Title: Artificial Intelligence or the Ultimate Tool for Conservatism

Author: Maciej Marcinowski

Citation style: Marcinowski Maciej. (2022). Artificial Intelligence or the Ultimate Tool for Conservatism. "Danube" (Vol. 13, iss. 1 (2022)), doi 10.2478/danb-2022-0001

DANUBE • 13(1) • pp. 1–12
DOI: 10.2478/danb-2022-0001

ARTIFICIAL INTELLIGENCE OR THE ULTIMATE TOOL FOR CONSERVATISM

Maciej Marcinowski[1]

Abstract

Artificial intelligence (AI) is foremost viewed as a technologically revolutionary tool, however, the author discusses here whether it is in fact a tool for socio-economic and legal conservatism, because its training data is always embedded in the past. The aim of this paper is to explain, exemplify and predict – whether and how – AI could cause discrimination, stagnation and uniformization by conserving what is relayed even by the most representative data. Furthermore, the author aims to propose possible legal barriers to these phenomena. The presented hypotheses are based upon empirical research and socio-economic or legal mechanisms, aiming to predict possible results of AI applications under specific conditions. Results indicate that the inherent AI conservatism could indeed cause severe discrimination, stagnation and uniformization, especially if its applications were to remain unquestioned and unregulated. Hopefully, the proposed legal solutions could limit the scope and effectiveness of AI conservatism, encouraging AI-related solutions.

Keywords

Artificial Intelligence, Conservatism, Human Rights, Policing, Targeting

I. Introduction

The aim of this paper is to explain why and exemplify how conservatism is inherently embedded in the artificial intelligence (AI), then to predict its possible consequences and magnifying conditions. These intertwining consequences could be identified foremost as discrimination, but also stagnation and uniformization of the legal and socio-economic status quo. The topic at hand is often hidden under a veil of mathematical neutrality, while it needs to be understood in political terms, because its consequences are aligned with specific political aims, i.e. with the freezing of legal and socio-economic dynamics, even at the cost of preserving the existent discrimination. The supposed AI neutrality seems to be almost axiomatic and the topic of unfair AI tends to be reduced to data policies. Thus, it seems that no AI-related solutions will be proposed unless we argue, propose and adopt

[1] University of Silesia, Bankowa 11b, 40-007 Katowice, Poland. E-mail: ma.marcinowski@gmail.com.

legal solutions, that would enforce a development of proper AI solutions. Hence, the author considers possible legal frameworks, which could divert the negative consequences of AI conservatism and hopefully redirect the AI research towards finding technical solutions. The present approach to AI conservatism seems also novel in respect of subjectivity, i.e. it is devoid of ethical considerations (e.g. inequality) and focused on the possible and existent legal infringements (e.g. abuse of human rights) or practical issues (e.g. legislative stagnation). The ubiquity of AI, its immense power and uninterpretability, the negligible costs of data, the promises of personalized satisfaction and profit optimization, the belief in neutrality of technology, and the subsequent lack of regulations or common awareness, make it highly relevant to identify AI in political terms and discuss its consequences.

The so called artificial intelligence (AI) or rather machine learning (ML) models are predominantly mathematical functions, that can be specifically trained to perform automatic problem solving, because: 1) a number of their adaptive parameters could be automatically optimized through a process of calculations called gradient descent; 2) given that the data which represents the problem and its solutions could be numerically encoded. Therefore, ML models are in essence uninterpretable (i.e. black boxes), but automatically executable maps that lead us blindly from the problem data to the problem solutions (and the more complete such a mapping is, the more accurate such a machine is). We need to realize, however, that it is extremely hard to train any ML model, while ensuring it will find a general solution to the problem – rather than simply memorize solutions – based on the known variants of the problem, thus guaranteeing the model will solve also the unknown yet variants of the problem. And yet, even if such a generalization succeeds, it is achieved through the sheer scale of this approximation, that the machine conducted with respect to the known variants of the problem. In other words, the unknown variants are simply gaps determined and swept under the known variants of the problem. Hence, even the best approximation – to some optimal set of parameters – has nothing to do with reality, for it has all to do with the way our problem is known and presented, i.e. with the data, while the data is always past and historical by definition.

