Inscrutability versus Privacy and Automation versus Labor in Human-Centered AI: Approaching Ethical Paradoxes and Directions for Research

Hagen Peukert
0000-0002-3228-316X
Universität Hamburg
Email: hagen.peukert@uni-hamburg.de

Abstract—From an analysis of ethical paradoxes based on the inscrutability feature of AI algorithms and resulting from recent advances in this field, this paper emphasizes the pressingness of dedicating research to the potential consequences on societal organization and interactions. With reference to Critical Theory that needs to be recombined with other socio-technical theories, new perspectives on future research is offered and discussed in light of privacy and labor market, their mutual influence as well as limitations.

I. INTRODUCTION

[6] OUTLINE possible challenges of managing human-centered Artificial Intelligence (AI). In the proposed model, the "frontiers of AI" (pp. 5, 14) are placed in a two dimensional space of scope and performance, which set the boundaries of three identified facets: autonomy, learning, and inscrutability. Moreover, the authors suggest further challenges likely to be relevant in the future; among them a reuniting frontier, i.e. ethical issues. In a wealth of literature, the later are referred to as the prevalent concern of the digital age and a major research gap that needs to be occupied on a much larger scale [8], [19], [14]. More specifically, violation of privacy, deep fakes, and accountability are burning issues evolving with AI and at the same time revealing different views on AI technology as well as conflicting at least partly with the prevalent value system established in the non-digital world.

In the digital world, the right for anonymity and the right of forgetting can no longer be guaranteed because algorithms are enabled to establish chains of correlations between pieces of information that can neither be foreseen nor fully understood by a human brain. Consequently, effective barriers to predict and protect these rights cannot be set up including the idea that AI is also employed to discover privacy violations itself. Similar to a trapdoor-one-way function in mathematics, the integrity of data can be manipulated by neural nets in such a way that it becomes impossible to clearly identify fake video clips and pictures from true material. Both, deep fakes and backtracking to a specific person, is realized but not prevented with AI. In other words, the issue at stake is the inscrutability condition as outlined in [6], which creates a paradox. On the one hand, humans have found a method that really helps them in getting work done; yet they are not capable of understanding how it really works. This inscrutability feeds back on the privacy of users. On the other hand, the lack of understanding AI methods also accounts for undermining fundamental rights mainly due to the inscrutability property, which is, by the same token, accountable for more job efficiency elsewhere. Thus, deep neural nets produce advantages and disadvantages that directly conflict each other. In analogy to nuclear fission, today it is still unclear if the advantages of AI outweigh the disadvantages. And this applies to the digital as well as to the analogous world.

The aforementioned paradox really consists of several ethical issues. Due to the inscrutability feature of AI systems, violated privacy has at least a direct and an indirect consequence. The direct consequence consists in massive collections on personality traits and unique identification vectors unknown to the users and future use. The indirect consequence is a substitution effect of human labor. AI algorithms collect data from users granting insights into processes and best practices on their work expertise while pretending help. The gained knowledge can then be used to oust human labor.

In this study, a critical perspective on these paradoxes will be provided. First the background of the theoretical framework, in which the identified AI paradoxes can be embedded, is given. Second, the nature of the paradoxes is analyzed and discussed. Last, the study closes with possible limitations and a short summary of the outcome.

II. THEORETICAL FRAMEWORK FOR ANALYSIS

The aim of the Frankfurt School was to scrutinize and challenge the existing power relations in questioning their underlying (often not explicitly known, but subconsciously assumed) preconditions, how these evolved over time and if these are still valid under the actual conditions that persist at present [24], [2], [17], [15]. Now, we find ourselves in a constantly readjusting societal value system. Socio-economic elites of a society have the power to influence belief and values in their favor and interests. Although elites cannot predict the outcome of their actions with certainty and must not consciously do so, the degree of impact is higher than for other societal groups. Put briefly, the area of influence is not absolute, yet it is higher, especially if considering the number of its members. These sociological findings form the theoretical backbone of what is referred to as Critical

Theory under the additional constraint that technology it- self establishes and strengthens power relations, i.e. it is tantamount to the means of production, capital goods, and financial assets in classical Marxian thinking. The two ethical paradoxes to be elaborated on in section III are model cases for this theoretical set up since technology is abused to (again) substitute labor. The substitution alone must not be negative. What is disadvantageous is how the socio-economic elite finds a way to keep the productivity gain from the substitution for themselves. From a sociological point of view, the substitution of white collar workers is especially demanding because it may reveal a difference to the first wave of machine automation, which stroke blue collar workers only.

