Research Statement

Manuel Gomez Rodriguez

January 10, 2024

My main research interest lies in the development of **human-centric machine learning**, a rapidly growing field of research that seeks to maximize the societal benefits of machine learning systems and minimize their potential harms, risks and burdens.

During the last years, my main focus has been on developing human-centric machine learning models and algorithms for evaluating, supporting and enhancing decision making processes where algorithmic and human decisions feed and influence each other. To this end, my research has introduced conceptual innovations and technical breakthroughs along three different dimensions:

1. Probabilistic, causal, and game-theoretic models of decision making in specific contexts.

2. Algorithms that leverage these models to enhance decision making with provable guarantees.

3. Observational and interventional studies to evaluate such models and algorithms in practice.

Moreover, these conceptual innovations and technical breakthroughs have often uncovered previously unexplored connections between fields, as exemplified by several of my current achievements and my vision for the future.

Current Achievements

My research contributions to human-centric machine learning can be categorized in several distinctive research themes. Since each of these themes entails their own technical challenges and application domains, in what follows, I discuss each of them separately.

Human-AI Teaming. In recent years, there has been an increasing interest in developing machine learning systems especially designed to team up with human experts in a wide range of critical domains, from medicine and drug discovery to candidate screening and criminal justice. The main promise is that, by working together, human experts and machine learning systems will achieve a considerably better performance than each of them would achieve on their own. In the last years, together with students and postdocs in my group at MPI-SWS, we have pursued two lines of research to fulfill this promise on prediction tasks.

In a first line of research, we have pursued the idea of algorithmic triage. Under algorithmic triage, a machine learning model does not predict all instances but instead defers some of the instances to human experts. As a result, one does not only have to find a machine learning model but also a triage policy which determines who predicts each instance. Here, one of the main challenges is that, for each potential triage policy, there is an optimal machine learning model, however, the triage policy is also something one seeks to optimize. In a sequence of papers, we have developed some of the first algorithms with theoretical guarantees to learn under algorithmic triage in regression [1], classification [2, 3] and reinforcement learning [4] settings. In these pieces of work, we have conducted observational experiments showing that, by using algorithmic triage, human experts and machine learning models consistently achieve better average performance than each of them would achieve on their own.

In a second line of research, we have pursued the design of machine learning models that adaptively limit experts' level of agency [5, 6]. We have advocated for machine learning models which, rather than providing single predictions, provide sets of predictions, namely prediction sets, and forcefully ask human experts to predict label values from these prediction sets.[1] The key rationale here is that, by using the theory of conformal prediction [7, 8] to construct the above prediction sets, we can precisely trade-off the probability that the ground truth label is not in the prediction set, which determines how frequently the systems will mislead human experts, and the size of the prediction set, which determines the difficulty of the prediction task the experts need to solve using the system. Further, we have developed several bandit algorithms that leverage the nested structure of the prediction sets provided by any conformal predictor and a natural counterfactual monotonicity assumption on the experts' predictions to find the conformal predictor under which experts would benefit the most from using such system very efficiently. Finally, we have conducted a large-scale human subject study that suggest that machine learning models that adaptively limit experts' level of agency may offer greater performance than those allowing experts to always exercise their own agency.

Fairness in Machine Learning. As algorithmic decisions become more consequential to individuals and society, there have been increasing concerns about the potential for unfairness of algorithmic decision making. These concerns have been supported by a number of empirical studies, which have provided, *e.g.*, evidence of racial discrimination in criminal justice [9], health [10] or advertisement [11]. As a result, there has been a growing interest in the field of fairness in machine learning, which aims to develop machine learning models whose outcomes do not have a disproportionally large adverse impact on particular groups of people sharing certain sensitive traits such a race or sex.

Together with students, postdocs and collaborators at MPI-SWS, we have been at the forefront of the field since its early days. In a sequence of papers, we were among the first to tackle the design of (margin-based) classifiers with fairness guarantees [12–15]. This work received immediate international recognition by means of a best paper award honorable mention at the 26th International World Wide Web Conference and, since then, it has stimulated a very large amount of follow-up work (∼3,000 citations). In addition, Bilal Zafar—the PhD student who led these sequence of papers—won an Otto Hahn Medal from the Max Planck Society for outstanding scientific achievements in their doctoral thesis research.