Let us observe, that whenever some machine makes a correct decision, to us it is a rational decision, but to the machine it is but a correctly calculated function. In other words, if we have successfully optimized our machine, then this function is a pattern of specific paths from the input to the output numbers, which is highly similar to some pattern of reality – in respect to some actions causing some reactions, that were encoded as these input and output numbers – hence this function may be used to automatically solve some pattern of reality. Or in yet simpler words, if we know some actions and encode these as a set of arbitrary input numbers, and if we know some reactions and encode these as a set of arbitrary output numbers, then to solve this causal problem is to find a function, such that completely maps one set to the other. However, we do not have a complete record about any causal relationship of actions and reactions, hence the best that we can achieve – through the automatic optimization of the function parameters – is some approximation roughly generalizing the record we have.

Since the dawn of science it is largely obvious that no future data exists. Hence, we are long used to this fact and its consequences. However, the ML models are not us and to them the consequences of relying on the past are different. That is to say, we can deduce or induce the future based on the past, finding new ideas and foreseeing new possibilities. The ML models can at best constantly approximate the future based on the past, therefore forecast only such events that already flow from the past. To aphorize this point, we could say that the human reasoning (and yet unobtained machine reasoning)[2] is less constrained by the past, than any – so to say – statistical reasoning. The above generalizations refer foremost to such ML models, that are specifically trained to make predictions. Hence, we should consider also the unpredictive ML models (tasked mainly with classification of data), and observe that the output of human mental endeavours is not as predetermined as in the case of these models, which have their outputs specified to exactly match our expectations. Therefore they cannot, but we can foresee things that never yet happened, because we can invent new laws, classes and systems, etc., then submit all of these to experiments and obtain new data, that we shall utilize to falsify or verify our hypothesis. While these ML models – that do not predict – were trained to approximate and repeat what is already known and considered correct (e.g. to classify the cloud images into some established categories of clouds, or to advertise a suburban housing area towards some automatically inferred categories of people who have historically preferred such property). The conservatism is understood here as any active or passive form of conservation of any status quo, i.e. of any current or recently past socio-economic reality. And conservation of any status quo is understood here to be as political and violent as any action contrary to any status quo[3]. Let us then observe the obvious, that all unbiased and representative data is past and representative of the status quo at the time of its collection. In other words, discrimination by the AI is quite often misinterpreted as a bias, while it is per se conservation of a discriminatory reality (which is most often correctly represented by the data). It is presumed correctly, that the ML models are objective and insensitive to human expectations – that is true because they are in essence mathematical and statistical or probabilistic models – but they are sensitive to the data collected by humans. The problem here considered, however, is not a biased data collection, but a perfectly unbiased data collection. As such data has to objectively represent the status quo of its collection, hence the hypothesis that ML is a tool for socio-economic conservatism, whenever ML models are applied to solve or automate solving of socially relevant problems. Of course, it may seem that social change is always inevitable, and it is unless the factors of social change are frozen – here, the ML is hypothesized as such a freezing factor (of mostly social and economic relationships).

The present hypothesis takes a problem of the ML based discrimination one step further than most research into the machine and data bias usually does. As it is important to realize, that whenever the status quo is already biased, then it will be at best prolonged and at worst reinforced by the ML applications. Because ML models are already applied in very sensitive areas of human life (e.g. they often constitute a core of the predictive policing

[2] Bottou (2014).