At this point, it is necessary to extend Critical Theory by yet another explanatory variable, that is, proximity, i.e. the degree of similarity to the interests of the socio-economic elite. The hypothesis here would be: The closer a socio-economic group is to the socio-economic elite; the more concessions are made. Within this approach, one would have to observe that the substitution of white collar workers by AI technology comes along with political action that pays more tribute to the well- being of e.g. office workers compared to what the industrial workers experienced in the first wave of machine automation in the last decades of the last century. For the German case, one may speculate how the "Hartz IV" acts differ from the more recent legislation called "Bürgergeld", which may have construed towards possible future firings coming along with AI engagement in public administrations. At the European level, new AI legislation (AI act) is in preparation, which also includes concessions in this direction [30], [21]. This alone may not be evidence enough for including proximity in Critical Theory, but it gives the incentive to carry out research in this direction.

A different perspective on the challenges revealed by the proposed paradoxes is laid out in Socio-Technical Theo- ries [25], [31], [32], [33], [1]. Here the focus is on the organization and the question of how socio-technical change is analyzed towards the optimal way taking into account em- ployees and their values. There are some more theories tying into the ethics of AI and society, which should be considered and be integrated into future research in this direction. These include *Privacy Calculus Theory* [23], [7], [20] to validate the assumptions of the preset theoretical material. This would offer an IS perspective on privacy issues if extended by findings of *Technological Network Analysis*. In addition, *Information Asymmetry Theory* [3] and certainly also as a cascade of sev- eral very influential theories, the *Unified Theory of Acceptance and Use of Technology* [40], are needed to further add on explanatory potential. One of its inputs can be contrasted to the simpler *Technology Acceptance Model* [12], [13], that in a similar manner modifies its predecessor. By the same token these theories could be enriched by or contrasted to old classics such as *Rational Choice Theory* or *Prospect Theory*, but also by more recent findings prevalent in *Organizational Culture Theory*, *Psychological Ownership Theory*, and *Work Systems Theory*.

III. ANALYZING AI ETHICS PARADOXES

This study is supposed to reveal a contrastive view on two seemingly ethical paradoxes that are selected purposefully for their differing contents, assumptions, and consequences for the society, but at the same time make interconnectedness visible, that is, the inscrutability paradox provokes the "'AI versus labor'" dilemma. A stepwise comparison will also reveal the methodological set up of analyzing the two paradoxes. Roughly, the dynamic equilibrium model [37] serves as an orientation. Yet the differentiation of paradox and dilemma applies to the economic context that appreciates logical para- doxes, but not to ethics per se for that moral dilemmas and paradoxes are synonymously used. An ethical paradox is recognized in situations, in which important values or norms are violated no matter how an individual behaves [11].

While the questions on "AI versus labor" need a quantitative analysis of secondary sources, the "inscrutability paradox" is for the most part a qualitative investigation of network designs and algorithms as well as a collection of results in the literature. For the former, to check the assumptions implicitly made by the claim AI substituted labor, data on the branches of unemployment and correlations with branches of dismissal are available by the federal Statistical Office and the Census Bureau. Additionally, data of recruiting branches, expatriates and qualified immigration needs to be aggregated (cp e.g. [22]). From there, it becomes clear for which qualifi- cations organizations aim at. Last, programs of work creation schemes are considered. More particularly, the duration and number of programs that are aligned to AI technology can here be determined. And to evaluate the quality of these programs, they are compared to educational tracks typically present at university curricula.

