HARVARD | BUSINESS | SCHOOL

# Soul and Machine (Learning)


Davide Proserpio
John R. Hauser
Xiao Liu
Tomomichi Amano
Alex Burnap
Tong Guo
Dokyun Lee
Randall Lewis
Kanishka Misra
Eric Schwarz
Artem Timoshenko
Lilei Xu
Hema Yoganarasimhan



**Working Paper 20-036**

# Soul and Machine (Learning)

Davide Proserpio
University of Southern California

John R. Hauser
Massachusetts Institute of Technology

Xiao Liu
New York University

Tomomichi Amano
Harvard Business School

Alex Burnap
Massachusetts Institute of Technology

Tong Guo
Duke University

Dokyun Lee
Carnegie Mellon University

Randall Lewis
Independent

Kanishka Misra
University of California San Diego

Eric Schwarz
University of Michigan

Artem Timoshenko
Northwestern University

Lilei Xu
Airbnb

Hema Yoganarasimhan
University of Washington

**Working Paper 20-036**



Funding for this research was provided in part by Harvard Business School.

# Soul and Machine (Learning)

Davide Proserpio, John R. Hauser, Xiao Liu, Tomomichi Amano, Alex Burnap, Tong Guo,

Dokyun Lee, Randall Lewis, Kanishka Misra, Eric Schwarz, Artem Timoshenko, Lilei Xu,

Hema Yoganarasimhan[*†]

[*] Authors affiliations: USC Marshall School of Business, MIT Sloan School of Management, NYU Stern School of Business, Harvard Business School, MIT Sloan School of Management, Duke Fuqua School of Business, CMU Tepper School of Business, Independent, UCSD Rady School of Management, University of Michigan Ross School of Business, Northwestern University Kellogg School of Management, Airbnb, UW Foster School of Business.

[†] Corresponding author: Davide Proserpio, email: proserpi@marshall.usc.edu

**Abstract**

Machine learning is bringing us self-driving cars, improved medical diagnostics and machine translation, but can it improve marketing decisions? It can. Machine learning models predict extremely well, are scalable to "big data," and are a natural fit to rich media such as text, images, audio, and video. Examples include identification of customer needs from online data, accurate prediction of consumer response to advertising, personalized pricing, and product recommendations. But without a soul, the applications of machine learning are limited. Consumer behavior and competitive strategies are nuanced and richly described by formal theory. To learn across applications, to be accurate for "what-if" and "but-for" applications, and to advance knowledge, machine learning needs theory and a soul. The brightest future is based on the synergy of what the machine can do well and what humans do well. We provide examples and predictions for the future.

# 1 The soul and machine framework

Marketing studies how consumers react to complex stimuli to make purchase decisions and how firms organize economic resources for profit-maximization. In today's information-rich environment, consumers rely not only on their personal experience, but also on shared experiences provided by other consumers and recommendations by the firms. Firms use real-time news and sensor signals combined with predictions of consumer response to design policies and choose the best tactics. Accurately predicting consumer reactions and competitor responses to marketing strategies remains a fundamental challenge.

Machine learning has in recent years made significant advances. Today we see progress in areas such as self-driving cars, automated conversational agents, medical diagnostics, machine translation, and financial fraud detection. Marketing practice has already benefited from many of these advances. Firms of all sizes employ production-level machine learning systems to improve advertising campaigns, recommendation engines, and assortment decisions. Marketing science as an academic field has likewise begun to leverage machine learning approaches to propose new powerful applications and insights. In this paper, we propose to step back and ask how can we best integrate machine learning to solve previously untenable marketing problems facing real companies?