[3] Sallustius (2010).

models), they are often under heavy scrutiny as to their biases and data representativeness[4]. And yet it is almost not at all considered, that even an unbiased model could be a tool for conservation of discriminatory mechanisms and biases underlying the society. While other socially relevant applications of ML may not seem the most crucial perhaps – when compared to the predictive policing tools – they are often the ones most negatively affecting the real world in socially relevant ways. For example, the ML based targeting and micro-targeting of social media delivered advertisements – which are also very much profit optimised[5] – is most probably a very significant factor conserving the unjust status quo and racial discrimination on the housing markets in the United States of America[6].

If we further consider that data may be purposefully biased, then there exists a risk of ML models being applied with stricter political motivations, i.a. as tools for reactionary or even reformist and revolutionary purposes. We should observe here, that it will be much easier to gather data very far reaching into the past (reactionary), than to create some data describing our desired future that never yet existed (reformist or revolutionary). And that the former could be very simply covered up by being fully representative and unbiased, but just slightly too historical, thus remain long unquestioned. However, the foremost danger is the overall risk of human rights abuse based upon or reinforced through purposeful ML applications. Therefore, the general problem – of the socially relevant tools applied without any social control – and its solutions, should not be strictly politically motivated, as they could negatively affect anyone in any possible way.

If we were thus to ask the question – can we prevent the ML from being utilized as tools objectively detrimental to the society and economy – the answer is yes. We could prevent, if we were to enact the ML based predictions of human behaviour, human profiling and targeting as illegal overall. And this solution may be especially useful, because targeting, etc., causes the value of personal data, which leads to the most serious data abuse and protection problems.

However, because it may be overall too harsh and hard to generalize and regulate such ML applications that would cause the considered here effects, then the best method could be to create a governmental agency tasked with ad hoc control and supervision of at least the most sensitive areas where the ML models may be applied. Yet the foremost natural solution will be to regulate and allow the citizens, governmental agencies and the non-governmental organizations to take legal actions against any entities applying ML in such ways as to cause the considered here negative effects.

Although the problem of AI conservatism is – in its essence – often considered by other authors[7], it is possibly never named as such, thus it is misidentified, except for hints seldom present in popular literature[8] and journalistic discussions[9], that alas are not peer-reviewed. Some authors, however, have undertaken a similar task and presented thus the problem of

[4] Lum and Isaac (2016), Ensign et al. (2018), Bennett-Moses and Chan (2018) Richardson et al. (2020).

[5] M. Ali et al. (2019a, 2019b).

[6] US Department of Housing and Urban Development (2019).

[7] Hoffmann (2019), Whittaker et al. (2018).

[8] O'Neil (2016).

[9] Doctorow (2020).

AI conservatism in respect to ML driven perpetuation of societal biases[10]. These papers are highly complementary to the present one, however, the author aimed to focus on systemic impacts and future consequences of ML conservatism – within a legal framework – and policies necessary to remedy and govern them as objectively practical/legal, rather than subjectively ethical problems.

II. Discrimination

Let us first consider the problem of ML based human profiling, targeting and behaviour predictions, as exemplary means of social conservation that already negatively affects the human rights, foremost through reinforcement of discrimination. During the past decade, multiple research was conducted and published as to the risks of predictive policing, which is foremost based on the machine learning models. The basic idea of predictive policing is to forecast where and when crimes shall most probably occur – based on statistical data supplied to the ML model – allowing hence for better planning and deployment of police resources. Here, we can already foresee the main and yet superficial problem, i.e. the unintentionally biased and unrepresentative data (due to low standards of collection or low rates of crime reporting)[11], but also intentionally skewed data (the so called dirty data)[12], which leak then into the police statistics and self perpetuate, either through the subsequent updates of the predictive system or human interference. Then, these systems are media covered as a useful method for lowering the crime rates (while they could lower the rates without lowering the crime) and spread often without any proper evaluation[13]. To observe that the above problems are indeed superficial, let us assume that the historical data utilized for the initial model training was unbiased and representative of the reality, and that this social reality was exemplary of some systemic discrimination. If we deploy and guide our police forces based on the ML predictions, then we are conserving the existing status quo, if we furthermore update our ML training databases based on the resulting actions and reactions of our forces, we are directly reinforcing the existing systemic discrimination.