The second method addressing the "inscrutability para- dox" is qualitative. As a starting point, a short collection of acknowledged results of big data analyses could be laid out (e.g. [26]). Some popular findings include results that happen to be technically correct, but are neither causal nor plausible, yet valid with respect to the algorithmic short cut through the data (e.g. of the form: people with green shirts, long hair, and ...have an 80 percent chance of getting a heart attack). If made public, some of these findings may feedback into future analysis and change the final result with unfair consequences for independent parties (e.g. sellers of green shirts). By use of these examples, an analysis of the algorithmic construct of a limited number of networks used for big data analysis will reveal that the inscrutability feature rather erodes privacy than enhancing it. Now, the same is done for net privacy issues (integrity, authentication, anonymity) and contrasted with e.g. profiling or fingerprinting. Last, privacy and inscrutability is brought together by showing what the application of the respective other algorithmic set up would mean for inscrutability and privacy.

The interrelatedness of privacy and the labor market is revealed when looking at the forecasted consequences on the white collar labor market. Now knowledge jobs, sales agents,

law consultants, and the like are affected on a large scale and much will depend on how fast respective AI technologies will be introduced in public and private organizations and if labor forces have sufficient time to shift to other areas of high-quality jobs. The average education time for these vocations are estimated to be five years. It is likely to depend on the age of the learner, too. This line of argumentation hints at the supposition of the paradox: the advantages of human-centered AI technologies in the service sector that are supposed to assist and add value to employed staff are traded off against the cost of their labor, that is, human workforce is potentially substituted by machine power whereas the profit of this substitution is roughly the cost of the employed. But different from the first wave of the machine revolution, which massively reached out to blue collar workers, the ethical paradox could be seen again in the inscrutability of deep learning. Without consciously knowing, service workers grant worldwide petabytes of data on processes, best practices, behaviors. Now, AI algorithms use this naturally grown knowledge, derive new patterns and routines from it. In fact, optimization functions produce even better results than their human counterpart would do. Put briefly, the human employee helps the machine to learn the specifics and secrets of its job. By doing so employees downsize and cut their own jobs while AI tools pretend to assist them, which they do, but in the background collect valuable data to get rid of the human workforce in the long run, which is prevented from transparency by the asynchronous inscrutability property. The knowledge gathered here is most of the time not even explicitly accessible to any human understanding (cf [35]).

This shows how human-centered AI technology in the digital sphere produces ethical paradoxes that further replicate serious consequences, even if unintentionally brought forward, for the interaction in the analogous world. First, it brings together the characteristics of inscrutability acknowledged as inherently preconditioned in all deep net's AI technology, on the one hand, and privacy issues, deep fakes, and accountability, on the other hand. Second, an ethical problem may also be seen in promoting the advantages of AI for the existing workforce while really exploiting their knowledge and preparing its substitution without making it transparent. This is not, as one may think at first glance, an inscrutability issue per se, but mediated by privacy violations since the inscrutability of the inner workings of a deep net are not what is hidden from the public, but the fact that more or less sensitive knowledge is gained from its aid. As a first working hypothesis it looks as though the first paradox seems to be inbuilt as the property of inscrutability that produces a dilemma no matter how it is framed. Inscrutability fosters anonymity, but conflicts with accountability and, at the same time, it discloses privacy (big data correlations), but facilitates disguise (deep fakes). The second paradox, however, is concerned with the social context, for which it is possible to make additional assumptions and find additional factors of influence. As such these assumptions and factors of influence should be modifiable towards the value system of our society. AI may disguise spying on labor skills

while pretending support, but it is still the decision of the members of the society to accept its consequences. In other words, if a machine substitutes human labor, we may decide that the substituted workforce continues to receive the full payment and may have more leisure time.

To validly evaluate possible consequences and developmental trends, it is to be further examined if additional assumptions could be made. They have to stand the test of plausibility and if possible have to be derived from the theoretical basis or should be included as a given fact. The latter is the case when the exact employment figures from Census of federal statistical offices, respective branches, and significant workforce shifts are taken into consideration. An example for a theoretical finding that could counterbalance an ethical evaluation is that in times of a high labor demand, the dismissed workforce is even better off if employed in newly created and emerging AI technology branches. This is the case in a Schumpeterean understanding. Yet, what could not be foreseen in the 1930ies, not even up until the turn of the century, was the velocity, in which these changes take place, and its unpredictable consequences for the societal set up (labor, wealth, values). As we have learned from the last burst of the Internet bubble in the early 2000th, interest rate reversals, or the financial crises with a 10-year recession of the American employment market later on that public institutions such as educational systems do not keep pace with the necessary requirements dictated by a digital world economy.