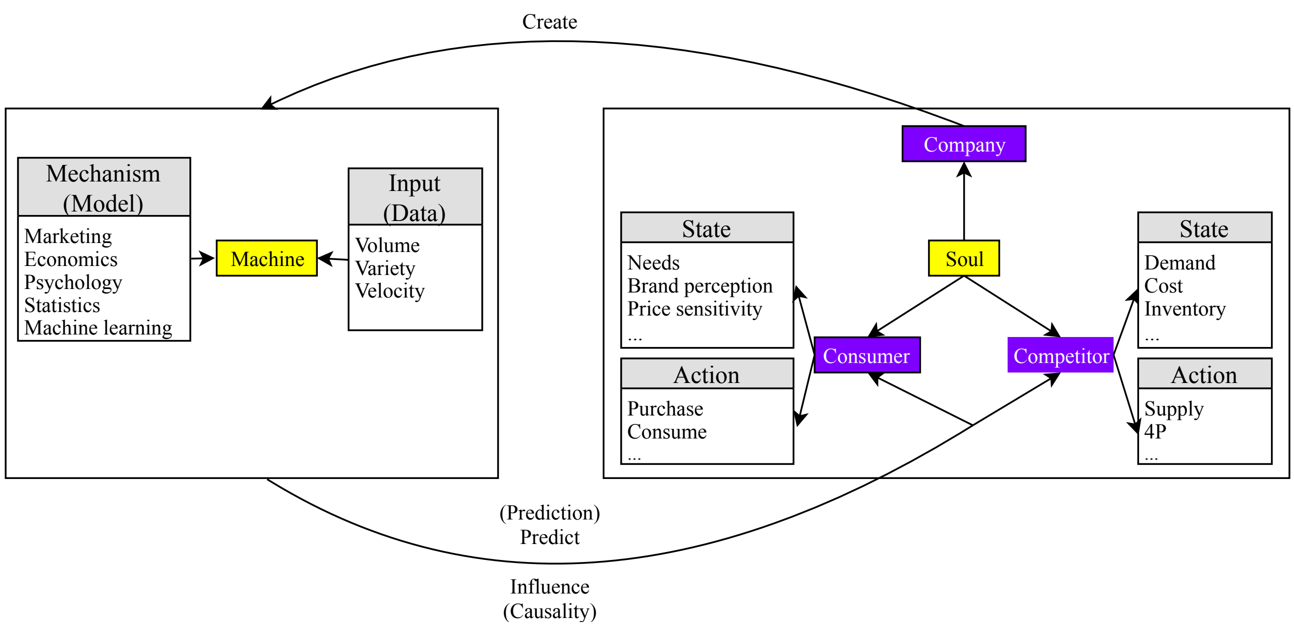
As a motivating example, marketing scientists have long dreamed of the “virtual market” - a machine representation of customers and competitors that could be queried as one would do with traditional market research. The intelligent machine would collect the data it needs to answer strategic questions and guide the managerial decisions. For example, if *Tantalizing Startup* proposes a new widget, the machine would use its growing database to identify (1) key customer groups interested in the widget, (2) key customer needs to improve the widget, (3) initial forecasts of sales of an improved widget and (4) competitors' moves. With these data, *Tantalizing Startup* could redesign the widget for consumer acceptance, make GO/NOGO decisions, and set an initial marketing strategy. *Tantalizing Startup* might still wish to query consumers and competitors with more traditional market research, but such research would be more directed following the prior provided by the machine.

Today's machine learning has offered better progress towards goals such as virtual markets than ever before, but machine learning is still evolving, and the integration of marketing science with machine learning concepts is still in its infancy. While we can query our machines, our machine representations do not yet encode the "soul" embedded in the marketing researcher, the consumer, the firm, and the market as a whole. The "soul" is our human intuition, our scientific expertise and our institutional knowledge that can define the questions and improve "machine" decisions about the market. For instance, the "soul" can guide the "machine" to account for irrational consumer choice behavior, violations of competitive markets, or optimality of firm-level strategic decisions—in short, the bread-and-butter of marketing science.

The "soul and machine" is the integration of marketing and machine learning that explicitly accounts for relationships among the company, the customers, competition, and the machine (Figure 1). The soul retains the judgment and intuition of marketing decision-makers; the machine retains data-driven patterns that represent marketing concepts such as "virtual consumers." "Machine" can process available information at scale and identify complicated patterns not perceivable by humans.

However, “Machine” cannot fully capture the soul of the consumer, whose wants and needs, powers of reason, biases and perceptions, and social relationships all influence what they will buy. The machine’s virtual customers capture as much as feasible about the consumer's soul; marketing scientists use theory, experience, and insight to augment the machine, to predict consumer preferences, and to uncover tactics and strategies to influence consumer decisions. Similarly, “Machine” cannot capture the managerial goals, and can only provide guidance for well-defined problems. Neoclassic economics assumes firms make fully rational decisions via complex optimizations, behavioral economics has documented managers’ bounded rationality, such as overconfidence, inattention, and loss aversion. But no model of the world can ever be complete.