The present problem is foremost a legal one, that is how to regulate and react to the ML applications, also pre-emptively (e.g. by standardizing data collections), knowing that they affect human rights. For example, the allegations of i.a. racial discrimination stated by the Department of Housing and Urban Development (HUD) against Facebook[14], correctly suggest that the Facebook's ad targeting could be purposefully biased by advertisers. Let us observe, then, that even if these allegations were not true, and it could not have been purposefully biased, yet the charges of discrimination would have been correct nonetheless. Because Facebook has to effectively support the existing racial disparities – if it is algorithmically ad targeting – because the data it is using to perform the targeted housing advertisement was, in this case, collected in a country where housing reflects

[10] Zajko (2020).

[11] Bennett-Moses and Chan (2018).

[12] Richardson et al. (2020).

[13] Bennett-Moses and Chan (2018).

[14] US Department of Housing and Urban Development (2019).

systemic racial discrimination[15]. The aforestated difference is not just a matter of phrasing the problem, as it is a matter of finding the root of the problem, which are overall the individualised ads as a vector of discrimination. In other words, whether the Facebook is blind to the race or sex, gender, age, ethnicity and religion, etc. – this is a superficial blindness – as it is not at all blind to the individuals, who are represented as data by their online habits[16], etc., while their habits and personalities are caused foremost by their environmental factors, i.e. factors of the very environment that may be systemically discriminating them. The author has to note here, that it is not known what kind of an algorithm is applied by the Facebook to perform its targeted advertisement, but considering the complexity of the problem and the available macro-targeting features[17], it is most probably a state of the art machine learning model.

It was recently proved[18], that the Facebook algorithms for ad delivery lead to many different forms of discrimination on the employment and housing markets, foremost through profit optimized delivery (further discussed in the next section). And because Facebook allows the business entities to macro-target their ads in discriminatory ways – they can e.g. specify such targeting parameters as demographic characteristics and geographical locations of the ad receivers – which point was also challenged by the HUD. However, what is especially important, the aforementioned researchers reported also, that the results for targeted ads delivery were nonetheless significantly skewed along both the racial and gender lines, despite neutral targeting parameters and neutral content. These observations directly support the present hypothesis, because only the targeting algorithm could have skewed these delivery results without human interference (i.a. assuming some automated ads content classification), by delivering the ads foremost to such people who were statistically most willing to respond and most probable to be interested in these ads. Let us observe, that if there exists some characteristic of human behaviour in social media, such that is highly correlated with being e.g. a police officer, then foremost these people who share such a characteristic will be targeted by the algorithm delivering police employment offers. This will be either deepening the social stereotypes or straightforward discriminatory, not only when the data is biased or unrepresentative, but also when the data is exactly representative of some biased social reality or unjust status quo. It should be hinted here, that these targeting ML models could correlate some determined and semantically recognizable input and output data points (method allowing for direct human control over the input and indirect control over the output), or they could take some raw input data (e.g. pixel and text ad's content as in the Facebook's case[19]), and map these into some predetermined output categories or undetermined[20] lists of ad receivers, etc.

[15] Feagin (1999).

[16] M. Ali et al. (2019a), Bachrach et al. (2012).

[17] M. Ali et al. (2019a, 2019b), US Department of Housing and Urban Development (2019).

[18] M. Ali et al. (2019a).

[19] M. Ali et al. (2019a).

[20] If the output is predetermined, we say it is a supervised ML, otherwise it is unsupervised.