More recently, together with students, postdocs and collaborators at MPI-SWS and MPI-IS, we have focused on settings in which there is a feedback loop between algorithmic and human decisions, which influence the data generation and collection process. In one piece of work [16], we have focused on learning accurate predictive models in scenarios where labels only exist conditional on certain decisions—if a loan is denied, there is not even an option for the individual to pay back the loan. We have shown that, in this selective labels setting, learning a predictor directly only from available labeled data is suboptimal in terms of both fairness and utility. To avoid this undesirable behavior, we have proposed to directly learn decision policies that maximize utility under fairness constraints

[1] There are many systems used everyday by experts that, under normal operation, limit experts' level of agency. For example, think of a pilot who is flying a plane. There are automated, adaptive systems that prevent the pilot from taking certain actions based on the monitoring of the environment.

and thereby take into account how decisions affect which data is observed in the future. In another piece of work [17], we have focused on designing ranking models that understand the long-term consequences of their proposed rankings and, more importantly, are able to avoid the undesirable ones. More specifically, we introduced a joint representation of rankings and user dynamics using Markov decision processes and, building upon this representation, introduced an efficient gradient-based algorithm to learn parameterized ranking models that trade off the immediate utility and the long-term welfare.

Finally, in a very recent piece of work, together with a student in my group at MPI-SWS [18], we have focused on selection processes in which an often intractable number of candidates is reduced to a shortlist of the most promising candidates using a screening classifier. We have shown that, even if a screening classifier is well-calibrated calibrated [19, 20], any threshold decision rule that uses such classifier may be biased against qualified candidates *within* demographic groups of interest. More specifically, it may shortlist one or more candidates from a group who are less likely to be qualified than one or more rejected candidates from the same group. As a consequence, it may perpetuate historical biases against minority groups by precluding the *best* candidates within these groups—the candidates who are more likely to be qualified—to be shortlisted. Then, we have introduced an efficient post-processing algorithm to minimally modify any given calibrated classifier so that it does not suffer from this type of within-group unfairness.

Strategic Machine Learning. As decision making is increasingly informed by data-driven predictive models, there is an increasing pressure on the decision makers to be transparent about the decision policies, the predictive models, and the features they use. However, individuals may be incentivized to use this knowledge to invest effort strategically in order to receive a beneficial decision. For example, in loan decisions, if a bank discloses that it will use credit card debt to decide whether it offers a loan to a customer, she may feel compelled to avoid credit card debt overall. Motivated by this observation, there has been a flurry of work in the emerging field of strategic machine learning in very recent years and, together with students, postdocs and collaborators at MPI-SWS, we have carried out some of the early work.

In a first piece of work [21], we were among the first to address the problem of finding decision policies that incentivize individuals to invest in forms of effort that increase the utility of the policy. This was in contrast with most, if not all, of the contemporary work, which had focused on the design of accurate machine learning models in the above mentioned strategic setting. One of our key ideas was the realization that individuals' investment of effort induces a change in their features and, under some technical assumptions, this change can be characterized analytically at a population level. Building upon this characterization, we have studied the hardness of the problem and identified a natural monotonicity assumption on the cost individuals pay to change features that allows for a highly effective polynomial time heuristic search algorithm to find optimal policies.

In a second piece of work [22], we have uncovered a previously unexplored connection between strategic machine learning and counterfactual explanations. In short, counterfactual explanations aim to help individuals subject to decisions informed by machine learning models understand what would have to change for these decisions to be beneficial ones. In this work, we have investigated how individuals may use the knowledge gained by counterfactual explanations to invest effort strategically and maximize their chances of receiving a beneficial decision. More specifically, we have developed several polynomial-time algorithms with approximation guarantees to find near-optimal policies and counterfactual explanations in such a strategic setting. Moreover, we have also shown that, by incorporating a matroid constraint into the problem formulation, one can increase the diversity of the near-optimal set of counterfactual explanations and incentivize individuals across the whole spectrum of the population to self improve.