As a consequence, for an ethical evaluation, it is not enough to show one positive path out of dilemmas, but to take into account all (thinkable) possible scenarios and a plausible estimation of their occurrence probability. Indeed, a scenario is more likely if a theoretical claim or a claim derived from a theoretical basis has proven right by past events, which harbor similar assumptions. Paraphrased as a research question, one could ask what are the ethical paradoxes that follow from the assumptions given in the literature, i.e. inscrutability and privacy, task automation and labor market? What is the nature of these paradoxes? Under which costs and assumptions could they be resolved? The answer of an ethical question needs to be contextualized in a social context, in which a set of values prevail. Typically, these values neither are of equal importance ("speaking about the dead" vs. murder) nor do they stay constant over time (adultery, piousness). Some values may change rather drastically if power relations or other dynamics overcome a critical limit (euthanasia, right of succession, role of men and women), others are more rigid and seem to be static (theft, right of possession, piety).

Societies (as groups of people of different sizes [38]) happen to converge on values and its members show an intuitive understanding [27], [18], [16]. Values are key to social cohesion. It implies bottom-up learning processes that emerge over long time spans and they consolidate subconsciously in the collective memory of a community. Also, an intuitive and entrenched understanding of values is necessary to rank values [36], [5], [10], [4], [28], [29]. It is this ranking together with changes in the understanding of values that leads to

conflicts and ethical paradoxes whereas changes are often driven by technological innovations (such as AI) or major scientific findings (evolution, solar system, relativity).

Hence to understand ethical paradoxes, it is essential to guarantee social cohesion and thus this research directly offers its practical usefulness. Adjustments in the value system is an ongoing process in all societies and if made transparent, it becomes more robust. The humble scientific value comprises two aspects. First, it consists of the systematic proof of the underlying assumptions and balancing the ranking of values, i.e. how can AI as a new driver of technological innovation and its consequences be embedded into the existing value system. Second the correct derivation of action alternatives from the AI placement must be considered.

There are no clear cut solutions for ethical problems, otherwise they would not be any. Solutions are sketched as appraisals of all alternatives, a careful balancing of all known pros and cons while taking into account the value system prevalent at the time of evaluation. This modus operandi is common in Ethics in general and in Ethics in Information Science in particular. And it has been done for several Topics in AI research [39]. However, along the lines of inscrutability in the condensed form of the frontiers model [6], there is no systematic ethical investigation deriving privacy violations from the inherently given inscrutability feature and arguing for a causal relation to effects on the labor market.

IV. Limitations

When viewed through the lens of Critical Theory, the "Labor versus Automation" paradox could turn out to be not a real ethical paradox because there seem to be legislative solutions that do not contradict to ethical values. Yet the entire chain of arguments cannot be overlooked from a superficial assessment. So the legislature in favor of white collar workers could have negative effects for others, e.g. blue collar workers. The assessment could turn out to be different if considering Socio-technical theories that focus on organizations. Other theoretical groundwork may come forward with ideas not yet considered. All of them need to be accounted for the final evaluation.

For the case of the "inscrutability versus privacy" paradox, there exist at least no obvious solution from which a major drawback on the values could be denied. So no matter what is done, it will always be ethically questionable. Both, inscrutability and privacy, reinforces the weaknesses of the respective other. As a preliminary thought experiment: Increasing transparency, i.e. decreasing inscrutability, also raises the negative effects on privacy. Strengthening privacy (e.g. Thor browsing) enhances inscrutability even further. In addition, it would further lead to a bias towards users unfamiliar with the technology. As identified in the above argument, technology is seen on the same analytical level as financial assets or means of production. So it feeds back into supporting distorted power relations.

A possible limitation for the examination of labor and automation could be seen in the general economic situation on the labor market. Throughout all service branches, organizations report a high deficit on qualified employees. If this situation continues as some outlooks suggest, the labor-automation dilemma would lose its ethical grasp. In this case, the net effect of automation is very likely to be positive with respect to labor substitution. And it would only be relevant for Critical Theory for other aspects such as the traditional thinking in this field, that is, alienation of the human being from nature, but no longer the machine human substitution as the source of societal unfairness.

