Effects of Algorithmic Decision-Making and Interpretability on Human Behavior: Experiments using Crowdsourcing

Avishek Anand[1], **Kilian Bizer**[2], **Alexander Erlei**[2], **Ujwal Gadiraju**[1],

Christian Heinze[3], **Lukas Meub**[2], **Wolfgang Nejdl**[1], **and Björn Steinrötter**[3]

[1] L3S Research Center, Leibniz Universität Hannover
[2] Chair of Economic Policy and SME Research, Georg-August-Universität Göttingen
[3] Institute of Legal Informatics, Leibniz Universität Hannover
[1]lastname@L3S.de
[2] lukas.meub@wiwi.uni-goettingen.de
[3] lastname@iri.uni-hannover.de

Abstract

Today algorithmic decision-making (ADM) is prevalent in several fields including medicine, the criminal justice system, financial markets etc. On the one hand, this is testament to the ever improving performance and capabilities of complex machine learning models. On the other hand, the increased complexity has resulted in a lack of transparency and interpretability which has led to critical decision-making models being deployed as functional black boxes. There is a general consensus that being able to explain the actions of such systems will help to address legal issues like transparency (ex ante) and compliance requirements (interim) as well as liability (ex post). Moreover it may build trust, expose biases and in turn lead to improved models. This has most recently led to research on extracting post-hoc explanations from black box classifiers and sequence generators in tasks like image captioning, text classification and machine translation.

However, there is no work yet that has investigated and revealed the impact of model explanations on the *nature of human decision-making*. We undertake a large scale study using crowd-sourcing as a means to measure how interpretability affects human-decision making using well understood principles of behavioral economics. To our knowledge this is the first of its kind of an inter-disciplinary study involving interpretability in ADM models.

Introduction

In the context of machine learning and more generally in algorithmic decision-making systems (ADMs) interpretability can be defined as "*the ability to explain or to present in understandable terms to a human*" (Doshi-Velez and Kim 2017). Inspite of the application of ADMs in a breadth of domains, for the most part, they are still used as black boxes which output a prediction, score or rankings without understanding partially or even completely how different features influence the model prediction. In such cases when an algorithm prioritizes information to predict, classify or rank, algorithmic transparency becomes an important feature to keep tabs on restricting discrimination and enhancing explainability-based trust in the system.

Why Interpretability?

Interpretability is often deemed critical to enable effective real-world deployment of intelligent systems, albeit highly context dependent (Weller 2017). For a researcher or developer, high interpretability is crucial to understand how their system/model is working, aiming to debug or improve it. For an end user, it provides a sense of what the system is doing and why, to enable prediction of what it might do in unforeseen circumstances and build trust in the technology. Additionally, adequate interpretability provides an expert (perhaps a regulator) the ability to audit a prediction or decision trail in detail and verify whether legal regulatory standards have been complied with. For example, explicit content for innocuous queries (for children) or to expose biases that may be hard to spot with quantitative measures.

Recent work has highlighted the opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks which avoid discrimination and enable explanation (Goodman and Flaxman 2016). Also, many regulatory policies now require or will require algorithmic transparency. Take for example the European Union's new *General Data Protection Regulation* (GDPR) which will take effect from 25 May 2018 onwards, that restricts automated individual decision-making which *significantly* affects users (Art. 22 GDPR). The law intends to create a right to explanation, whereby a user can ask for an explanation of an algorithmic decision that was made about them (Art. 12, 13(2) lit. f, 14(2) lit. g GDPR).

But how is human decision-making affected when ADMs are accompanied with explanations? How does it affect acceptability of ADMs ? Does it increase trust in the ADMs ? We intend to initiate large scale studies using crowdsourcing based on behavioral economics in order to understand how and if human decision-making is impacted when ADMs are accompanied with explanations.

Why Behavioral Economics?

