patient self's long-term predictions.

Temporal Cognition

These errors are most likely correlated with a hyperbolic or quasi-hyperbolic time function [1,2,3]. Conceptually, an actor's behavior can be affected by the amount of time until the reward, and their behavior may also change overtime leading to a hyperbolic or quasi-hyperbolic function of temporal cognition.[3] While we did see this with Peysakhovich's [1] work, he simplified the time parameter to a binary system where the initial decision was the only decision—reward now or reward later. Fudenberg and Levine [2] also discussed this, in detail, but considered the wavering long-term decision to be atypical and discretely irrational. However, Evans, Stuhlmuller, and Goodman [3] find this behavior normative and analyze its potential predictability. They see an actor's behavior as a complex permutation of variables that each have a unique value [3]. Time and uncertainty play large roles in the determination factor of an actor's choices and main goal. A notorious example utilized in papers on RL and behavioral predictions is the time value problem of behavioral economics, as seen in [2,3]. Evans, Stuhlmuller, and Goodman precisely summarize this thought experiment by stating that "The agent takes \$100 now over \$110 tomorrow, but would prefer to take \$110 after 31 days to \$100 after 30 days" [3, p. 2].

Time can create a substantial paradigm shift in the actor's choices, as shown in the previous example, but becomes less impactful when other variables are introduced. Probabilistic inferences arise from varying knowledge of an actor's choices [3]. Up until this point, it has been assumed that the actor had full

knowledge of their decisions, but this is an unrealistic expectation. An actor's behavior can be influenced by heuristics and naivety, which leads to probabilistic inferences. This is where the actor bases their decision off of preconceived biases—or they just guess what they should do next [3].

Evolutionary Intuition and Deliberation

The predictive models that see the largest impact from this attribute are those that involve multiple actors. In Game Theory, cooperative games allow two or more actors to cooperate for the maximized reward or they defect for a chance at a larger, maximized individual reward [4]. Here, uncertainty is at its highest, because one actor may not always know the future choices of another actor. This level of uncertainty could lead to irrational or possibly unpredictable behavior. In contradiction, Adam Bear and David Rand [4] believe that time increases the predictability of such behavior. They hypothesize that an actor with higher levels of uncertainty overtime will align with evolutionary principles. Fundamentally, there is an equilibrium between deliberative costs and deliberative behavior, which means that decision making becomes more difficult, so the actor reverts to intuition [4]. This can help determine if the actor will cooperate or defect against the other players. The following figure from [4] shows the correlations between deliberation, intuition, deliberative costs, and time:

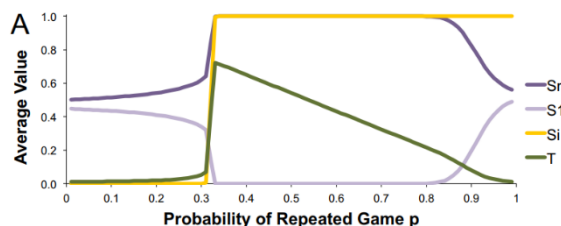


Figure 1: Probability of Cooperation [4]

This figure from Bear and Rand's work [4] can be confusing when taken out of context, but this imagery holds a lot of weight and encompasses one of their main theories. Let us assume there is an actor that has the chance to cooperate or defect against another actor in a one-off game that has a chance to be repeated with the same actors. All four lines correlate to the probability of cooperation with 0 being no chance of cooperation and 1.0 being guaranteed cooperation. T is the cost of deliberation and Si is the probability of intuitive cooperation. Furthermore, S1 corresponds to the deliberative behavior associated with one-off games and Sr corresponds to the deliberative behavior associated with repeated games. Even though deliberative behavior has a large predictability factor,

guaranteed one-off games and guaranteed repeatable games make deliberative behavior unpredictable. Alternatively, the intuitive behavior is only unpredictable when there is approximately a 32% chance of a one-off game repeating. Note that there is also a correlation between the accelerative increase in deliberative costs and intuitive cooperation as stated earlier [4].

CONCLUSION

Predictive modeling offers a view into future outcomes but leaves room for improvement. Many factors come into play when looking at behavioral analytics, and without any scientific insights, some choices appear random and unpredictable. Unknown continuous variables with parametric bounds embody the smoking gun between unpredictable and predictable decision making. Variables, such as the level of deliberation over intuition and the hyperbolic value of time vs uncertainty, mitigate the issues associated with unpredictable behavior. This is where these papers agree; newly discovered variables increase the accuracy of behavioral predictions.

Yet, contradicting arguments about the relationship between unpredictable and irrational behavior still prevail. The abstract hypotheses associated with cognitive interpretations hold inconsistencies that need further study. Specifically, the language in which researchers use to discuss choice actions vary from researcher to researcher. While this does not invalidate their analytical work, it shows how little we still know about choices and freewill. A better understanding of cognitive functions will give insight into potential gaps in predictive modeling and may one day lead to more robust and efficient IRL models.

REFERENCES

- [1] A. Peysakhovich, "Reinforcement Learning and Inverse Reinforcement Learning with System 1 and System 2," *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 409–415, 2019.
- [2] D. Fudenberg and D. K. Levine, "A Dual-Self Model of Impulse Control," *American Economic Review*, vol. 96, no. 5, pp. 1449–1476, 2006.
- [3] O. Evans, A. Stuhlmüller, and N. D. Goodman, "Learning the preferences of ignorant, inconsistent agents," *AAAI*, pp. 323–329, 2016.
- [4] A. Bear and D. G. Rand, "Intuition, deliberation, and the evolution of cooperation," *Proceedings of the National Academy of Sciences*, vol. 113, no. 4, pp. 936–941, 2016.