Building Provably Reliable, Safe, and Secure AI

When a hospital's machine-learning model recommends against a medical procedure, a bank's algorithm denies a loan, a large language model answers a legal question, we might ask: Would the outputs of these models change if the inputs were slightly different? Could a malicious actor manipulate the outcomes? If these models are trained on sensitive records and shared externally, could they leak private information? How can we be sure about the correctness of their outputs? As AI systems become critical infrastructure in high-stakes domains, practitioners need reliable answers to these questions to deploy them confidently and at scale. Yet, unlike other mainstream technologies, there are few aspects of AI systems' functionality we can be sure about, limiting their adoption. My research develops formal foundations to ensure AI systems are provably reliable, safe, and secure, enabling their responsible adoption in high-stakes settings.

To achieve this, I work at the intersection of theory and practice as an **applied theorist:** I leverage the formal toolboxes of **statistical theory, information theory, and optimization** to design methods with **practically relevant and useful guarantees** for (i) *evaluating* AI system functionality and (ii) *building* the systems with assurances from the ground up. I identify existing theoretical tools, develop, and adapt them to be applicable and relevant in practice. In the past, I have applied this approach to various aspects of AI functionality where principled evaluations and guarantees are crucial, such as **privacy, security, fairness, contestability, and recourse**. To address the challenges faced by practitioners in the real world, I joined Lausanne University Hospital for my postdoctoral research, where I continue applying my translational approach to ensure responsible development of **AI systems in healthcare.** Let me summarize my past work across three high-level areas:

1. **Foundations of practical privacy protection.** Machine-learning models memorize their training data, making it possible for malicious actors to extract sensitive information from models and their outputs. I investigated differential privacy—a standard method for provably protecting against data leakage through controlled noise addition—and found it has both beneficial and negative side effects: although it makes model performance more predictable overall by providing strong information-theoretic generalization guarantees [NeurIPS'22], it also causes arbitrary decisions at the individual level by increasing predictive multiplicity [FAccT'23]. I developed methods for making differential privacy more interpretable and useful by leveraging new numeric methods for analyzing its guarantees [NeurIPS'24]. Beyond differential privacy, I showed how information leakage can disproportionately affect different demographic groups and how to address this problem using principled statistical evaluations [PETS'22], and used information-theoretic analyses to establish the fundamental limits of obfuscation-based methods which scramble individual records [ICML'24].
2. **Adversarial robustness, contestability, and recourse.** When automated systems are involved in high-stakes decision-making processes, e.g., in approving loans, benefits, care, jobs, or granting parole, people should have legitimate ways to influence and contest unfair or wrong automated decisions. I systematized and formalized ways in which people attempt to obtain better outcomes from automated systems [FAccT'20], and developed methods to simulate these strategic behaviors using discrete optimization [NDSS'23]. I used mixed-integer linear programming to build tools for formally certifying when such interventions are possible [ICLR'24], finding that existing tools for algorithmic recourse—recommending actions that should enable people to rectify a negative outcome, e.g., get a loan once rejected—often fail. The methods of this line of work are general and are applicable to multiple problems: in security, to test models for robustness to manipulations; in explainability, to interpret model predictions using counterfactual "what if" explanations; and in fairness, to analyze whether predictions change when demographic attributes change.
3. **Real-world applications in healthcare AI.** As a research scientist at Lausanne University Hospital, I have been collaborating with medical informatics practitioners and clinicians to ensure reliability, safety, and security of AI in healthcare. I have been developing guidelines for trustworthy AI in medicine [JMIR'25], and evaluated realistic deployment scenarios for large language models in healthcare [NeurIPS WS'24] through a collaboration with 30 clinicians across 11 hospital departments. I surveyed and systematized methods for evaluating and ensuring privacy and utility of generative models for medical synthetic data generation [npj Digital Medicine'25], finding widespread issues such as the usage of conflicting metrics and privacy evaluations that have been shown to be ineffective. Recently, I have solved a long-standing open problem: ensuring that machine learning models trained on sensitive data provably satisfy interpretable privacy risk requirements using a novel analysis of differential privacy [NeurIPS'25].