Figure 1: The Soul and the Machine Framework




To predict customer wants, needs, and behavior and, when feasible, competitors' response, the marketing researcher feeds the machine with data and models. The data come in structured and unstructured formats, in enormous volume and at increasing frequency. Such data are now ubiquitous and readily available: Consumers generate prodigious unstructured data that is shared online through social media and review platforms. They made decisions across categories, over time, and in a varying social context. Firms develop data infrastructures to record their own and competitive actions. The machine uses visual, auditory, and textual data across many platforms. Recent advances in computer science and engineering, provide the means to peer into these data, to peer into the soul of consumers and competitors. The ability to analyze and understand unstructured data is leading to a quantum leap in our ability to understand, influence, and better the lives of consumers.

Data alone are not sufficient. The machine requires a mechanism to transform raw data into actionable insights. Extant theory in marketing, economics, psychology, statistics, engineering, design science, and machine learning lay the foundation of the mechanisms. Historically, marketing scientists relied on consumer choice theories to predict demand and promotion sensitivity as well as equilibrium system of equations to estimate competitor's pricing response in simplified markets. Machine learning improves predictions with traditional data, but the improvements are slight and, often at the expense of interpretability. Big data often hinders the capability of traditional models because such models are not sufficiently scalable nor flexible for unstructured data. Machine learning, including natural language processing, computer vision, and high-dimensional statistics, works naturally with these data and enable marketing scientists to glimpse into the consumer's and competitor's souls. Traditional survey-based research will always have a role, but that role is being diminished.

The soul and machine are synergistically iterative. The machine predicts and influences consumers and competitors while consumers and competitors produce more data to be fed into the machine. The outputs of the machine analysis are interpreted by the managers who turn insights and recommendations into actions. Among the many current applications are the ability to track a brand's image using UGC alone, customize marketing communications to each consumer in each context, model and understand the aesthetics of a product, or recommend products that enable consumers to explore their own preferences. But the machine is still in its infancy and lacks many important features such as interpretability and auditability. We argue that big changes are coming and many more applications of the machine will be available soon. We start by discussing what the machine and the soul can now do, and then move to the future. Marketing science theory and engineering, combined with machine learning capabilities, is changing the way in which we think about marketing.

# 2 An example

Timoshenko and Hauser (2019) propose a machine learning approach to leverage user-generated content (UGC) to identify customer needs. Consumers (the soul) have needs and wants, which are abstract context-dependent statements describing the benefits, in consumers' own words, that they seek to obtain from a product or service. Consumers generate content, such as product reviews, to express their needs and wants. Companies seek to better segment the market, identify strategic dimensions for differentiation, and make efficient marketing decisions by uncovering customer needs from the UGC. The machine is a deep-neural-network-based machine-learning model, which identifies relevant content and removes redundancy from a large UGC corpus in a cost-effective and scalable manner. But the human analysts are key players. Professional analysts manually review the selected, informative content and formulate customer needs. Manual review is necessary because customer needs are highly context-dependent and semantically abstract. Natural language processing is not yet ready to find the nuggets of customer needs in the identified sentences. The hybrid machine learning approach allows companies to replace traditional interviews and focus groups. The authors provide several successful examples and more are being created monthly.

# 3 Where are we going? What will Soul and Machine be like in the future?

The rise and popularity of machine learning methods and algorithms are providing marketers with a variety of new opportunities that will lead to new, and often breakthrough, applications. We provide an outlook on where marketing and machine learning are headed.

## 3.1 The soul

### 3.1.1 Going virtual

With the rise in popularity of machine learning and new technology, we predict a virtual future for marketing. The virtual customer, a longtime dream for marketing, is coming. But to use the virtual customer, companies must understand and model how consumers shape markets along with limitations of a virtual customer.

Economically successful firms depend on their ability to profitably develop new products and services that satisfy consumer demand. Implicit in this statement is the assumption that new products and services can be feasibly developed and delivered by the firm. But feasibility is governed by both hard and soft constraints that are scattered across the firm's divisions.

For example, the design of a new automobile has hard constraints on engineering, such as governmentally-regulated crashworthiness standards, and soft constraints on industrial design, such as compatibility with brand recognition and the human intuition that related automotive models come from the same "family." This generates coordination challenges across actors within the firm and often requires trade-offs between different decisions (good aesthetics does not always translate to good aerodynamics, for example).

Marketing is no stranger here. Good marketing decisions require understanding considerations across the firm positioning, brand strategy, and promotions. All of these can be ineffective if engineering cannot create the product, or manufacturing and supply bottleneck production. While these coordination challenges are partially ameliorated by processes such as "stage-gates" and cross-functional teams, the constant back-and-forth between intertwined actors still accounts for the majority of overall development time.