Various external factors shape the design and effects of algorithmic decision-making systems and ultimately define the adequate implementation of interpretability measures. Besides being constrained by the institutional and regulatory framework, an optimal design further anticipates behavioral aspects of human-agent interaction (Mosier and Skitka 2018). We argue that only an interdisciplinary approach allows to analyze these factors comprehensively. Introducing behavioral economics offers such an integrative approach, that could substantially advance prevailing discussions in manifold dimensions. Over the last decades, behavioral economists have developed progressively detailed and sophisticated models of human behavior. This process has yielded a rich set of meticulous experimental methods and inherently diverse theoretical models (Kagel and Roth 2016; Camerer, Loewenstein, and Rabin 2011). While these models of human behavior need to account for the progress in artificial intelligence (Camerer 2017; Marwala and Hurwitz 2017), they enable a sound analysis of ADM systems increasingly penetrating into society. Specifically, we aim to examine how human behavior changes in human-agent environments and whether these changes have repercussions for economic outcomes. For instance, we are interested in total productive activity, the frequency of economically relevant interactions, cooperation and coordination activity or changes in overall as well as individual welfare. The use of pertinent economic models enables to generalize empirical findings and subsequently derive inferences about effects in our outcomes of interest. Consequently, certain ADM design and regulatory choices can be evaluated on relevant societal dimensions using straightforward counterfactuals (Kleinberg et al. 2017). Our approach therefore promises evidence that supports the design of economic policy measures with consequences for constructing machine-learning systems (Athey 2017; 2018).

To arrive at a suitable research design integrating behavioral economic science, our work in progress focuses on the effects of interpretability in human-agent interaction. For instance, explicitly quantifying the economic value of interpretability and identifying beneficiaries has implications for both the design of ADM systems and regulatory choices. We rely on ultimatum bargaining - a prominent working-horse in experimental economics - to derive novel insights with respect to the influence of ADM systems and interpretability on human behavior. Overall, we ask: *Does the introduction of ADM systems influence human decision-making in a straightforward bargaining context? How do ADM systems adapt to these presumably new behavioral patterns?* Beyond those rather general considerations, we specifically focus on interpretability to examine, e.g.: *Does increased interpretability influence established behavioral concepts such as acceptance, reciprocity or fairness concerns? Does it increase the quantity of economically relevant interactions and subsequently affect overall welfare?*

Interpretability of ML Models

Interpretability in Machine Learning has been studied for a long time in classical machine learning as a desirable property to have while chosing a certain model family under *interpretability by design* like decision trees, falling rule lists etc. However, the success of Neural networks (NN) and other expressive yet complex ML models have only intensified the discussion on post-hoc interpretability or interpreting already built models.

Consequently, interpretability of these complex models has been studied in various other domains to better understand decisions made by the network – image classification and captioning (Xu et al. 2015; Dabkowski and Gal 2017; Simonyan, Vedaldi, and Zisserman 2013), sequence to sequence modeling (Alvarez-Melis and Jaakkola 2017; Li et al. 2015), recommender systems (Chang, Harper, and Terveen 2016) etc. Interpretable models can be categorized into two broad classes: *model introspective* and *model agnostic*. Model introspection refers to interpretable models, such as decision trees, rules (Letham et al. 2015), additive models (Caruana et al. 2015) and attention-based networks (Xu et al. 2015). Instead of supporting models that are functionally black-boxes, such as an arbitrary neural network or random forests with thousands of trees, these approaches use models in which there is the possibility of meaningfully inspecting model components directly e.g. a path in a decision tree, a single rule, or the weight of a specific feature in a linear model.

Model agnostic approaches on the other hand extract post-hoc explanations by treating the original model as a black box either by learning from the output of the black box model, or perturbing the inputs, or both (Ribeiro, Singh, and Guestrin 2016; Koh and Liang 2017). Model agnostic interpretability is of two types: local and global. *Local interpretability* refers to the explanations used to describe a single decision of the model. There are also other notions of interpretability, and for a more comprehensive description of the approaches we point the readers to (Lipton 2016).