There is still much work to be done. The provable privacy guarantees we can provide are still overly pessimistic for, e.g., the needs of medical practitioners, and thus still hurt utility more than necessary. We have barely scratched the surface of evaluating the robustness of models in high-stakes domains to manipulations or input fluctuations. I aim to continue working in these areas, and collaborate broadly with researchers and practitioners across disciplines to stay grounded to real-world challenges in high-stakes domains. Moreover, I aim to expand to cover other aspects of AI functionality which require assurances, such as uncertainty quantification and explainability. I detail more on past work and future plans next.

Impact. I presented my research at multiple invited seminars both in industry (e.g., Microsoft, Google) and academic institutions (e.g., MIT, University of Toronto, University College London, European Organization for Research and Treatment of Cancer). My work has received a spotlight distinction at the 2024 International Conference on Learning Representations (ICLR) and a best paper award at the 2024 NeurIPS Generative AI for Health Workshop. Building on my results on interpretable privacy-preserving machine learning [NeurIPS'24], I secured a USD 125,000 Swiss National Science Foundation Spark grant as PI to operationalize the findings, and substantially contributed to successful proposals for a USD 28,000,000 EU Horizon Europe project on synthetic data, and a USD 1,250,000 Swiss National Science Foundation project on language modeling in healthcare. My work on principled evaluation of a Twitter image processing system, based on my research on evaluating machine learning under manipulations, has won the first place at Twitter's 2021 algorithmic bias challenge, and has received media coverage (e.g., in The Guardian, Wired, and The Verge).

Foundations of practical privacy protection

Machine learning models can memorize and potentially leak information about their training data [Carlini et al., 2021]. Consequently, if these models are trained on sensitive data, privacy concerns emerge. This is particularly critical in high-stakes domains such as healthcare, finance, and criminal justice, where training data often contains personal information protected by laws and ethical guidelines. In this line of work, I study several aspects of privacy in machine learning from two related perspectives. First, understanding the intrinsic capabilities and limits of methods that aim to protect privacy. Second, assessing the interaction and trade-offs between privacy, utility, and fairness.

Unequal distribution of privacy vulnerabilities. To better understand who is most affected by privacy risks, I led a study investigating how vulnerable different groups are to attacks against privacy [PETS'22]. We found that some demographic groups are substantially more vulnerable to having their private information exposed than others. This work provided the first analysis of these disparities in privacy risks and developed methods for measuring them effectively. This observation has legal consequences: we show that determining whether models or their outputs constitute personal data under, e.g., GDPR, requires disaggregated analyses across demographic groups, in contrast to the established practice of evaluating the risk on average across the population.

A standard way to ensure that ML models cannot leak information about their sensitive training data is a formal framework called differential privacy [Dwork et al., 2006, 2014]. To achieve this privacy protection, one has to add a controlled level of random noise during model training [Chaudhuri et al., 2011; Abadi et al., 2016].

Differential privacy, reliability, and arbitrary predictions. I led a study which uncovered a side effect of this approach: The introduction of randomness leads to *arbitrary* model predictions [FAccT'23]. Concretely, some predictions end up depending entirely on the random noise rather than meaningful patterns in the data—technically, making the *Rashomon set*, the set of plausible models that the training algorithm could output, more noisy. Theoretically and empirically, we showed that stronger privacy protections lead to more arbitrary decisions, affecting different demographic groups unequally. This poses a fundamental concern for using such methods in critical applications in domains such as healthcare. In another study I have co-led, we empirically demonstrated that this noise also has a surprising benefit: ensuring that multiple aspects of model behavior remain consistent between training and deployment, i.e., strong generalization guarantees beyond just average training and test-time loss [NeurIPS'22]. We used this finding to build more reliable deep neural networks which provide a provable guarantee that the performance observed at train time is what one gets at test time, a property that in general does not hold for deep neural networks.