III. Stagnation

Let us secondly consider the problem of ML based human profiling and targeting as exemplary means of political and economic conservation, that already negatively affect our democratic institutions, foremost through political polarization and stagnation. It was recently proved[21] that the Facebook's algorithms are ad targeting based rather on the individual preferences of ad receivers, than based on advertisers' suggestions in form of macro-targeting. And since Facebook subsidizes such ads which are most relevant to their audience (where relevancy is a composite of action rates – like engagement or feedback), it supports political polarization (as a financially optimal solution) and creates informational filter bubbles – in other words – conserves the political status quo.

Let us observe then, that if the political brands pay less for their adverts streamed by the algorithm towards their supporters, than for the same adverts automatically targeted towards the supporters of their opponents, hence we can easily generalize, that it may be also true for brands overall. In other words, the ML based and profit optimised ads targeting could be effectively blockading the markets on levels that are already occupied by the main and well established brands, that is to say, targeting may be obstructing the fair competition for smaller and new business entities.

Let us proceed then with a following thought experiment, about the future constitutional institutions of some model democratic country, as to foresee its crises in respect of the possible ML applications in the socially relevant areas. Assuming that e.g. the blockchain technology could make the online electronic voting completely safe[22], we may easily imagine that the direct democracy shall be brought back and perhaps surpass its own previous forms. We should not assume, however, that such a direct democracy will make the representative institutions obsolete, as the modern problems are highly complex and such is the legislation, hence professional politicians or elective legislators are necessary. Thus it is more probable, that the future democracies will rather resemble the late Roman Republic with respect to its mechanism and balances of direct democracy. Hence, we should assume some democratically elected representative body tasked with preparing, initiating and enacting the legislation (e.g. the Roman Senate was such a democratic institution; however the system of senatorial elections favoured the aristocrats and oligarchs). Then we should assume some self-regulated and independent, but quorum and majority constrained online-assembly of citizens, who will directly and electronically vote legislation already passed by the representative body in favour or disfavour (the Plebeian Council was such an institution obligatorily consulted by the Senate, while its legal status was guaranteed by the elective Tribunes of the Plebs). The above constitutional system seems like the most natural and balanced form of direct democracy, and allows for direct analogies to the Roman constitutional law and its crises. For example, if it were a rule, that for anyone to pass any law it is obligatory to secure the legislative support of both bodies, then it will be quite heavily contested if anyone tried to pass legislation without such consent of both assemblies, yet it will not be nearly as strongly opposed for the online

[21] M. Ali et al. (2019b).

[22] Kshetri and Voas (2018).

public assembly to act out of order, i.e. to initiate and pass the law without petitioning it first to the representative body[23]. Furthermore, not only the minor or technical, but the overall standard legislation will be – so to say – routinely rubber stamped as a package by the online assembly, because of the large amounts of legislation to be obligatorily considered at every session, unless some opposition is gathered specifically to vote in disfavour.

Hence we can easily foresee the first crisis, stagnation, as the legislative agenda for the online-assemblies will be advertised foremost via the social media. Therefore, for political factions to secure quorums and pass any laws at all, these ads will be heavily targeted towards such people, who already very often and very positively engage in the online-assemblies (i.e. reliably rubber stamp every legislative package). Thus the price of adverts targeting these people will lower itself, hence create a self-perpetuating cycle of more constant support at lower prices, which shall envelop also its initiators. Obviously, the initial targeting will be representative of the given status quo, and then it shall narrow and reinforce itself as the above explained self-perpetuating cycle.

Another crisis may seem the reverse, yet is quite complementary and also stagnant. However, this crisis may be more intuitively foreseen because on the emotional level humans are far more likely to engage with high-arousal negative and positive political information (i.e. causing anger, anxiety and awe), rather than simply positive, neutral or sad content[24]. And, as more engagement means a lower price, and lower price means more range, then any meaningful or important legislative agenda could be endlessly paralysed by political factions. Because, the least expensive advert is such that is the most relevant to the receiver, i.e. an advert emotionally negative to him and produced by the faction he already supports (in this case an advert that informs and encourages him to vote against some legislative agenda). Here, the assumption is that any vote against something is far easier to load with high-arousal negative emotions – than a vote in favour of something with high-arousal positive emotions – hence it offers a better chance of supporters mobilisation. Thus, the tools of power struggle that the political factions shall reach to solve this dilemma, will be political polarization of the public, as they will have to narrate any important legislation they propose is written against someone or something, just to balance the negativity that could be far more easily gathered to stop rather than pass legislation. The best known example supporting this assumption is the famously failed Equal Rights Amendment to the constitution of the United States of America. We should observe here, that the polarization itself constitutes a form and tool of political stagnation, hence the above described mechanism shall be another self-perpetuating cycle.