Interpretability and Human Decision-Making

Interpretability is no end in itself. The effects of interpretability remain ambiguous even if one learns about the effectiveness of interpretability measures as obtained by studies like (Garcia et al. 2009; Gacto, Alcala, and Herrera 2011). Rather, to resolve this ambiguity, one needs to ask in how far variation in interpretability transfers into variation in behavior.

For instance, additional explanations could foster a more trustful environment that motivates fruitful human-agent interactions. However, providing additional information might conversely result in an erosion of trust due to a more thorough scrutiny with respect to agent recommendations. Consider an agent supporting a physician (expert) in diagnosing a patient's (consumer) MRI scan. The physician might generally trust the agent based on positive experience and common knowledge about its superiority; thus reaching higher accuracy in his diagnosis. In contrast, learning about unfamiliar features used by the agent might cause distrust and has the physician stick to her own assessment. This hypothesis stems from evidence gathered by observing human interaction (Keller and Staelin 1987; Grimmelikhuijsen et al. 2013; Cramer et al. 2008; Ditto et al. 1998). Hence, increased interpretability might diminish the efficiency of such economically vital consumer-expert interactions.

The consideration above illustrates only one distinct case with inherent ambiguity regarding the effects of introducing increased interpretability. Besides trust, one might think of concepts established in behavioral economics like acceptance, accountability or social-preferences. Further, to obtain a more thorough understanding of increased interpretability, one needs to not only evaluate its effects on the end-user, but rather also consider regulators, developers or consumers. Such a comprehensive approach poses several challenges to the design of experiments and respective modeling of human behavior. Our work in progress relies on ultimatum bargaining to derive novel insight with respect to our considerations outlined above.

Crowdsourcing Methodology

Over the last decade, microtask crowdsourcing platforms such as Amazon's Mechanical Turk[1] and CrowdFlower[2] have been used to support or replicate findings from psychology and behavioral research, and also to run human-centered experiments on a large scale (Mason and Suri 2012; Crump, McDonnell, and Gureckis 2013; Chandler, Mueller, and Paolacci 2014; Gadiraju et al. 2017). Previous works have established that crowdsourcing platforms can be reliably leveraged to conduct large scale behavioral experiments that can be ecologically valid.

Ultimatum Bargaining Experiment

Ultimatum bargaining represents one of the most prominent games researched in experimental economics (Gueth, Schmittberger, and Schwarze 1982). Although it seems quite simple, understanding behavior in this framework remains complex even after decades of research (Gueth and Kocher 2014; van Damme et al. 2014). However, there is a rich literature allowing to integrate and evaluate the relevance of our findings. Literature on automated, though not artificial intelligent, agents from computer science and economics, makes the ultimatum game an optimal working horse to test our hypothesis. Our basic framework replicates the simplest design of the ultimatum game. A proposer X decides on the distribution of a pie with size p. X receives x and the responder Y receives y, where $x, y \geq 0$ and $x + y = p$. In a sequential process, the responder Y learns about the proposal (x, y) and either accepts $\delta(x, y) = 1$ or rejects $\delta(x, y) = 0$. Payoffs are given by $\delta(x, y)x$ and $\delta(x, y)y$, i.e. if the responder Y rejects both earn nothing.

A straightforward solution of the game - merely based on monetary outcomes - implies that responder Y should accept all positive offers, which gives $\delta(x, y) = 1$ for $y > 0$.[3] This is anticipated by the proposer X, which has him offer the minimal positive amount. In consequence, X receives almost the whole pie p and Y receives little more than nothing. However, actual behavior observed in prior experiments shows that the optimal offer by the proposer amounts to 40 to 50% of the pie. This might for example reflect fairness concerns or merely strategic thinking avoiding punishment by the responder who rejects offers perceived as unfair (Camerer 2003).