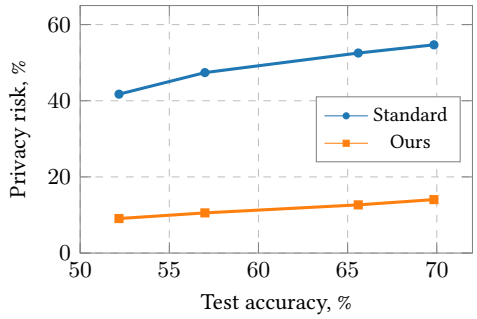

Figure 1: We perform privacy-preserving (differentially private) fine-tuning of a language model in a text classification task. Our method shows that the maximum privacy risk is substantially lower than what the standard method indicates. As a result, to calibrate the model to a given level of risk, we need to add substantially less noise, thus model's accuracy increases from 52% to 70% at even lower privacy risk [NeurIPS'24].

Achieving better utility with interpretable privacy guarantees. To address the challenges of noise addition, I co-led the developement of new "attack-aware" methods for calibrating the level of noise [NeurIPS'24]. In practice, we often want to evaluate the *operational privacy risks*: what is the maximum possible success rate of various attacks that aim to infer information about the training data. In contrast, the standard practice in differential privacy is to use the privacy parameter known as ε, which is not immediately interpretable in operational attack-aware terms. We showed that existing methods for mapping ε to operational risk are extremely ineffective. As a result of this inefficiency, if one needs to ensure that the models trained with differential privacy ensure a certain level of operational risk—as is often required in tightly regulated domains such as healthcare and finance—we need to add more noise than is necessary, often destroying the utility of models. To address this, we developed techniques to calibrate the noise in a way that directly provides guarantees on a level of interpretable operational privacy risk. This approach significantly reduces the amount of noise needed at the same level of risk, leading to better model performance without compromising on privacy, e.g., an improvement from 52% to 70% accuracy in a language modeling task (see Figure 1). I further led a study adapting this approach [NeurIPS'25] to capture the notions of risk that appear in data-protection guidelines—with more details next.

Future directions: *Bridging the theory and practice of data privacy protection.* Privacy of training data in machine learning is a systematic issue in high-stakes settings, which typically involve sensitive data. Although differential privacy is a mature theory in computer science, as I have learned at Lausanne University Hospital, it is rarely used in practice even when privacy guarantees are legally and ethically mandated. A critical barrier to adoption of differential privacy in high-stakes domains is that the added noise results in significant side effects such as degradation of model performance. In this direction, I will use my experience with differential privacy [NeurIPS'22; PETS'22; FAccT'23; NeurIPS'24; NeurIPS'25] to **(i)** introduce and study context-specific relaxations of differential privacy that provide meaningful formal privacy guarantees while carefully defining and measuring utility beyond simple accuracy metrics. A critical challenge is that standard differential privacy can have harmful effects on fairness—for instance, disproportionately degrading performance for minority groups or even effectively removing underrepresented populations from datasets. I will develop methods that take into account the joint privacy, utility,

and fairness considerations, ensuring that privacy protections do not come at the cost of exacerbating existing disparities. Additionally, **(ii)** I will develop new relaxed approaches to defining privacy that do not require randomness—instead relying on obfuscation of records—using our recent principled information-theoretic formulations [ICML'24]. These two directions would provide a sweet spot for improving adoption of privacy-preserving learning in practice, as they both would reduce the inherent performance degradation while explicitly addressing fairness and reliability concerns, yet still provide useful privacy guarantees.

Adversarial robustness, contestability, and recourse

In a series of works [NeurIPS WS'18a; FAccT'20], I showed how techniques traditionally viewed as "attacks" on ML systems—manipulating the inputs to a model in order to achieve a desired output—can actually serve either as tools for legitimate contestation of unfair or harmful decisions, or providing recourse to denied resources. If Alice has her loan application denied, what does she need to change about her application to get it approved? For complex ML-based decision-making systems, this is not trivial to answer, and requires algorithmic approaches. These works were among the first in an emerging line of research on contesting and affecting algorithmic systems externally, now known as algorithmic collective action.