Concerning the limitations of inscrutability and privacy issues, it is not possible to take into consideration all prevalent architectures of neural nets for a technical analysis. Due to lack of evidence, it is still open if the inscrutability feature of neural nets could also be used to enhance privacy without paying off on accountability (e.g. by widely establishing cryptographic solutions [9]), so that inscrutability is not misused as a data collector, but as a data protector. However, there are no references in the relevant literature that really makes this theoretical possibility plausible.

V. Summary

The paper at hand set out to identify two relevant ethical paradoxes that come along with human-centered AI technology. It turns out that the inscrutability property of AI technology as produced by deep neural nets and by an increasing blurring of the ground truth input, invokes privacy violations. Privacy violations, on the other side, make it possible to collect huge amounts of data, which enable a machine view on services and processes that is not accessible to a human brain. These views are exploited as shortcuts bearing large efficiency gains for carrying out these services and thus making human work obsolete with seemingly dramatic consequences for the labor market. From this situation new paradoxes emerge from the very moment that humans allow machine exploitation without being able to grasp the full account of such decision. The consequences cannot be understood for the inscrutability argument; privacy stretches out over observations as to what is clicked, when and where. In fact, we see some kind of a cascade following from inscrutability over privacy to the engagement of labor. In addition to identifying the paradoxes, a specific theoretical context, in which these phenomena could best be studied is given and some obvious limitations to these approaches are also outlined.

References

[1] Adman, P., and Warren, L. 2000. "Participatory sociotechnical design of organizations and information systems - an adaptation of ethics methodology," Journal of Information Technology (15:1), pp. 39-51.

[2] Adorno, T. W. 1972. „Zur Logik der Sozialwissenschaften," in Gesammelte Schriften. Band 8: Soziologische Schriften I, Suhrkamp: Frankfurt am Main, pp. 547–565.

[3] Akerlof, G. A. 1970. "The market for 'lemons': Quality uncertainty and the market mechanism", The Quarterly Journal of Economics, pp. 488-500.

[4] Allen, V. L. and Levine, J: M. 1971. "Social support and conformity: The role of independent assessment of reality," Journal of Experimental Social Psychology (7), pp. 48-58.

[5] Asch, S. E. 1951. "Effects of group pressure upon the modification and distortion of judgments," in Groups, leadership and men, H. Guetzkow (ed.), Pittsburgh: Carnegie, pp.177-190.

[6] Berente, N., Gu, B., and Recker, J. 2021. "Managing Artificial Intelligence," MIS Quarterly (45:3), pp. 1-41.

[7] Blatchford, C. 2000. "Information security, business and the internet — Part 1," Network Security 2000(1), pp. 8–12.

[8] Bechmann, A. and Bowker, C. G. 2019. "Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social meda," Big Data & Society, pp. 1-11.

[9] Chaum, D. 1992. "Achieving Electronic Privacy, Scientific American," pp. 66-101, URL: https://chaum.com/wp-content/uploads/2021/12/ScientificAmerican-AEP.pdf.

[10] Cialdini, R. B. and Trost, M. R. 1998. "Social influence: Social norms, conformity, and compliance," in Handbook of social psychology D. T. Gilbert, S. T. Fiske and G. Lindzey (eds), Boston: McGraw-Hill, pp. 151-192.

[11] Cohen, B. 1967. "An Ethical Paradox," Mind (76:302), pp. 250–259.

[12] Davis, F. D. (1989). "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Quarterly 13(3), 319-339.

[13] Davis, F. D., Bagozzi, R. P. and Warshaw, P. R. (1989). "User acceptance of computer technology: A comparison of two theoretical models," Management Science (35:8), pp. 982-1003.

[14] DCC (Digital Curation Center). 2020. "The Role of Data in AI: Report for the Data Governance Working Group of the Global Partnership of AI," School of Informatics, University of Edinburgh.

[15] Habermas, J. 1981. Der philosophische Diskurs der Moderne. Zwölf Vorlesungen, Frankfurt am Main: Suhrkamp.

[16] Hogg, M. A., Sherman, D. K., Dierselhuis, J., Maitner, A. T. and Moffitt, G. 2007. "Uncertainty, entitativity, and group identification," Journal of Experimental Psychology (43), pp. 135-142.