The last crisis could be perhaps the worst, unless social media entities are under full scrutiny of all governments (not just of some one government where the entity providing a given platform is registered), and unless their data, algorithms and machine learning models are in full open access to the public. Because, the targeting – of ads informing that there is a vote about some matter – may be purposefully biased to serve or blockade any

[23] Plutarchus (2012).

[24] Marcus et al. (2000), Berger and Milkman (2012).

legislative agenda by these social media entities (and sometimes by third and alien parties, like the infamous Cambridge Analytica Ltd).

Except for this last crisis which is quite certain, the above considered crises deal solely with the problems of status quo conservation and information bubbles (that could be caused by machine modelled and profit optimised targeting) in respect of the most probable directly democratic institutions, because, in the author's opinion, these are the obvious consequences we can systemically and constitutionally hypothesise, while the rest belong to the future.

IV. Uniformization

One of the ML effects upon the socio-economic situation, stems from conservatism, but should be rather described as uniformization, such that occurs when some idea, product, brand, issue, etc., is meant to appeal to as many receivers as possible, hence it has to be irrelevant on all of the polarized planes of their interests (e.g. it has to avoid their politically motivated animosity at any cost)[25], as proved by many politically motivated boycotts of product brands. And since, so to say, there is not much room for style manoeuvring when the goals are utilitarian, then the usual outcome is uniformization (the statistically obvious example are these very common beverages, which are uniformly perfected not to have any specific flavour, hence never to cause any distaste). And good or ill, such uniformization may be foremost observed as the prevalent popular culture, which already is and shall be in the future far more strengthened (further uniformized) by applications of ML based targeting, profiling and behaviour predicting or even music composing, etc. Inevitably, because the point of popular culture products is exactly utilitarian, i.e. it is meant to appease as many people as possible, hence by its own definition[26], it has to either influence and reinforce similarities of their tastes, or at least never to cause any negative reactions, thus it cannot be radical or essential, while these goals may be easily reinforced by means of ML modelling, exactly as they already are reinforced by profits optimization.

One could argue also, that the ML models applied in law shall similarly lead to more uniform applications of the law in practice (i.e. uniform in accordance to the past practices), thus cause stagnation and conservation of the already existing legal mechanisms, practices, beliefs and theories. Hence, the probability of any new precedences and ideas could greatly diminish with further automation of the legal analyses, which shall inevitably (for technical reasons already explained) cast all the past judgements onto the new ones, while the society changes (therefore the society may also slow down its rate of change, or to the contrary, it could accelerate its pace violently). Let us consider here a simple empirical example of a rare, because open source and interpretable ML approach, called the Public Safety Assessment (PSA) model[27], that was independently evaluated by many USA jurisdictions, and is being utilized by the New Jersey Judiciary for risk assessments and release or detention decisions. There are two important factors to consider here: 1) the model was based upon accurate and objective statistical data representative of the

[25] Cossío-Silva et al. (2019).

[26] Horkheimer and Adorno (2002).

[27] American Civil Liberties Union of New Jersey et al. (2016).