Robustness of ML models to realistic input manipulations. Taking a step back to examine this problem from first principles, I identified that the core technical problem needed to understand whether decision-making systems can be legitimately affected is the same technical problem as evaluating robustness to adversarial manipulations [ICML WS'21]. Most tools available in the robustness literature, however, are not adapted to realistic constraints faced by people seeking recourse or contestation, and are thus not useful to evaluate whether model decisions can be affected in practice. Specifically, the standard approach is to start from a real input and find a perturbation which flips the model's decision at minimal, e.g., L_2 distance, to the initial input. Although this approach is mathematically convenient, it does not model realistic constraints well. For example, a realistic change to actionable features about a lending profile of a person to change the model's loan decision need not have a small L_2 distance to the initial input, and might instead be constrained by feasibility (e.g., one cannot decrease their age), effort (e.g., changing employment might be costly), or time (e.g., building credit history takes years). To address this issue, I co-led studies [NeurIPS WS'18b; NDSS'23] to design optimization methods which evaluate robustness of models to realistic, practical input manipulations. We do so by incorporating generic cost constraints as opposed to the standard geometric distances, and posing the problem as graph search.

Certifying algorithmic preclusion. Most recently, I discovered a fundamental issue: ML models deployed in high-stakes settings make predictions using features that people cannot ever meaningfully change, effectively "fixing" the negative outcomes [ICLR'24] *(spotlight distinction)*. Even though there exist certain methods for testing for responsiveness in such contexts [Ustun et al., 2019; Mothilal et al., 2020], we showed that existing methods systematically miss the cases in which outcomes cannot actually be changed or return actions that are actually infeasible (see Figure 2). For example, a loan-approval system might use factors that an applicant has no practical way to improve, permanently excluding them from access to credit once denied. To address this, we developed methods to formally verify whether people can actually achieve better outcomes through feasible actions (i.e., realistic changes they can make, as opposed to infeasible manipulations like decreasing their age), introducing the concept of "reachable sets"—the collection of all situations a person could reasonably achieve with realistic actions. We then used mixed-integer linear programming solvers to certify feasibility within those reachable sets in discrete domains. In work currently in submission, we have developed extensions of the method to support continuous and high-dimensional feature spaces, enabling certification for real-world models [preprint'25].

	Method	
Failure mode	DiCE	**Ours**
Loophole	34.4%	0.00%
Blindspot	21.0%	0.00%

Figure 2: Common methods (here we show DiCE [Mothilal et al., 2020]) to find *counterfactual explanations* or *recourse*—actions through which people can change the model's negative decision (loan denied) into a positive one (loan accepted)—often fail. They either find loopholes (infeasible actions, e.g., decrease age), or have blindspots (do not find an action when one exists) [ICLR'24].

The cited approaches [NDSS'23; ICLR'24] rely on different low-level algorithms, yet share the common principle of modeling realistic constraints faced by decision subjects. These algorithms are thus applicable beyond the problem of recourse. As case studies, we applied the methods to evaluate robustness to adversarial manipulations in content moderation, stability to natural measurement noise in kidney transplant risk models, counterfactual fairness—how much do predictions change under changes to sensitive attributes—and counterfactual explanations—explanations showing what needs to be changed about the input to yield a different model prediction—in recidivism prediction [preprint'25].

Future directions: *Certifying reliability and safety.* Regulations, e.g., those covering software as a medical device, require demonstrable evidence that high-risk AI systems perform reliably. For instance, a reliable medical model's prediction should not change erratically due to minor, irrelevant fluctuations in its input (e.g., sensor noise) or arbitrary choices during its training process (e.g., the random seed for weight initialization). Yet, standard performance metrics such as accuracy cannot, by definition, capture such notions of stability. This is a critical issue in ML deployments in high-stakes settings. Without guarantees on stability, model predictions are fragile and potentially untrustworthy. My research established two approaches to formally certify two distinct types of model stability. First, using techniques from [NDSS'23; ICLR'24; preprint'25], we can provide certifiable guarantees of a model's stability to input perturbations, proving that for a given range of changes to its input, its output will remain within a bounded region. Second, using another technique [FAccT'23], we can certify a model's stability to train-time interventions—arbitrariness—determining if a prediction is a robust outcome of the learning process

or an artifact of a specific random initialization. In the future research, I aim to apply these techniques to language models and predictive models for high-stakes tasks to **(i)** certify the input and training stability, and **(ii)** to produce counterfactual explanations and counterfactual fairness evaluations. The certificates of stability can provide reliable answers to questions like "the diagnosis would have been the same even if the patient's blood pressure was 5% higher" or "would this transplant risk model had a different prediction if the patient's age was different." This enables operational uncertainty quantification for abstaining when the model is uncertain (a prediction's instability indicates high uncertainty). Overall, these frameworks will enable practitioners to verify and attest to reliability before deployment, and support regulatory compliance.