Machine Learning for Counterfactual Reasoning. There is empirical evidence that humans improve their decision making skills by means of counterfactual reasoning—reasoning about what might have been had they made alternative decisions to those they actually took [23, 24]. However, in sequential decision making processes where multiple, dependent decisions are made sequentially over time, the number of alternatives a human may need to reason about can be intractably large, particularly if there is uncertainty on the dynamics of the environment. In the last years, together with students and postdocs in my group at MPI-SWS, we have initiated the development of machine learning models and algorithms to help humans conduct counterfactual reasoning in such settings.

In a first line of work, we have focused on Markov decision processes and developed algorithms to identify optimal sequences of alternative decisions that, in comparison with the factual decisions, would have counterfactually led to better outcomes. In a first piece of work [25], we have focused on discrete state spaces and developed an algorithm based on dynamic programming that is guaranteed to solve the problem in polynomial time. In a second piece of work [26], we have shown that, in continuous state spaces, we cannot generally expect to solve the problem in polynomial time. However, under a natural form of Lipschitz continuity of the environment's dynamics, we have developed a practical A^* algorithm that is guaranteed to return the optimal solution to the problem. In both pieces of work, using real data from the medical domain, we have shown that the alternative sequences found by these algorithms can provide valuable insights to enhance sequential decision making.

In a second line of work, we have focused on temporal point processes, a popular type of processes for modeling discrete event data in continuous time, and developed a first-of-a-kind temporal point process model that, in contrast with existing models, can be used to answer counterfactual questions about these type of processes [27]. For example, in epidemiology, assume that, during a pandemic, a government decides to implement business restrictions every time the weekly incidence—the (relative) number of new cases—is larger than certain threshold but unfortunately the incidence nevertheless spirals out of control. Our model could help the government understand retrospectively to what extent the incidence would have grown had a lower threshold been implemented.

Other achievements. Around the time I received tenure at MPI-SWS, the COVID-19 pandemic started and I quickly branched out from my current focus at that time and, together with students, postdocs and collaborators from multiple MPIs, EPFL and ETH, we worked on several projects related to COVID-19 at a furious pace. We have developed an epidemiological model based on temporal point processes that can be used to predict the spread of epidemics at an unprecedented spatiotemporal resolution [28], a privacy-preserving and inclusive system for epidemic risk assessment and notification [29], and a group testing method based on dynamic programming that is specifically designed to use the information provided by contact tracing [30].

Previous Achievements

During my PhD and my postdoctoral work, my research focused on understanding, predicting and controlling information diffusion over the Web and social media. This led to a series of papers [31–43], which significantly advanced the state of the art in the network inference and influence maximization problems. Two of these papers [32, 37] received immediate international recognition by means of a best paper award honorable mention at KDD, the flagship conference in data mining, and an outstanding best paper award at NeurIPS, the flagship conference in machine learning. Since then, these series of papers have stimulated a large amount of follow-up work (>3,000 citations) and have been the methodological basis for two journal papers on malaria in collaboration with epidemiologists at Imperial College and others [44, 45].

After I joined MPI-SWS as a tenure-track faculty, I realized that my doctoral and postdoctoral work on network inference and influence maximization leveraged particular instances of a more general and powerful type of random processes, marked temporal point processes, which could be potentially used to design a new generation of models and algorithms to predict and optimize the functioning of social, information and networked systems. In the years that followed, I leveraged this realization to lead the design of:

(i) probabilistic models based on marked temporal point process to predict information propagation [46–48], product competition [49], opinion dynamics [50], information reliability [51–53], knowledge content [54], and spatiotemporal processes [55, 56]. In all cases, by exploiting fine grained user data, the models provide more accurate predictions than the state of the art.

(ii) a series of efficient off-line and online algorithms with provable guarantees to steer information dissemination [57–63], detect and precent the spread of misinformation [64, 65], and design spaced repetition algorithms for efficient memorization [66–68]. These algorithms exploit an alternative representation of marked temporal point processes using SDEs with jumps and establish a previously unexplored connection between optimal control of SDEs with jumps and marked temporal point processes.

At the beginning of 2020, my work on marked temporal point processes made my case for tenure at MPI-SWS. However, around the same time, I started to shift focus spearheaded by an ERC Starting Grant on “Human-Centric Machine Learning”.

