Defective truth. AI or HI ideological imprints and political biases?

Adrian LESENCIUC

"Henri Coandă" Air Force Academy, Brasov, Romania

adrian.lesenciuc@afahc.ro, a.lesenciuc@yahoo.fr

Abstract: Artificial Intelligence, the AI 2.0 version (Pan, 2016), implies continuous adaptation to the information environment. The development of AI is generated by research and development requirements and by the need for an optimal response to the changing information environment. The change in the information environment entails the development of AI and, consequently, the development of information networks understood as *human-machine hybrid-augmented intelligence*. This dynamic is not reduced to the information or physical dimension, but to the cognitive one (JCOIE, 2018), which can be affected by the information flows necessary for the decision-making process. Overall, these are a few sources of AI-generated corruption of truth. The first of them is related to the generalization process through which statistical algorithms create instructions to be able to build the artificial neural network. The second concerns the human selection of samples on which statistical algorithms are applied to produce learning and the selection of principles on which information filtering occurs. Both produce trust-twisting errors, similar to those that operate in prejudice, stereotyping, and discrimination and leave ideological imprints on how AI operates. This article aims to analyse from the perspective of AI ethics, the forms of truth falsification through the process of machine learning specific to AI. In this respect, an interpretive/ qualitative meta-analysis of primary studies regarding the political biases of AI is proposed.

Keywords: AI Ethics, Machine Learning, Training Examples, Social Representations, Political Biases, Meta-analysis.

Adevăruri viciate. Amprente ideologice și biasuri politice ale inteligenței artificiale sau ale inteligenței umane?

Rezumat: Inteligența artificială în versiunea IA 2.0 (Pan, 2016) presupune adaptarea continuă la mediul informațional. Dezvoltarea IA este generată de cerințele de cercetare și dezvoltare și de nevoia unui răspuns optim la mediul informațional în schimbare. Schimbarea mediului informațional presupune dezvoltarea IA și, în consecință, dezvoltarea rețelelor informaționale înțelese ca inteligență hibridă *om-mașină* (*human-machine hybrid-augmented intelligence).* Această dinamică nu se reduce la dimensiunea informațională sau fizică, ci la cea cognitivă (JCOIE, 2018), care poate fi afectată de fluxurile informaționale necesare procesului decizional. În general, există câteva surse de viciere a adevărului generate de IA. Prima dintre ele este legată de procesul de generalizare prin care algoritmii statistici creează instrucțiuni pentru a putea construi rețeaua neuronală artificială. A doua se referă la selecția umană a eșantioanelor pe care sunt aplicați algoritmi statistici pentru a produce învățare și selecția principiilor pe baza cărora are loc filtrarea informațiilor. Ambele produc erori care distorsionează încrederea, similare cu cele care operează cu prejudecăți, stereotipuri și discriminare și lasă amprente ideologice asupra modului în care funcționează IA. Acest articol își propune să analizeze, din perspectiva eticii IA, formele de falsificare a adevărului prin procesul de *învățare automată (machine learning)*, specifice inteligenței artificiale. În acest sens, se propune o meta-analiză interpretativă/ calitativă, a studiilor primare privind biasurile politice ale IA.

Cuvinte cheie: etica IA, învățare automată (machine learning), exemple de antrenament (training examples), reprezentări sociale, biasuri politice, meta-analiză.

1. Introduction to AI 2.0

To identify the sources of political biases in the most well-known form of artificial intelligence, ChatGPT, an effort was made to look for sources of truth falsification by artificial intelligence based on models of human intelligence. In this respect, a short history of AI versions for reasons of reduction to the essence of the *machine learning* process was presented, which depends on external sources, i.e. human intelligence that selects *training examples* and that allows *ground-truth* calibration. The two sources depend on the information environment as a whole,

https://doi.org/10.33436/v34i3y202401

which include two types of definitions: from the field of AI studies and from the military. In both cases, the most vulnerable level from the perspective of falsification of truth is the human one (the cognitive level), which requires going through psychological and sociological studies that identify patterns of truth altering related to the limits of the human mind and the processes of simplification and generalization in judging and understanding. After following this theoretical approach, document analysis was selected as a meta-method of research (meta-analysis) on various studies that have catalogued the political biases of ChatGPT, in an effort to identify the source of truth alteration by AI.