New Jersey criminal justice system; 2) the New Jersey law states, that: “If the court enters an order that is contrary to a recommendation made in a risk assessment when determining a method of release or setting release conditions, the court shall provide an explanation in the document that authorizes the eligible defendant’s release”[28]. Hence, the authors of the New Jersey Pretrial Justice Manual objected, that: “Judges may be hesitant, therefore, to depart from Pretrial Services recommendations, either due to fear of being held responsible for a defendant’s misconduct if they release on lesser conditions than recommended, or desire to avoid additional paperwork”[29]. While we should raise a slightly different objection to this approach, that such a law will conserve and perpetuate the past and contemporary racial and socio-economic discrimination, and uniformly draw the whole New Jersey criminal justice system into a stagnant past, encompassing some decades of many dubious quality risk assessments and release or detention decisions. We should thus observe, that the aforediscussed problems of personal discrimination, systemic stagnation and legal uniformization are often closely interacting, despite causing a myriad of different harms, because having a common root in AI conservatism.

V. Conclusion

If we agree that the possibility of ML applications having some effects on society is true. And if we agree that such effects may be legally and politically meaningful. Then we should also agree that research into the statistical relevancy of such effects is needed. And if these effects were to be proven statistically relevant. Then we should agree that some solutions will have to be implemented on the international level as to ensure their full effectiveness.

And the simplest solution may be a general prohibition of human behaviour predictions, human profiling and human targeting. Such a solution could be significantly useful also because targeting causes the value of personal data, which leads to the most serious data abuse and protection problems. Such prohibition should not however interfere with these ML applications, where the ML is utilized to process some data on demand of its source, unless this ML application will be utilized neither in the private interest of the data source (e.g. of the customer), nor of the provider of this ML application (e.g. of some scientific institution), i.e. unless this data will be processed in the interest of some third party – unless the customer agrees (with the exception of possible consumer rights infringements and human rights abuse, especially that the later are inalienable).

Then, however, because it may be overall too harsh and hard to generalize and regulate such ML applications that would cause the considered here effects, the other simplest method is to create a governmental agency tasked with ad hoc control and supervision of at least the most sensitive areas where the ML models may be applied. Yet the foremost natural and organic solution will be to regulate the range of remedies that could be claimed by the citizens, governmental agencies and the non-governmental organizations – who should be hence allowed to take legal action – against any entities applying ML models

[28] New Jersey Legislature (2015).

[29] American Civil Liberties Union of New Jersey et al. (2016), p. 11.

in such ways that cause the considered here negative effects. In these cases, where the prediction of human behaviour, human profiling and targeting will be allowed, both of the last two solutions are necessary and complementary. First of all, there is already a need for planned and organized supervision of such ML databases and models – that are applied in the most socially sensitive areas (e.g. where they could affect human rights) – in respect of their purposeful biases and lack of representativeness (here, a specific and specialized governmental agency is necessary). Secondly, the above made point is also true for such areas of human life, that are exemplary of discrimination and other abuse of human rights (e.g. the policing in some countries), and could be conserved or even reinforced by the ML applications (e.g. if these countries apply predictive policing) – hence a specialized governmental agency armed in both active and proactive tools is necessary. Thirdly, some form of independent civil control is always necessary, in case the specialized agency is not properly authorized, funded or is politicized. Therefore the citizens, their non-governmental organizations and governmental agencies (but also the unspecified governmental agencies if the area of their authority is affected), should be allowed to undertake a specifically regulated legal action, since the problem is itself highly specific.

Finally, we may summarize that no ML model could ever be used to solve or cure a broken reality, for it shall always conserve or reinforce that reality based on the data, hence it should be used only there, where the data and reality are already sane. This is true for the problem of human rights abuse, foremost discrimination, while the problems of stagnation and uniformization could occur despite a sane reality, as a consequence of inevitable mechanisms underlying the society, that could be leveraged by ML applications.

References

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., Rieke, A. (2019a). *Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes*. arXiv:1904.02095v5 [cs.CY], 1–17.