Vision for the Future

Had the physician initiated the antibiotic treatment for sepsis a day earlier, the patient would have recovered. Had they taken an earlier train to the airport, they would not have missed their flight. Had I clicked on the attachment of that strange email, my computer would have been hacked. These are examples of counterfactual reasoning, a type of reasoning that is epitomized by the phrase “what might have been,” which implicates a juxtaposition of an imagined versus factual reality [23, 69].

Counterfactual reasoning is tightly connected to the way we attribute causality and responsibility [70], and it has been shown to play a significant role in the ability that humans have to learn from limited past experience [24] and improve their decision making skills [71]. Is counterfactual reasoning a human capacity that machines cannot have? Surprisingly, recent advances at the interface of causality and machine learning have demonstrated that it is possible to build machines that perform and benefit from counterfactual reasoning, in a way similarly as humans do [72, 73]. However, these advances have predominantly comprised machine learning systems designed to operate autonomously, without human supervision. In the incoming years, my goal is to develop human-centric machine learning models and algorithms for automated decision support that are able to perform and benefit from counterfactual reasoning. In pursuit of its goal, together with students and collaborators, we will introduce several fundamental innovations and technical breakthroughs, which I discuss next.

Human-AI Complementarity. In a decision making process, whenever a machine learning model is used to predict the value of an outcome of interest, it has been widely agreed that the model should also provide a confidence score along with each prediction [19, 74]. Then, the decision maker is supposed to use the confidence score to calibrate how much to trust the prediction. However, multiple lines of empirical evidence have recently shown that decision makers have difficulties at developing a good sense on when to trust a prediction using confidence scores, based on the way they are computed today [75, 76]. As an immediate consequence, it is not yet clear how to ensure that, after receiving automated decision support from a machine learning model, the average quality of the decisions made by decision makers does not worsen. In my research, we will use counterfactual reasoning to first understand why and then design automated decision support systems that can only increase and never decrease the average quality of human decisions. Our starting point will be our recent work on human-aligned calibration [77], where we have shown that, on settings with binary decisions and outcomes, a (rational) decision maker can only make optimal decisions if the confidence scores satisfy a natural counterfactual monotonicity property.

Computational and Data Requirements. Using counterfactual reasoning, we will reduce the computational and data requirements of machine learning models underpinning automated decision support systems. To this end, we will view each machine learning model as a different (causal) intervention in the decision making process. Under this view, given decision making data gathered under one intervention, we will use the structural similarity and shared properties between this intervention and other alternative interventions to generate counterfactual decision making data. Finally, we will leverage this counterfactual decision making data to find the optimal machine learning model much more efficiently. Our starting point will be our on-going work on counterfactual prediction sets [6], where we have demonstrated that, for automated decision support systems based on conformal prediction, it is possible to achieve an exponential gain in terms of computational and data efficiency by leveraging counterfactual decision making data.

Algorithmic Harm. The concept of harm is a central tenet of ethical principles, codes and laws. For example, the bioethical principle "first, do no harm" [78], asserts that a doctor's moral responsibility to benefit patients is superseded by their responsibility not to harm them [79]. In this context, we do not find any reason to make an exception with automated decision support systems based on machine learning systems. Motivated by this observation, we will develop mechanisms to measure and control how much harm an automated decision support system based on a machine learning model may cause. Building upon the counterfactual comparative account of harm from the field of philosophy [80], we will say that an automated decision support system is harmful on a specific decision making instance if and only if the decision maker would have made a better decision on their own. Here, we will focus on automated decision support systems for multiclass classification tasks [5, 81] and matching tasks [82, 83].

To showcase and evaluate the above machine learning models and algorithms, we will go beyond the status quo, which relies only on observational experiments for evaluation, and conduct human subject studies with laypersons using Prolific, an online research platform that facilitates the recruitment and management of human participants. By doing so, we will demonstrate that counterfactual reasoning can bring practical benefits to machine learning for decision support.

References

[1] A. De, P. Koley, N. Ganguly, and M. Gomez-Rodriguez. Regression under human assistance. In *AAAI*, 2020.

[2] Abir De, Nastaran Okati, Ali Zarezade, and Manuel Gomez-Rodriguez. Classification under human assistance. In *AAAI*, 2021.

[3] Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable learning under triage. In *NeurIPS*, 2021.

