Galit Shmueli, PhD
Tsing Hua Distinguished Professor and Institute Director
Institute of Service Science, College of Technology Management
National Tsing Hua University, Taiwan

"Improving" Prediction of Human Behavior Using Behavior Modification

Large internet platforms that collect behavioral big data predict user behavior for internal purposes and for third parties (advertisers, insurers, security forces, political consulting firms) who utilize the predictions for personalization, targeting, and other decision-making. Data science researchers design algorithms, models, and approaches to improve prediction. Prediction is also improved with larger and richer data. We describe how, beyond improving algorithms and data, platforms can stealthily achieve better prediction accuracy by “pushing” users' outcomes towards their predicted values, using behavior modification techniques, thereby demonstrating more certain predictions. The better the platform can make users conform to their predicted outcomes, the more it can boast its predictive accuracy. Hence, platforms are incentivized to “make predictions true”. Such apparent “improved” prediction can unintentionally result from employing reinforcement learning algorithms that combine prediction and behavior modification. This strategy is absent from the machine learning and statistics literature. Investigating its properties requires integrating causal with predictive notation. To this end, we incorporate Pearl's causal do(.) operator into the predictive vocabulary. We then decompose the expected prediction error given behavior modification, and identify the components impacting predictive power. Our derivation elucidates implications of such behavior modification to data scientists, platforms, their customers, and the humans whose behavior is manipulated. Behavior modification can make users' behavior more predictable and even more homogeneous; yet this apparent predictability might not generalize when customers use predictions in practice. Outcomes pushed towards their predictions can be at odds with customers' intentions, and harmful to manipulated users.