Ali, M., Sapiezynski, Korolova, A., Mislove, A., Rieke, A. (2019b). *Ad Delivery Algorithms: The Hidden Arbiters of Political Messaging*. arXiv:1912.04255v3 [cs.CY], 1–16.

American Civil Liberties Union of New Jersey, National Association of Criminal Defense Lawyers and State of New Jersey Office of the Public Defender (2016). *The New Jersey Pretrial Justice Manual*. Retrieved April 14, 2021, from https://www.nacdl.org/getattachment/50e0c53b-6641-4a79-8b49-c733def39e37/the-new-jersey-pretrial-justice-manual.pdf.

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., Stillwell, D. (2012). Personality and patterns of Facebook usage. In Contractor, N., Uzzi, B., Macy, M., Nejdl, W. (ed.). *WebSci '12: Proceedings of the 4th Annual ACM Web Science Conference*. New York: Association for Computing Machinery.

Bennett-Moses, L., Chan, J. (2018). Algorithmic prediction in policing: assumptions, evaluation, and accountability. *Policing and Society*, 28, 806–822.

Berger, J., Milkman, K. L. (2012). What Makes online Content Viral. *Journal of Marketing Research*, 49(2), 192–205.

Bottou, L. (2014). From machine learning to machine reasoning. *Machine Learning*, 94, 133–134.

Cossío-Silva, F. J., Revilla-Camacho, M. A., Palacios-Florencio, B., Benítez, D. G. (2019). How to face a political boycott: the relevance of entrepreneurs' awareness. *International Entrepreneurship and Management Journal*, 15, 321–339.

Doctorow, C. (2020). *Our Neophobic, Conservative AI Overlords Want Everything to Stay the Same*. Retrieved April 14, 2021, from https://blog.lareviewofbooks.org/provocations/neophobic-conservative-ai-overlords-want-everything-stay/.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., Venkatasubramanian, S. (2018). Runaway Feedback Loops in Predictive Policing. *Proceedings of Machine Learning Research*, 81, 1–12.

Feagin, J. R. (1999). Excluding blacks and others from housing: The foundation of white racism. *A Journal of Policy Development and Research*, 4(3), 79–91.

Hoffmann, L. A., (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7), 900–915.

Horkheimer, M., Adorno, T. W., (2002). *Dialectic of Enlightenment*. Stanford: Stanford University Press, 94–136.

Kshetri, N., Voas, J. (2018). Blockchain-Enabled E-Voting. *IEEE Software*, 35 (4), 95–99.

Lum, K., Isaac, W. (2016). To predict and serve. *Significance*, 13(5), 14–19.

Marcus, G. E., MacKuen, M., Neuman, R. W. (2000). *Affective Intelligence and Political Judgment*. Chicago: University of Chicago Press, 126–129.

New Jersey Legislature. 2015. *New Jersey Revised Statutes § 2A:162-23*. Retrieved April 14, 2021, from https://law.justia.com/codes/new-jersey/2015/title-2a/section-2a-162-23/.

O'Neil, C. (2016). *Weapons of Math Destruction*. New York: Crown Books.

Plutarchus, L. M. (2012). *Parallel Lives*. New York: Start Publishing LLC, 621–626.

Richardson, R., Schultz, J. M., Crawford, K. (2020). Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *New York University Law Review*, 94, 15–55.

Sallustius, G. C. (2010). *Catiline's Conspiracy, The Jugurthine War, Histories*. Oxford: Oxford University Press, 134.

US Department of Housing and Urban Development. (2019). *HUD v. Facebook Inc., HUD ALJ No. FHEO No. 01-18-0323-8*. Retrieved April 14, 2021, from https://www.hud.gov/sites/dfiles/Main/documents/HUD_v_Facebook.pdf.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., Schwartz, O. (2018). *AI Now Report 2018*. New York: AI Now.

Zajko, M. (2020). *Conservative AI and social inequality: Conceptualizing alternatives to bias through social theory*. arXiv:2007.08666v1 [cs.CY], 1–21.