[4] Vahid Balazadeh Meresht, Abir De, Adish Singla, and Manuel Gomez-Rodriguez. Learning to switch among agents in a team via 2-layer markov decision processes. *TMLR*, 2022.

[5] Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez-Rodriguez. Improving expert predictions with conformal prediction. In *ICML*, 2023.

[6] Eleni Straitouri and Manuel Gomez-Rodriguez. Designing decision support systems using counterfactual prediction sets. *arXiv preprint arXiv:2306.03928*, 2023.

[7] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[8] Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 2023.

[9] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 2017.

[10] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 2019.

[11] Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 2013.

[12] Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*, 2017.

[13] Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna Gummadi. Training fair classifiers. In *AISTATS*, 2017.

[14] B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. Gummadi. Fairness constraints: A flexible approach for fair classification. *JMLR*, 2019.

[15] Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference: Learning with cost-effective notions of fairness. In *the alternative sequence*, 2017.

[16] N. Kilbertus, M. Gomez-Rodriguez, B. Schölkopf, K. Muandet, and I. Valera. Fair decisions despite imperfect predictions. In *AISTATS*, 2020.

[17] B. Tabibian, V. Gómez, A. De, B. Schölkopf, and M. Gomez-Rodriguez. Consequential ranking algorithms and long-term welfare. In *UAI*, 2020.

[18] Nastaran Okati, Stratis Tsirtsis, and Manuel Gomez-Rodriguez. On the within-group fairness of screening classifiers. In *ICML*, 2023.

[19] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[20] Lequn Wang, Thorsten Joachims, and Manuel Gomez-Rodriguez. Improving screening processes via calibrated subset selection. In *ICML*, 2022.

[21] S. Tsirtsis, B. Tabibian, M. Khajehnejad, B. Schölkopf, A. Singla, and M. Gomez-Rodriguez. Optimal decision making under strategic behavior. *Management Science*, 2023.

[22] Stratis Tsirtsis and Manuel Gomez-Rodriguez. Decisions, counterfactual explanations and strategic behavior. In *NeurIPS*, 2020.

[23] Neal J Roese. Counterfactual thinking. *Psychological bulletin*, 1997.

[24] Neal J Roese and Kai Epstude. The functional theory of counterfactual thinking: New evidence, new challenges, new insights. In *Advances in experimental social psychology*. 2017.

[25] Stratis Tsirtsis, Abir De, and Manuel Gomez-Rodriguez. Counterfactual explanations in sequential decision making under uncertainty. In *NeurIPS*, 2021.

[26] Stratis Tsirtsis and Manuel Gomez-Rodriguez. Finding counterfactually optimal action sequences in continuous state spaces. In *NeurIPS*, 2023.

[27] Kimia Noorbakhsh and Manuel Gomez-Rodriguez. Counterfactual temporal point processes. In *NeurIPS*, 2022.

[28] Lars Lorch, Heiner Kremer, William Trouleau, Stratis Tsirtsis, Aron Szanto, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Quantifying the effects of contact tracing, testing, and containment measures in the presence of infection hotspots. *ACM TSAS*, 2022.

[29] Gilles Barthe, Roberta De Viti, Peter Druschel, Deepak Garg, Manuel Gomez-Rodriguez, Pierfrancesco Ingo, Heiner Kremer, Matthew Lentz, Lars Lorch, Aastha Mehta, et al. Listening to bluetooth beacons for epidemic risk mitigation. *Scientific Reports*, 2022.

[30] Stratis Tsirtsis, Abir De, Lars Lorch, and Manuel Gomez-Rodriguez. Pooled testing of traced contacts under superspreading dynamics. *PLOS Computational Biology*, 2022.

[31] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *ICML*, 2011.

[32] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring Networks of Diffusion and Influence. In *KDD*, 2010.

[33] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. *ACM TKDD*, 2012.

[34] M. Gomez-Rodriguez, J. Leskovec, and B. Schölkopf. Structure and Dynamics of Information Pathways in On-line Media. In *WSDM*, 2013.

[35] M. Gomez-Rodriguez and B. Schölkopf. Influence maximization in continuous time diffusion networks. In *ICML*, 2012.