Big and new data sources offer new opportunities to change and improve this process. Marketing analytics has already advanced how traditional marketing decisions are made, but combining the soul and the machine will lead to greater integration of marketing decisions across all actors (silos) in the firm. Transfer learning, a machine learning approach that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem, provides one potential answer. For example, new product development will rely more and more on virtual tasks moving towards what we call a "virtual new product development" stage that precedes the (physical world) new product development stage. Machines collect, store, and analyze data about hundreds or thousands of products and services created by different firms, and transfer this knowledge to the development of new products. Machines create a seamless way to communicate across silos to drastically reduce the costs and speed the creation of new products.

### 3.1.2 Less is more

Consumers are often overwhelmed and paralyzed by the myriad of choices. Streaming services, such as Netflix and Amazon Prime, provide (too) many choices, and consumers often complain they spend more time deciding what to watch than watching. Despite many efforts by machine learning to represent consumer preferences, marketers are still scrambling to find the right solutions and tradeoff between personalization and privacy. Indeed, the latest thinking is that consumers often learn their preferences as they search with new developments, including diversity, novelty, and serendipity being used in recommendation systems.

While pure machine learning predictions can improve recommendations, marketing science theory (the soul) is providing insight towards breakthroughs. Behavioral theories provide insights on how many choices customers should be presented with, how much advertising is too much, and in which sequence, should recommendations be shown.

These behavioral theories can be further and perhaps more deeply studied using online platforms that capture data throughout the consumer journey. For example, the theory of consumer learning can be tested for new insights during consumer search (Bronnenberg, Kim, Mela 2016). Behavioral insights can then be encoded toward algorithms that, for instance, make recommendations to help consumers learn their preferences as they search (Dzyabura and Hauser 2019). By leveraging the increasing availability of data and machine learning, a new generation of recommendation systems will help consumers. In addition, real-time adaptive algorithms based on multi-arm bandits, optimal experimentation, and causal inferences are helping marketers decide which choices to offer, when to offer them, and how to offer them. In doing so, marketers will be able to better target consumers leading to a better experience and facilitate a living style where "less is more."

### 3.1.3 Marketing as a cross-disciplinary field

Marketing as an isolated discipline may soon disappear. Historically, the marketing discipline created its own theories, algorithms, and methods. Perhaps, the most well-known are the choice model, the Bass model, and conjoint analysis (Guadagni and Little, 1983; Bass, 1969; Green and Srinivasan, 1978). While drawing from and affecting many other fields, marketing, for many years, has been a discipline on its own. Engineering science has been around since pioneering work by Frank Bass, Paul Green, and John Little; economic and econometric theory has provided structure for many problems; behavioral science has helped marketing to see the consumer and more than rational automation. But the growth of electronic media has made marketing central to the economy (Google, Facebook, and others are, in many instances, advertising platforms). It is common to read marketing papers that contain words such as machine learning, deep learning, predictions and machine-learning papers that contain words such as advertising, targeting, segmentation, and product development. Marketing scientists evolve and learn to embrace advances from any field and adapt them to solve important marketing problems. The machine is the new frontier.

Those in machine learning are developing new algorithms and new ways of thinking, but without a soul. Without the rich theoretical tradition of marketing science, this approach may miss key phenomena such as nuances of consumer search, behavioral decision, and choice.

## 3.2 The machine

### 3.2.1 Theory-driven applications

We have seen many successful applications of AI and ML to challenging games like Chess (Deep Blue by IBM), Shogi (Bonaza), and Go (AlphaGo by Google). These applications share two key features that form the foundation of their success. First, researchers started with a theoretical model of the structure of the game, which allowed them to break down the problem into a series of empirical tasks that then can be solved using advanced machine learning techniques. Second, the models were then trained on a wide range of data, which allowed the model to learn an optimal response function (or policy) for any board configuration. For example, in the case of AlphaGo, the research team trained the model extensively on both human and computer players. The algorithm learned to respond to and extremely broad range possible scenarios that it might encounter.

Many marketing and economics problems also have a game structure or optimization objective that can benefit from a solution concept that mirrors those of AlphaGo or Bonaza (Taddy, 2018). The two most important types of problems that marketers face are: (1) substantive questions, e.g., what is the ROI of a marketing intervention, and (2) prescriptive questions, e.g., how can a two-sided platform design a selling mechanism to maximize its revenues? An important issue in these types of questions is policy evaluation (for substantive questions) or counterfactual evaluation (for prescriptive questions), i.e., understanding how some outcomes of interest would evolve under a different treatment regime or a new data generating process than the one observed in the data.