Real-world applications in healthcare AI

In my current role at Lausanne University Hospital, I am translating the insights from prior work into healthcare practice. This work spans multiple collaborative projects, including solving open problems in privacy-preserving synthetic data generation.

AI in healthcare—understanding which use cases are appropriate and which are not. I am contributing to a large-scale participatory assessment of Large Language Model applications at Lausanne University Hospital [NeurIPS WS'24] *(best paper award at NeurIPS GenAI in Healthcare workshop)*, working with thirty stakeholders across 11 departments to identify potential use cases and assess their feasibility. We found that the applications that are most commonly studied in the academic literature, such as clinical decision-support systems, are the least likely to be deployed due to constraints with respect to infrastructure, data protection, and medical-device regulation. In contrast, applications focused on administrative tasks, such as documentation and workflow optimization, are more feasible for near-term deployment. This gap between research priorities and practical deployment possibilities is important because it highlights a fundamental misalignment: the field invests heavily in studying applications that hospitals cannot readily adopt, while more deployable use cases remain understudied.

Reliable and privacy-preserving medical generative modeling. Another major focus of mine has been evaluating medical generative models which generate *synthetic data* designed to mimic real patient data while preserving privacy. Through a scoping review [npj Digital Medicine'25], my collaborators and I identified significant gaps in how the biomedical field evaluates privacy and utility of synthetic data, demonstrating the lack of consensus and lack of reliable tools for evaluating both utility and privacy in synthetic data. We also conducted empirical studies that challenge the widely-held assumption that synthetic data can effectively reduce bias in real datasets. Our research on clinical prediction tasks [MIE'24a] demonstrated that augmenting real data with synthetic data to reduce bias rarely outperforms simple baseline approaches.

Moreover, our review found that privacy risks in synthetic data are often underestimated. In response, I led a study which solved a long-standing open problem: how to use differential privacy to create private-by-design generative models that are compatible with existing data-protection guidelines, e.g., from the International Standards Organisation (ISO), or European Medicines Agency's guidelines on the implementation of GDPR. For this, I adapted the tools from my Ph.D. research on attack-aware noise calibration [NeurIPS'24] to ensure that we can calibrate algorithms to interpretable notions of risk that often appear in the data-protection guidelines, such as re-identification, inference, and data reconstruction risks, in a way that allows for high utility [NeurIPS'25]. We showed that with prior methods for doing so, it is rarely possible to attain reasonable utility and reasonable level of privacy risk at the same time, something that our method solves (see Figure 1). Consequently, certain long-standing intuitions that were formed about differential privacy in the last decades of translation efforts—that it is not possible to have a low level of interpretable risk with high utility—turned out to be artifacts of inefficient ways to analyze differentially private algorithms.

Future work: *Interdisciplinary collaborations.* Responsible AI research cannot be done in isolation from the domains where AI is deployed. My approach, established through my work at Lausanne University Hospital, centers on participatory methods that involve domain experts, affected communities, and regulators. For example, my assessment of language models [NeurIPS WS'24] revealed that these systems are likely inappropriate due to regulatory, practical, and ethical constraints—insights only possible through collaboration with 30 stakeholders across 11 hospital departments. This methodology ensures my technical contributions address broader ethical issues. I aim to continue using such approaches in my future work.

References

Bogdan Kulynych, Yao-Yuan Yang, Yaodong Yu, Jarosław Błasiok, and Preetum Nakkiran. What You See is What You Get: Principled deep learning via distributional generalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 2022a.

Bogdan Kulynych, Hsiang Hsu, Carmela Troncoso, and Flavio P Calmon. Arbitrary decisions are a hidden cost of differentially-private training. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2023.