1.1. AI versions

The almost 70-year-long history of the artificial intelligence (AI) concept has been known and intensively debated in articles that analyse, in fact, the current phase of the emergence of advanced technologies that have allowed large masses of users to interact with the chatbot produced by OpenAI from San Francisco, ChatGPT, and make it the star of simulated dialogue in the virtual environment. Research interest in various areas of knowledge has arisen as a result of the fact that a chatbot has been offered for free use to Internet users. Actually, AI is limited neither to simulated conversation software applications nor to applications like Generative Pre-trained Transformer (GPT), including the aforementioned ChatGPT star, which possess artificial intelligence. The issue of the new generation of artificial intelligence, AI 2.0, is not recent, and it has been continuously updated (Li et al., 2013; Leng et al., 2024), even if the reference element of the version remains *deep learning*. Some studies design the horizon of AI 3.0 version, which involve generative artificial intelligence (Gilbert et al., 2024; Howell et al., 2024). It was not generated by the appearance of this chatbot. The perspective on the dynamics of this generation was correctly addressed by researchers. On the one hand, the complex qualitative transformation of the information environment on the other hand, the increasing social demands have made it possible for AI to evolve and become a partner in the human-computer dialogue, with the role of stimulating human intelligence for on different levels: (1) only in human-machine dialogue; (2) by including humans and machines in networks and (3) by creating more complex ecosystem-like systems, such as intelligent cities (Pan, 2016).

Based on efforts to standardize the AI 2.0 concept and on the contribution of The High-Level Expert Group on Artificial Intelligence of European Commission (AI HLEG), Samoili et al. proposed in 2021 a standardized framework to define Artificial Intelligence system as a software – AI HLEG (2018) considers the possibility that AI systems also include hardware – that generate outputs for goals defined by HI, "influencing the environments they interact with" (Samoili et al., 2021).

As for the emergence of the information environment, this became possible due to the development of new technologies adapted to the steps towards the superior version of artificial intelligence, the quantitative development of data – one of the great challenges of AI was the transformation of big data into knowledge, a problem solved among others by the AlphaGo application developed by DeepMind – and their qualitative extension, meaning cross-checking of their viability. In short, it was possible due to the use of different media channels (including new media), that generated the Chinese understanding of them as *cross-media computing*, extended through further studies to the concept of *cross-media intelligence* (Pan, 2016). The decisive step towards the phase in which AI 2.0 entered the public debate, beyond the November 2022 moment of ChatGPT exposition, consists of reaching the stages of hybrid intelligence, defined as *human-machine hybrid-augmented intelligence*, which allowed, in the quantitative and qualitative comparison of different forms of intelligence in cooperation (and, at the same time, in competition), the development of intelligent autonomous systems, using artificial neural networks based on *machine learning* process.

1.2. Machine learning process

This process has moved beyond what it was 25 years ago, involving the automatic and autonomous development of computational algorithms through experience, but still relates to the same kind of understanding of learning as the enrichment of performance following learning experience (Mitchell, 1997). From a current perspective, the *machine learning* process is defined in relation to factuality and the truth value associated to it to find an approximate ground-truth: (1) the *learning* or *training* process involves the use of learning algorithms; (2) the *training data* are those used in the previously used process; (3) each example in the training data is called a *training example*, and their totality is a *training set*, and (4) the result of the learning process is a *ground truth* (or *fact*), as it was called by Zhou (2021).

From this definition of the *machine learning* process that is the basis of artificial intelligence, including AI 2.0 version, there results two limits regarding the corruption of the truth. The first is the selection of *training examples* that must be representative, with the lowest margin of error, at the level of factual reference reality. The second one results from the ability of the *ground-truth* approximation algorithm. Both sources of error depend on the dimensions of the reference information environment, therefore a chapter dedicated to it is strictly necessary. Both sources of truth altering are added to the classic errors in *machine learning* process, regardless of the types of learning and the types of learning algorithms (Cîrnu et al., 2023; Rotună et al., 2022). In the study regarding the comparative analysis of *machine learning* algorithms, Cîrnu et al. (2023) highlight the need to evaluate the model performances with different data sets, but in the case of these subtle forms of truth altering, evaluation is necessary but not sufficient to eliminate biases.