[17] Horkheimer, M. 1988. „Traditionelle und kritische Theorie," in Gesammelte Schriften. Band 4: Schriften 1936–1941. Frankfurt am Main: Fischer.

[18] Janis, I. L. 1982. Groupthink, Boston: Hougthon Mifflin.

[19] Jaton, F. (2021): Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. Big Data & Society, pp. 1-15.

[20] Jozani, M., Ayaburi, E., Ko, M., and Choo, K. K. R. (2020). Privacy concerns and benefits of engagement with social media-enabled apps: A privacy calculus perspective. Computers in Human Behavior (107), 106260.

[21] Kerkmann, C. 2022. Künstliche Intelligenz: Wirtschaft warnt vor „massiven Einschränkungen" durch AI Act. URL am 23. Januar 2023: https://www.handelsblatt.com/technik/it-internet/eu-regulierung-kuenstliche-intelligenz-wirtschaft-warntvor-massiven-einschraenkungen-durch-ai-act/28850684.html.

[22] Kropp, P., Theuer, S. and Fritzsche, B. 2018. „Immer mehr Tätigkeiten werden durch Digitalisierung ersetzbar: Aktualisierte Substituierbarkeitspotenziale in Thüringen, IAB-Regional. IAB Sachsen-Anhalt-Thüringen," No. 02/2018, Institut für Arbeitsmarkt und Berufsforschung (IAB), Nürnberg.

[23] Laufer, R. S. and Wolfe, M. 1977. „Privacy as a Concept and a Social Issue: A Multidimensional Developmental Theory," Journal of Social Issues (33), pp. 22-42.

[24] Marcuse, H. 1965. „Philosophie und kritische Theorie," in Kultur und Gesellschaft I, Frankfurt am Main: Suhrkamp, pp. 102–127.

[25] Markus, L. 1983. "Power, politics, and MIS implementation," Communications of the ACM (26:6), pp. 430-444.

[26] McFarland, D. A., and McFarland, H. R. 2015. „Big Data and the danger of being precisely inaccurate,". Big Data & Society (2:2). https://doi.org/10.1177/2053951715602495 .

[27] Milgram, S. 1974. Obedience to authority: An experimental view, New York: Harper & Row.

[28] Moscovici, S. 1976. Social Influence and social change, London: Academic Press.

[29] Moscovici, S. 1980. "Toward a theory of conversion behavior," in Advances in experimental social psychology (13), L. Berkowitz (ed.), pp. 209-234.

[30] Müller, A. 2022. „Der Artificial Intelligence Act der EU: Ein risikobasierter Ansatz zur Regulierung von Künstlicher Intelligenz." Zeitschrift für Europarecht (1).

[31] Mumford, E. 1983. Designing human systems, the ETHICS approach. Manchester Business School, Manchester, U.K.

[32] Mumford, E. 2000, "Socio-technical design: An Unfulfilled Promise or a future Opportunity". in Organizational and Social Perspectives on Information Technology, Baskerville, R., Stage, J., and DeGross, J.I. (eds), Boston: Kluwer academic Publications, pp 33-46.

[33] Mumford, E., (2003), Redesigning Human Systems, Hershey: IRM Press.

[34] Mumford, E. and Weir, M., (1979). Computer Systems and work Design: The ETHICS Method, New York: Wiley & Sons.

[35] Paaß, G. and Hecker, D. 2020. Künstliche Intelligenz – Was steckt hinter der Technologie der Zukunft?, Springer Vieweg.

[36] Sherif, M. 1936. The psychology of social norms, New York: Harper & Row.

[37] Smith, W. and Lewis, M. (2011). "Toward a theory of paradox: A dynamic equilibrium model of organizing". Academy of Management Review, (36:2), pp. 381-403.

[38] Turner, J. C. 1991. Social Influence, Buckingham: Open University Press.

[39] Vallor, S. 2016. Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford: University Press.

[40] Venkatesh,V., Morris, M. G.;Davis, G. B.;Davis, F. D. 2003. "User acceptance of information technology: Toward a unified view", MIS Quarterly (27:3), 425-478.