Experimental Setup

We will carry out a large scale ultimatum bargaining experiment by recruiting workers from a crowdsourcing platform. Workers will play the roles of proposers and responders under the following different between-subjects treatment conditions, to understand the effects of automated decision-making and interpretability on human behavior. We will follow guidelines from previous works to ensure reliable participation of crowd workers (Gadiraju et al. 2015).

I: *Human-Human Interactions.* This condition follows the simplest design of the ultimatum game as described earlier, consisting of a proposer and responder (roles that will be fulfilled by randomly paired workers recruited from the crowdsourcing platform). We will record the interactions between N unique (proposer, responder) pairs, i.e., the offers made by the proposer and whether they are accepted or rejected by the responder. Following this, the proposer and responder will independently complete certain personality related questionnaires.

II: *Human-Machine Interactions.* Using the N human-human interactions and features engineered from condition **I**, we will train a machine learning model that can classify whether a bid from a proposer is likely to be accepted. In this condition, proposers will be given the opportunity to use the machine learning model as an algorithmic decision-making system that can aid them in making a proposal. The proposers will be allowed to probe the ADM system with proposals and the system would report the likelihood of the proposal being accepted. The proposers will be allowed to probe the ADM system any number of times, but can only make a proposal to the responders once. The responders will be made aware of the fact that the proposer have a ADM system at their disposal to help them in making a proposal. Once again we will record the interactions between N unique and distinct (proposer, responder) pairs. These interactions between the proposers with the ADM system, as well as with the responders will provide us valuable insights on the effects of ADM on human behavior and how *trust* manifests and fluctuates via such interactions.

III: *Human-Machine Interactions with Proposers as Observers.* This condition is similar to **II**, except that the proposers will not be allowed to probe the ADM system but will only observe the proposals made by the system on her behalf. The responders will be conveyed that the offer being made is from an ADM acting on behalf of the proposer.

IV: *Human-Machine Interactions with Explanations.* This condition is virtually identical to **II**, except that proposers in this case will be aided with explanations alongside likelihood estimates to enhance interpretability when they probe the ADM system. Note that we consider model-introspective variants of interpretability where access to an already built model is provided. This will allow us to understand the role of interpretability in shaping human behavior while interacting with ADM systems.

V, VI, VII: *ADM Learned from Human-Machine Interactions.* To analyze the impact of the type of interactions that the ADM is learned from, we will train a similar machine learning model by using the interactions in condition **II**, that can aid a proposer in making an offer to the responder. This will allow us to investigate the impact of the type of interaction data (human-human versus human-machine) that the ADM is learned from, on the entailing observations of human behavior. Thus, the conditions **V**, **VI** and **VII** are repetitions of **II**, **III** and **IV** except for the interactions that the ADM is learned from.

[1] https://www.mturk.com/

[2] https://www.crowdflower.com

[3] While this represents the weakly dominant strategy for Y, all distributions (x, y) can be established as equilibrium outcomes. For multiple equilibria consider a certain threshold $\bar{y}$ for acceptance by the responder Y, such that $[(x, y), \delta(\tilde{x}, \tilde{y}) = 1]$ if $\tilde{y} \geq y$ and $\delta(\tilde{x}, \tilde{y}) = 0$ otherwise.

References

Alvarez-Melis, D., and Jaakkola, T. S. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *arXiv preprint arXiv:1707.01943*.

Athey, S. 2017. Beyond prediction: Using big data for policy problems. *Science* 355:483–485.

Athey, S. 2018. The impact of machine learning on economics. *Economics of Artificial Intelligence*.

Camerer, C.; Loewenstein, G.; and Rabin, M. 2011. *Advances in Behavioral Economics*. Princeton, NJ: Princeton University Press.

Camerer, C. 2003. *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Cambridge University Press.

Camerer, C. 2017. Artificial intelligence and behavioral economics. *Economics of Artificial Intelligence*.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. ACM.

Chandler, J.; Mueller, P.; and Paolacci, G. 2014. Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods* 46(1):112–130.