Bogdan Kulynych, Juan Felipe Gomez, Georgios Kaissis, Flavio du Pin Calmon, and Carmela Troncoso. Attack-aware noise calibration for differential privacy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. Disparate vulnerability to membership inference attacks. *Proceedings on Privacy Enhancing Technologies (PETS)*, 1, 2022b.

Theresa Stadler, **Bogdan Kulynych**, Michael Gastpar, Nicolas Papernot, and Carmela Troncoso. The fundamental limits of least-privilege learning. In *International Conference on Machine Learning (ICML)*. PMLR, 2024.

Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. POTs: Protective Optimization Technologies. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, pages 177–188, 2020.

Klim Kireev, **Bogdan Kulynych**, and Carmela Troncoso. Adversarial robustness for tabular data through cost and utility awareness. In *Proceedings of the Network and Distributed System Security (NDSS) Symposium*, 2023.

Avni Kothari, **Bogdan Kulynych**, Tsui-Wei Weng, and Berk Ustun. Prediction without preclusion: Recourse verification with reachable sets. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Georg Starke, Felix Gille, Alberto Termine, Yves Saint James Aquino, Ricardo Chavarriaga, Andrea Ferrario, Janna Hastings, Karin Jongsma, Philipp Kellmeyer, **Bogdan Kulynych**, Emily Postan, Elise Racine, Derya Sahin, Paulina Tomaszewska, Karina Vold, Jamie Webb, Alessandro Facchini, and Marcello Ienca. Finding consensus on trust in AI in healthcare: recommendations from a panel of international experts. *Journal of Medical Informatics (JMIR)*, 2024.

Giorgia Carra, **Bogdan Kulynych**, François Bastardot, Daniel E Kaufmann, Noémie Boillat-Blanco, and Jean Louis Raisaro. Participatory assessment of large language model applications in an academic medical center. In *NeurIPS GenAI for Health Workshop*, 2024.

Bayrem Kaabachi, Jérémie Despraz, Thierry Meurers, Karen Otte, Mehmed Halilovic, **Bogdan Kulynych**, Fabian Prasser, and Jean Louis Raisaro. A scoping review of privacy and utility metrics in medical synthetic data. *npj Digital Medicine*, 2025.

Bogdan Kulynych, Juan Felipe Gomez, Georgios Kaissis, Jamie Hayes, Borja Balle, Flavio du Pin Calmon, and Jean Louis Raisaro. Unifying re-identification, attribute inference, and data reconstruction risks in differential privacy. In *Advances in Neural Information Processing Systems (NeurIPS, to appear)*, 2025.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 2014.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC conference on computer and communications security*, 2016.

Rebekah Overdorf, **Bogdan Kulynych**, Ero Balsa, Carmela Troncoso, and Seda Gürses. Questioning the assumptions behind fairness solutions. *NeurIPS Critiquing and Correcting Trends in ML Workshop*, 2018.

Kendra Albert, Maggie Delano, **Bogdan Kulynych**, and Ram Shankar Siva Kumar. Adversarial for good? how the adversarial ml community's values impede socially beneficial uses of attacks. *ICML The Prospects and Perils of Adversrial Machine Learning Workshop*, 2021.

Bogdan Kulynych, Jamie Hayes, Nikita Samarin, and Carmela Troncoso. Evading classifiers in discrete domains with provable optimality guarantees. *NeurIPS Workshop on Security and Privacy in ML*, 2018.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

Seung Hyun Cheon, Meredith Stewart, **Bogdan Kulynych**, Tsui-Wei Weng, and Berk Ustun. Statistical inference for responsiveness verification. *arXiv preprint arXiv:2507.02169*, 2025.

Nina Wahler, Bayrem Kaabachi, **Bogdan Kulynych**, Jérémie Despraz, Christian Simon, and Jean Louis Raisaro. Evaluating synthetic data augmentation to correct for data imbalance in realistic clinical prediction settings. In *Digital Health and Informatics Innovations for Sustainable Health Care Systems (Medical Informatics Europe)*, pages 929–933. IOS Press, 2024.