1.3. Defective truth as an AI ethics issue

The study of ethical issues regarding AI is a field of applied ethics almost as vast as the field of AI. This branch of ethics is rooted in an issue that predates the AI studies, focused on the ethical effects of implementing new technologies. Regarding aspects of AI ethics, most current articles deal with the study of ethical principles applicable to the area of interest (AI HLEG, 2018) and to the ethical practices (Pokholkova et al., 2024), and some of them to methods of translating principles into practices (Mittelstadt et al., 2016; Morley et al., 2023). Considering the principles of AI ethics, Morley et al. (2020) define the *machine learning* process in relation to a series of necessary and mandatory requirements, more precisely: beneficence, non-maleficence, autonomy, justice and explicability. The last one is important in relationship with the truth altering through AI as (i) traceability, that requires documentation regarding data sets and processes, (ii) explainability, regarding both AI processes and associated human decisions and (iii) interpretability (Morley et al., 2020). The same researchers consider explicability to be the „All-Encompassing Principle", which is not a first-order moral principle, but a second-order one (Morley et al., 2020) in the predominant mechanist and statistical perspectives in the field of AI. Once AI is seen as a socio-technical system or as a mixed AI-HI system, the issue of explicability, especially that of interpretability, becomes more and more important in ethical terms. It is, at the same time, more difficult to be analysed via quantitative methods and tools (Taddeo et al., 2024). Less numerous than quantitative research methods that propose measurement techniques suitable for AI, there are also qualitative studies that propose working guidelines (Franzke, 2022). Among the ethical concerns regarding the translation of principles into moral practices, there are misguided evidence, which emphasizes that the conclusions depend on the data used in defining training sets more as reliability than as neutrality, and unfair outcomes, which refer to the discriminatory potential of AI actions generated by the disproportionate impact of AI truth altering on a specific group or category of people (Mittelstadt et al., 2016).

Regarding the state-of-the-art in the ethical aspect of AI political biases, the studies are fewer, and they refer to a very small extent to HI dimension. The article of Uwe Peters (2022), for example, is illustrative of the algorithmic political bias analysis, although the general analysis framework assumes the consideration of the conscious (explicit) and unconscious (implicit) dimension regarding the political biases in human cognition. At the same level of algorithmic

political bias, the article of Tavishi Choudhary (2024) proposes, based on three tests, a comparative analysis of four AI models: ChatGP-4, Claude, Google Gemini, and Perplexity, highlighting liberal bias in the case of the first two, neutrality in the third case and conservative bias in the case of the fourth model. This study emphasizes predefined trends in AI models and advocates for building trust and integrity.

In relation to these aspects analysed by specialized literature, our study aims at highlighting political biases, regardless of whether they are generated by the algorithms or by the training sets, based on a meta-analysis of several researches carried out in relation to the most well-known and analysed AI model, ChatGPT. In order to identify the possible sources of truth altering through HI, the study was continued by considering the information environment (IE) as a whole, including AI-HI mix, focusing on the cognitive level of IE.

2. Information environment

2.1. Information environment in the field of AI studies

In the field of AI studies, the information environment is defined by the ternary structure CPH (cyber, physics, human), in which the first dimension generates the dynamics of new computational paradigm, including "perception fusion, "man-in-the-loop", augmented reality, and cross-media computing" (Pan, 2016). In essence, this structure is centred on the human dimension (*human-centric*), and the perspective of the complex exploitation of the information environment was developed through numerous researches, through which were identified: the general framework of the definition, starting from the basic subsystems, the technologies related, the applications used and the key characteristics, respectively the variety of similar terms or with a close semantic coverage are, such as: human-cyber-psychical system (HCPS), cyber-physical-human system (CPHS), human-in-the-loop cyber-physical system (HiLCPS), social cyber-physical system (SCPS) or cyber-physical social system (CPSS) (Wang et al., 2022). CPSS are considered the engine of the new computational paradigm, which involves the use of all dimensions of the information environment and the opening towards computational ubiquity or ubiquitous computing (as a perspective-shifting the focus to the first of the dimensions of the emerging information environment) resulting in a series of U-applications, U-systems, U-objects and U-services (Zeng et al., 2020).

2.2. Information environment in the military

In the security system, the information environment is understood as an operating environment serving as both an information medium and a resource. A different concept is not discussed, but only a special importance assigned to it. The information environment comprises a tangible component (the physical network through which information is transmitted) and an intangible component (made up of the information itself and the decision-making process) and overlaps almost perfectly with the similar concept in the field of AI. The CPH or CPS design naturally corresponds to the physical-information-cognitive (PIC) military perspective, in which the physical domain is assimilated to the real world and the communication channels through which the transfer of informational flows is carried out, and the information domain is associated to the contents of the previously mentioned flows (Cordray & Romanych, 2005). The cognitive domain is associated to the decision area or to the individual or collective consciousness that makes decisions. Collection, processing and dissemination of information lead to decision making, after completing the previous stages in the hierarchical cognitive model, i.e. information management, knowledge transmission and creation of shared understanding (JCOIE, 2018).

