Public Comment on the *Conception of the development of the AI field in Ukraine*

Bogdan Kulynych

May 2020

In this part of the comment, I address the parts of the *Conception* related to criminal justice.

Systems of "prevention of socially dangerous phenomena."

Modeling of future life outcomes (e.g., financial risks in lending to individuals, risks of recidivism, etc.) using methods of statistics and machine learning (the so-called "actuarial" approach to decision making) have a long history of criticism in academic and legal communities. First, such approaches are based on historical data. In social contexts, models based on historical data necessarily reproduce the patterns that arise due to social structure [1] (as also mentioned in "The problem of relevance, bias [...] in data," Chapter 7 of the *Conception.*) The use of such models in decision-making in justice for individuals or groups (e.g., prediction of recidivism, prediction of geographical outbreaks of crime), forms a feedback loop, which reproduces and entrenches the historical patterns found in the data, worsening social stratification [2–4].

Second, the most common methods of statistics and machine learning in decision-making are based on finding correlations between the historical results of classification and description of the classified person in data [5]. These models are based on observational data, and in general, without the use of special methods and without additional assumptions, cannot establish causal links between the description of a person and their classification [6]. Predicting the future with correlation-based models can be accurate in certain contexts, but is not reliable in general [7] [8, Sec. 5.9]. Robustness of the prediction is not the most important factor in other areas where AI methods are successful (e.g., targeted online advertising), because the cost of individual errors is small (e.g., advertising targeting system showed the user advertising that does not interest them). Robustness of the prediction is much more important when errors dramatically affect people's lives: in the case of systems for predicting recidivism, the error leads to unjustified inclusion or subsequent detention in the penitentiary system.

One of the most compelling arguments against the use of AI in decision-making is the following mass collaboration of 160 scientists under the guidance of Princeton University professor Matthew Salganik [9]. Having received high-quality training data on more than 4,242 families in the United States, with 12,942 descriptive characteristics collected over 15 years, more than a hundred researchers were tasked with predicting the lives of children in these families. The models were evaluated on part of the data to which the researchers did not have access during training, which ensured the impartiality of the evaluation. Participants used a variety of machine learning methods and statistics, but none of the models showed accurate results.

The COMPAS system provides models to predict recidivism for pre-trial proceedings and parole. This system is a stark example of both phenomena: predictions discriminate against vulnerable populations [10, 11], and do not have a significant advantage in accuracy over human-based estimates, even if people are not experts in the field [12]. In addition, the opacity of the system and its decisions hinders the analysis and understanding of these problems by decision-making bodies. The field of AI is developing rapidly in the world, and advances in image recognition and language understanding encourage applications in other areas. But the fact that Ukraine is behind other countries in the spread of AI applications has a positive side: we can learn from the mistakes of others.

Therefore, I propose to acknowledge the large amount of ethical, legal, social, and technological critique of prediction systems in justice and make the following changes to the *Conception*:

- I propose to completely remove the point in Section 8 (Justice): “Prevention of socially dangerous phenomena by analyzing existing data with the help of AI (e.g., COMPAS)”. The general wording of “prevention of socially dangerous phenomena” conceals algorithmic systems such as prediction of recidivism or geographical outbreaks of crime. As noted earlier, such systems do not comply with the principles of non-discrimination set out in one paragraph above in the *Conception* itself. I recommend to completely remove the item, because any data-driven decision-making system in the field of justice will have the effect of reinforcing historical patterns in the data.

- I propose to extend the paragraph on non-discrimination in the list of ethical principles of Section 8 (Justice) with the following clause: “non-discrimination, namely the prevention of any discrimination between individuals or groups of persons, **both in individual decisions as well as in long-term consequences of AI systems.**” In this proposal, I aim to clarify that in order to assess ”non-discrimination” it is necessary to take into account the consequences of the operation of decision-making systems not only at one point in time, but also in the long run.

- I propose to replace the point in Section 8 (Justice): “Judgment in cases of minor complexity (by mutual agreement of the parties) on the basis of analysis carried out through AI current legislation and case law” with “Development of transparent and interpretable instruments based on AI for assisting judges in making decisions in cases of minor complexity (by mutual agreement of the parties).” I propose this change, firstly, to avoid problems of system opacity, as in the case of the COMPAS system, and secondly, to avoid promoting the concept of developing fully automated systems that lead to a feedback loop problem.

References

[1] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[2] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

[3] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*, pages 62–76, 2018.

[4] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pages 160–171, 2018.

[5] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[6] Judea Pearl. Correlation and causation–the logic of co-habitation. *Written for the European Journal of Personality, Special Issue*, pages 1–4, 2012.

[7] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

[8] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice.* OTexts, 2018.

[9] Matthew J Salganik, Ian Lundberg, Alexander T Kindel, Caitlin E Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M Altschul, Jennie E Brand, Nicole Bohme Carnegie, Ryan James Compton, et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, 2020.

[10] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[11] Jeff Larson and Julia Angwin. Technical response to northpointe. `ttps://www.propublica.org/article/technical-response-to-northpointe`.

[12] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.