To solve these types of problems, marketers use (1) a theoretical framework that can be used to develop optimal policies (e.g., game structure, optimization rule), and (2) sufficient randomization in the data-generating process.

With a theoretical framework, a dataset with sufficient randomization, and a robust model, marketing scientists can better evaluate the efficacy of counterfactual policies. Often, marketing scientists seek to break down the problem into a series of many small empirical tasks that then can be solved. An optimal way to do so, and that was not possible until just recently, is for marketers to break down the problem into a series of many small empirical tasks that then can be solved. Machine-learning methods can play an important role here because they have powerful non-parametric properties, and they are scalable to extremely large dimensions and datasets. This gives them high predictive accuracy and when combined with good data (i.e., with enough randomization), they are excellent tools for forming and evaluating counterfactual policies.

A small but growing stream of literature in marketing has adopted a combination of theory-driven frameworks and machine-learning methods to answer important substantive and prescriptive questions. For example, Rafieian and Yoganarasimhan (2018) adopt this approach to examine the incentives of mobile ad-networks to engage in micro-targeting. More recently, Rafieian (2019a) and Rafieian (2019b) study the general problem of adaptive ad-sequencing in non-strategic and strategic environments. Both these papers start with a theoretical framework that allows the researcher to break the problem into smaller empirical pieces, which are then solved using advanced machine learning and structural methods. Problems that we previously solved in small and static settings, can be now defined for large-scale datasets and dynamic settings. Problems like real-time optimal pricing at scale will be solved in milliseconds leading to applications such as digital shelf labels, or in-store coupons and promotions.

### 3.2.2 Data-driven causal discovery

Thanks to a growing body of machine learning tools, marketers could gain a 360 degree-perspective of consumers based on what they view (visual cues from the market environment), say (textual reviews and verbal feedback), and experience (interactions with products and services), and then respond at scale in a timely fashion. Undoubtedly, the machine is empowering us as human beings: eyesight sharpened, voice amplified, and speed boosted.

While breakthroughs were made available through machine learning and deep learning tools in feature extraction from unstructured data and predictive tasks, the causal nature of most marketing questions calls for creative use and modification of machine learning toolbox. Recognizing that the causal inference problem is, in its essence, a missing data problem, marketers can exploit and augment machine learning tools and adapt them to causal tasks.

We are seeing an emergence of models for data-driven causal discovery. Some early examples, not related to marketing, include automated discovery for genomic drivers of tumors and brain functional connectivity. Marketing science will soon see the growth and proliferation of machines learning algorithms for large-scale hypothesis generations and hypothesis testing. Hypothesis generation algorithms will be able to formulate hypothesis by analyzing large scale datasets, identify patterns, similarities, and differences among groups of customers, products, or firms. Then, hypothesis testing algorithms will analyze these hypotheses and proceed to search and discover potential candidates for experimentation, non-equivalent control groups, further modeling, instrumental variables, local shocks that create exogenous variation, and improved insight and theory.

### 3.2.3 Learning while doing

Ehrenberg (1994) describes the empirical-theoretical-empirical-theoretical (ETET) approach to management science. This approach consists of the repetition of four basic steps: 1) establish an empirically well-grounded theory, 2) test the theory, 3) deduce new conjectural theory from the test, and 4) test the new theory. Marketing papers already utilize machine learning to scale theory testing. But it can serve as navigator to discover new and interesting patterns that could lead to new and improved theories to be tested.

We are witnessing an increase in data availability that has no precedent (social network, user-generated content, games, etc.). This new data, combined with machine learning, allows us to test existing theories that were untestable just a few years ago, and to revise, reformulate, or generate completely new theories that can then be tested. This process is augmented by methods such as reinforcement learning algorithms that are models that continue to learn from their actions and the data. (But like everything in machine learning, reinforcement learning algorithms need soul—human intervention, reasoning, and insight.)

Consider search models. Formal models commonly assume either sequential or fixed sample size search (De los Santos et al., 2012). While both models lead to important insights, a rich behavioral literature suggests that consumers us a consider-then-choose process (e.g., Payne 1976) and that at least the consideration phase is heuristic (Payne, Bettman, and Johnson 1988). For example, basic consumer behaviors (e.g., consumer recall, or changes in consumer's information set) imply a more complex process. Machine learning (and Bayesian statistics) is providing the means to infer the actual consideration process

(Gilbride and Allenby 2004; Hauser, et al. 2010). Such insight will accelerate with advanced machine learning and new data availability. Marketers will be able to start with simple models and then learn from them more involved and systematic patterns that can then be re-applied to the model to make it more realistic and complex. These approaches will help marketers to better understand and model consumer behavior at an unprecedented scale.