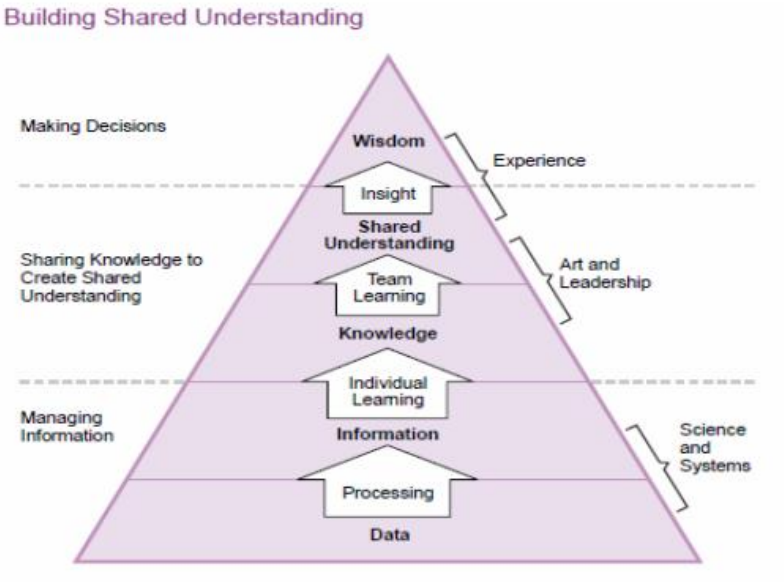

Figure 1. The cognitive hierarchical model used in building shared understanding, taken from *JCOIE* (2018)

The hierarchical cognitive model of the information environment (Figure 1) in military perspective discriminates between the possibilities of involving artificial intelligence, whose neural networks are useful in processing data and transforming it into information, being able to contribute to individual learning, except for the high phases of the process.

2.3. The cognitive level of the information environment

If the dynamic information environment is the one that determines the dynamics of AI, certainly the engine of change is not the physical or informational dimension, but the cognitive one, located at the highest stage of processing, shared understanding and responsible decision-making. Obviously, in order to reach this stage of AI involvement in decision-making, the sources of corrupting the truth must be reduced or eliminated, i.e. the base of *training examples* must be expanded, and the possibilities of *ground-truth* approximation must be increased. Moreover, it must be taken into account the fact that, in the context of the growing interest in AI and of a high rate of *veracity* and *equivocality* that artificial intelligence provides, there is a clear tendency to use it in decision-making, including in the political decision (Ciupercă et al., 2022), which can be dangerous, but it can also open the discussion on the political neutrality of artificial intelligence. The effects of such a high degree of informational accuracy can influence human perceptions and consequently affect human performance (Samuel et al., 2022), which means that once the high-accuracy digital output is not adapted to specifically human analogue processes, a third source of the corruption of truth follows from this. Specialists have begun to test the ability of AI to adapt to human cognition - for example, they propose an Adaptive Cognitive Fit (ACF), whose application framework involves an additional phase of comparing the AI decision with the human decision (Figure 2) -, but this design only contributes to reducing the gap between decision-making processes that assume shared understanding and ethics and those based on digital accuracy, expressed through the attributes of *veracity* and *equivocality*. Ethical issues are extremely complex and require an independent study, developed on the framework of the current analysis. They are not limited, for example, only to the ethics of using AI in the field of health in general, or in eHealth in particular (Gheorghe-Moisii et al., 2024), or to ethical considerations regarding governmental and social responsibility in e-Government applications (Dumitrache et al., 2023), but they refer to learning errors, generated by forms of truth altering or the human inclinations, which can be reduced with the consideration of the ACF model.