Chang, S.; Harper, F. M.; and Terveen, L. G. 2016. Crowd-based personalized natural language explanations for recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, 175–182. New York, NY, USA: ACM.

Cramer, H.; Evers, V.; Ramlal, S.; van Someren, M.; Rutledge, L.; Stash, N.; Aroyo, L.; and Wielinga, B. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18(455):456–496.

Crump, M. J.; McDonnell, J. V.; and Gureckis, T. M. 2013. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one* 8(3):e57410.

Dabkowski, P., and Gal, Y. 2017. Real time image saliency for black box classifiers. *arXiv preprint arXiv:1705.07857*.

Ditto, P.; Scepansky, J.; Munro, G.; Apanovitch, A. M.; and Lockhart, L. 1998. Motivated sensitivity to preference-inconsistent information. *Journal of Personality and Social Psychology* 75(1):53–69.

Doshi-Velez, F., and Kim, B. 2017. Towards a rigorous science of interpretable machine learning.

Gacto, M.; Alcala, R.; and Herrera, F. 2011. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Information Sciences* 181:43404360.

Gadiraju, U.; Kawase, R.; Dietze, S.; and Demartini, G. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1631–1640. ACM.

Gadiraju, U.; Möller, S.; Nöllenburg, M.; Saupe, D.; Egger-Lampl, S.; Archambault, D.; and Fisher, B. 2017. Crowdsourcing versus the laboratory: Towards human-centered experiments using the crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Springer. 6–26.

Garcia, S.; Fernandez, A.; Luengo, J.; and Herrera, F. 2009. A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing* 13.

Goodman, B., and Flaxman, S. 2016. European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*.

Grimmelikhuijsen, S.; Porumbescu, G.; Hong, B.; and Im, T. 2013. The effect of transparency on trust in government: A crossnational comparative experiment. *Public Administration Review* 73(4):575–586.

Gueth, W., and Kocher, M. 2014. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization* 108:396–409.

Gueth, W.; Schmittberger, R.; and Schwarze, B. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization* 3(4):367–388.

Kagel, J., and Roth, A. 2016. *The Handbook of Experimental Economics, Volume 2*. Princeton, NJ: Princeton University Press.

Keller, K., and Staelin, R. 1987. Effects of quality and quantity of information on decision effectiveness. *Journal of Consumer Research* 14:200–213.

Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2017. Human decisions and machine predictions. *The Quarterly Journal of Economics* 133(11):237293.

Koh, P. W., and Liang, P. 2017. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*.

Letham, B.; Rudin, C.; McCormick, T. H.; Madigan, D.; et al. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9(3):1350–1371.

Li, J.; Chen, X.; Hovy, E.; and Jurafsky, D. 2015. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*.

Lipton, Z. C. 2016. The mythos of model interpretability. *ICML Workshop on Human Interpretability of Machine Learning*.

Marwala, T., and Hurwitz, E. 2017. Artificial intelligence and economic theories. *arXiv:1703.0659*.

Mason, W., and Suri, S. 2012. Conducting behavioral research on amazons mechanical turk. *Behavior research methods* 44(1):1–23.

Mosier, K., and Skitka, L. 2018. Human decision makers and automated decision aids: Made for each other? In *Automation and Human Performance: Theory and Applications*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

van Damme, E.; Binmore, K.; Roth, A.; Samuelson, L.; Winter, E.; Bolton, G.; Ockenfels, A.; Dufwenberg, M.; Kirchsteiger, G.; Gneezy, U.; Kocher, M.; Sutter, M.; Sanfey, A.; Kliemt, H.; Selten, R.; Nagel, R.; and Azar, O. 2014. How werner gueth's ultimatum game shaped our understanding of social behavior. *Journal of Economic Behavior & Organization* 108:292–318.

Weller, A. 2017. Challenges for transparency. *arXiv preprint arXiv:1708.01870*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2048–2057.