### 3.2.4 Machine learning for social good

Many measures used by governments to make decisions (economic indices, consumer sentiment, etc.) are survey driven. Inflation, for example, is calculated by the Bureau of Labor Statistics by manually collecting 80,000 prices each month. U.S. unemployment is calculated monthly by using the Current Population Survey (CPS) of about 60,000 households. These methods are costly and are also difficult to be repeated at high frequency, and therefore they are often imprecise and revised over time. Indeed, new indices use contingent-value surveys to capture that portion of the GDP that is based on free goods such as search engines and social media (Brynjolfsson, Collis, and Eggers 2019).

Machine learning can augment or substitute for survey-based techniques and will do so soon. Machines can scrape at high frequency to collect publicly available information about consumers, firms, jobs, social media, etc., which can be used to generate indices in real-time. With careful development, these measures will be more precise and able to better predict the economic conditions of geographic areas at high granularity, from Zip Codes to cities to states and nations.

# 4 Conclusions

Everywhere we turn, we hear about machine learning (or a special case, deep learning). It is tempting to think that everything will be automated and that the singularity is upon us. But without a soul, without human insight, the capabilities of the machine will be limited. The true advances come from the soul and the machine. The soul is the consumer's complex and not always rational behavior (or at least our models do not yet capture it). The soul is the reaction of competitors in a complex market where not actions can be modeled. The soul is the marketing scientist who provides theory to direct the path of the development. And the soul is the marketing manager who must know when to trust the machine and when to trust instinct.

# References

Aksoy-Pierson M, Allon G, Federgruen A (2013) Price competition under mixed multinomial logit demand functions. *Management Science* 59 (8):1817-1835.

Bass, Frank M. 1969. A new product growth for model consumer durables. *Management science* 15(5) 215–227.

Bronnenberg, B. J., Kim, J. B., & Mela, C. F. (2016). Zooming in on choice: How do consumers search for cameras online?. *Marketing Science*, *35*(5), 693-712.

Brynjolfsson, Erik, Avinash Collis, and Felix Eggers (2019): Using Massive Online Choice Experiments to Measure Changes in Well-being,Proceedings of the National Academy of Sciences of the United States of America(PNAS),https://doi.org/10.1073/pnas.1815663116

De los Santos, Babur, Ali Horta¸csu, Matthijs R Wildenbeest. 2012. Testing models of consumer search using data on web browsing and purchasing behavior. *American Economic Review* 102(6) 2955–80.

Ehrenberg, Andrew SC. 1994. Theory or well-based results: which comes first? *Research traditions in marketing*. Springer, 79–131.

Gilbride, T., & Allenby, G. M. (2004, Summer). A choice model with conjunctive, disjunctive, and compensatory screening rules. Marketing Science, 23(3), 391–406.

Green, Paul E, Venkatachary Srinivasan. 1978. Conjoint analysis in consumer research: issues and outlook. *Journal of consumer research* 5(2) 103–123.

Guadagni, Peter M, John DC Little. 1983. A logit model of brand choice calibrated on scanner data. *Marketing science* 2(3) 203–238.

Hauser, J. R., Toubia, O., Evgeniou, T., Dzyabura, D., & Befurt, R. (2010, June). Cognitive simplicity and consideration sets. Journal of Marketing Research, 47, 485–496.

Rafieian, Omid. 2019a. Optimizing user engagement through adaptive ad sequencing. University of Washington, Working paper.

Rafieian, Omid. 2019b. Revenue-optimal dynamic auctions for adaptive ad sequencing. University of Washington, Working paper.

Rafieian, Omid, Hema Yoganarasimhan. 2018. Targeting and privacy in mobile advertising. *Available at SSRN 3163806* .

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search. Organizational Behavior and Human Performance, 16, 366–387.

Payne, J.W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 534–552.

Taddy, Matt. 2018. The technological elements of artificial intelligence. Tech. rep., National Bureau of Economic Research.

Timoshenko, Artem, John R Hauser. 2019. Identifying customer needs from user-generated content. *Marketing Science* 38(1) 1–20.