[36] M. Gomez-Rodriguez and B. Schölkopf. Submodular inference of diffusion networks from multiple trees. In *ICML*, 2012.

[37] N. Du, L. Song, M. Gomez-Rodriguez, and H. Zha. Scalable influence estimation in continuous-time diffusion networks. In *NeurIPS*, 2013.

[38] M. Gomez-Rodriguez, J. Leskovec, and B. Schölkopf. Modeling information propagation with survival theory. In *ICML*, 2013.

[39] H. Daneshmand, M. Gomez-Rodriguez, L. Song, and B. Schölkopf. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In *ICML*, 2014.

[40] Manuel Gomez-Rodriguez, Le Song, Hadi Daneshmand, and B. Schoelkopf. Estimating diffusion networks: Recovery conditions, sample complexity & soft-thresholding algorithm. *JMLR*, 2016.

[41] Manuel Gomez-Rodriguez, Le Song, Nan Du, Hongyuang Zha, and B. Schoelkopf. Influence estimation and maximization in continuous-time diffusion networks. *ACM TIST*, 2016.

[42] Nan Du, Yingyu Liang, Maria-Florina Balcan, Manuel Gomez-Rodriguez, Hongyuan Zha, and Le Song. Scalable influence maximization for multiple products in continuous-time diffusion networks. *JMLR*, 2017.

[43] M. Backes, M. Gomez-Rodriguez, P. Manoharan, and B. Surma. Reconciling privacy and utility in continuous-time diffusion networks. In *CSF*, 2017.

[44] I. Routledge, S. Lai, K. Battle, A. Ghani, M. Gomez-Rodriguez, K. Gustafson, S. Mishra, J. Proctor, A. Tatem, Z. Li, et al. Tracking progress towards malaria elimination in china: estimates of reproduction numbers and their spatiotemporal variation. *Nature Communications*, 2019.

[45] Isobel Routledge, Shengjie Lai, Katherine E Battle, Azra C Ghani, Manuel Gomez-Rodriguez, Kyle B Gustafson, Swapnil Mishra, Juliette Unwin, Joshua L Proctor, Andrew J Tatem, et al. Tracking progress towards malaria elimination in china: Individual-level estimates of transmission and its spatiotemporal variation using a diffusion network approach. *PLoS Computational Biology*, 2020.

[46] M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha, and L. Song. Coevolve: A joint point process model for information diffusion and network co-evolution. In *NeurIPS*, 2015.

[47] M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha, and L. Song. Coevolve: A joint point process model for information diffusion and network co-evolution. *JMLR*, 2017.

[48] C. Mavroforakis, I. Valera, and M. Gomez-Rodriguez. Modeling the dynamics of online learning activity. In *WWW*, 2017.

[49] I. Valera and M. Gomez-Rodriguez. Modeling adoption and usage of competing products. In *ICDM*, 2015.

[50] A. De, I. Valera, N. Ganguly, S. Bhattacharya, and M. Gomez-Rodriguez. Learning and forecasting opinion dynamics in social networks. In *NeurIPS*, 2016.

[51] M. Farajtabar, M. Gomez-Rodriguez, N. Du, M. Zamani, H. Zha, and L. Song. Back to the past: Source identification in diffusion networks from partially observed cascades. In *AISTATS*, 2015.

[52] Behzad Tabibian, Isabel Valera, Mehrdad Farajtabar, Le Song, Bernhard Schoelkopf, and Manuel Gomez-Rodriguez. Distilling information reliability and source trustworthiness from digital traces. In *WWW*, 2017.

[53] U. Upadhyay, A. De, and M. Gomez-Rodriguez. On the complexity of opinions and online discussions. In *WSDM*, 2019.

[54] U. Upadhyay, I. Valera, and M. Gomez-Rodriguez. Uncovering the dynamics of crowdlearning and the value of knowledge. In *WSDM*, 2017.

[55] N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song. Recurrent Marked Temporal Point Process: Embedding Event History to Vector. In *KDD*, 2016.

[56] Martin Jankowiak and Manuel Gomez-Rodriguez. Uncovering the spatiotemporal patterns of collective social activity. In *ICDM*, 2017.

[57] M. Farajtabar, N. Du, M. Gomez-Rodriguez, I. Valera, H. Zha, and L. Song. Shaping social activity by incentivizing users. In *NeurIPS*, 2014.

[58] M. Karimi, E. Tavakoli, M. Farajtabar, L. Song, and M. Gomez-Rodriguez. Smart Broadcasting: Do you want to be seen? In *KDD*, 2016.

[59] U. Upahdyay, A. De, and M. Gomez-Rodriguez. Deep reinforcement learning of marked temporal point processes. In *NeurIPS*, 2018.

[60] Ali Zarezade, Abir De, Hamid Rabiee, and Manuel Gomez-Rodriguez. Cheshire: An online algorithm for activity maximization in social networks. In *Allerton*, 2017.

[61] A. Zarezade, A. De, U. Upadhyay, H. Rabiee, and M. Gomez-Rodriguez. Steering social activity: A stochastic optimal control point of view. *JMLR*, 2018.

[62] A. Zarezade, U. Upadhyay, H. Rabiee, and M. Gomez-Rodriguez. Redqueen: An online algorithm for smart broadcasting in social networks. In *WSDM*, 2017.

[63] Khashayar Gatmiry and Manuel Gomez-Rodriguez. The network visibility problem. *ACM TOIS*, 2021.

[64] J. Kim, B. Tabibian, A. Oh, B. Schoelkopf, and M. Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *WSDM*, 2018.

[65] S. Tschiatschek, A. Singla, M. Gomez-Rodriguez, A. Merchant, and A. Krause. Fake news detection in social networks via crowd signals. In *WWW*, 2018.

[66] B. Tabibian, U. Upadhyay, A. De, A. Zarezade, B. Schoelkopf, and M. Gomez-Rodriguez. Optimizing human learning via spaced repetition optimization. *PNAS*, 2018.

[67] A. Hunziker, Y. Chen, O. Mac Aodha, M. Gomez-Rodriguez, A. Krause, P. Perona, Y. Yue, and A. Singla. Teaching multiple concepts to forgetful learners. 2019.

[68] Utkarsh Upadhyay, Graham Lancashire, Christoph Moser, and Manuel Gomez-Rodriguez. Large-scale randomized experiments reveals that machine learning-based instruction helps people memorize more effectively. *npj Science of Learning*, 2021.

[69] Ruth MJ Byrne. Precis of the rational imagination: How people create alternatives to reality. *Behavioral and Brain Sciences*, 2007.

[70] David A Lagnado, Tobias Gerstenberg, and Ro'i Zultan. Causal responsibility and counterfactuals. *Cognitive science*, 2013.

[71] Keith D Markman, Matthew N McMullen, and Ronald A Elizaga. Counterfactual thinking, persistence, and performance: A test of the reflection and evaluation model. *Journal of Experimental Social Psychology*, 2008.

[72] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

[73] Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*. 2022.

[74] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, 2001.

[75] Gal Yona, Amir Feder, and Itay Laish. Useful confidence measures: Beyond the max score. *arXiv preprint arXiv:2210.14070*, 2022.

[76] Chirag Gupta and Aaditya Ramdas. Top-label calibration and multiclass-to-binary reductions. *arXiv preprint arXiv:2107.08353*, 2021.

[77] Nina L Corvelo Benz and Manuel Gomez-Rodriguez. Human-aligned calibration for ai-assisted decision making. In *NeurIPS*, 2023.

[78] Cedric M Smith. Origin and uses of primum non nocere—above all, do no harm! *The Journal of Clinical Pharmacology*, 2005.

[79] W David Ross. *Foundations of ethics*. Read Books Ltd, 2011.

[80] Joel Feinberg. Wrongful life and the counterfactual element in harming. *Social Philosophy and Policy*, 1986.

[81] Varun Babbar, Umang Bhatt, and Adrian Weller. On the utility of prediction sets in human-ai teams. In *IJCAI*, 2022.

[82] Kirk Bansak, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner, Duncan Lawrence, and Jeremy Weinstein. Improving refugee integration through data-driven algorithmic assignment. *Science*, 2018.

[83] Narges Ahani, Tommy Andersson, Alessandro Martinello, Alexander Teytelboym, and Andrew C Trapp. Placement optimization in refugee resettlement. *Operations Research*, 2021.