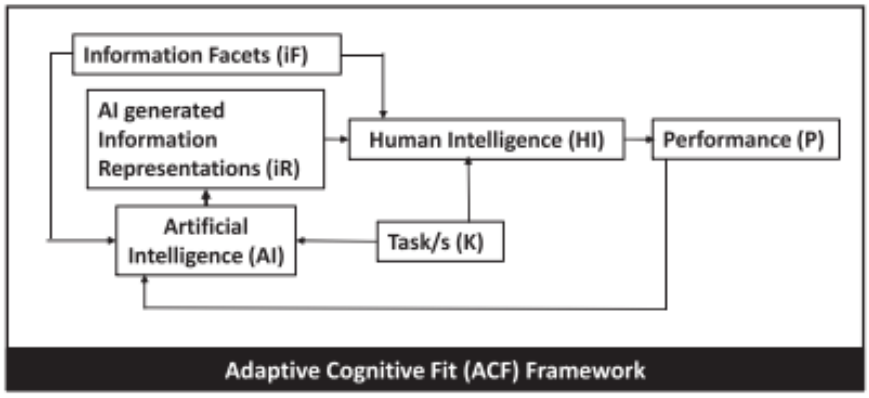

Figure 2. Implementation of ACF in increasing decision-making performance of AI, model taken from (Samuel et al., 2022)

Even if one of the drivers of the AI evolution and transition to 2.0 version is the information environment itself, and this information environment is centred on the cognitive dimension (human or social, in different terminologies), a certain level of cognitive dissonance continues to persist, and the need of keeping both AI and HI working for a high level of performance is the reality of the current version of artificial intelligence. The minimum requirement for AI-HI collaboration is that the truths should not be corrupted, not as an implication of AI - the accuracy class of statistical algorithms in *ground-truth* approximation can and is constantly improved - but as an implication of HI.

Through this chapter, it is underlined that the role of the cognitive dimension is fundamental in IE (regardless the paradigm in which it is analysed), and the truth altering depends not only on algorithms, but especially on the HI component in the AI-HI system. The analysis continues with the investigation of psycho-behavioural patterns in the truth altering in the AI-HI. The case study focuses on the displacement from the required neutrality in relation to other trends than those related to algorithms in the case of the most well-known AI model, following a meta-analysis of primary recent studies.

3. Altered information environment/altered truth through AI

3.1. Truth-altering in psycho-behavioural sciences

The main source of approximations in the CPS systems or in the AI-HI binomial is the human being. Consequently, it is important to understand the process of the corruption of truth starting with this source of vitiation. In fact, one of the old theories regarding the information environment is the one that attributes the main role to the social environment. One of the initiators of the cognitive consistency theories, psychologist Fritz Heider (1958), found that the social observer functions as a naive psychologist, looking for causes of the observed behaviour. His assumptions were followed up by important studies in the field of social psychology. Therefore, a touch of subjectivity in the attribution causes a nuance of meanings (transforming them from purely denotative to connotative meanings) but equally adds a nuance of psychological comfort to people, which explains the world resorting to cognitive shortcuts and to similar models avoiding the tension caused by cognitive inconsistency. It is obvious that the attribution process, assuming a mental path that tends to reduce tension, i.e. resorting to explanatory paths as a shortcut of the process, leads to attribution errors, some deliberate – through ideological alignment with a type of thinking – others motivated by this very tendency of conservation of energy, which manifests itself as “laziness of thought”. Attributions therefore represent a form of simplification in the judgment of the observer, an adaptation to what Walter Lippmann (1922) called “the images in our minds”, which determines the construction of a simplified version of the world and the transformation of the environment itself into a quasi-environment susceptible to satisfy the need to understand the world, as a schematization of the world (Yzerbyt & Schadron, 2002). This is how the issue of

stereotypes arises in the scientific debate. The stereotypes based on simplification precede reasoning processes and, unfortunately, sometimes even replace them.

From the psycho-social perspective, the process of simplification was later analysed by numerous researchers, including Gordon W. Allport (1954), who considers it necessary to reduce the much too complex environment by simplification, in order to create the possibility of encompassing reality and its appropriate mapping. Further studies by Tajfel & Wilkes (1963) substantially contribute to the understanding of the process of categorization, which entails the use of categories as stereotypes by emphasizing similarities and differences, which leads to a second important process in corrupting the truth, namely generalization. The line of models and theories will be abandoned, as well of validations in the field of psychology of cognition, underlying the two prejudicial processes, simplification and generalization, which contribute in line with the cognitive consistency theories to stereotypical thinking and, implicitly, to the evaluation based on this thinking. Stereotypes originate from a perceptual illusion encouraged by the natural tendency to conserve mental energy. They are anchored in reality (Yzerbyt & Schadron, 2002) and contribute to the creation of an alternative reality, exploitable through mass media. However, as long as numerous current studies position AI in the proximity of mass media (for example, de Lima-Santos & Ceron, 2022), by impacting media markets, by searching and exploiting consumer preferences in this market, by taking over some journalistic tasks, but especially by personalizing the content to be communicated (Hermann, 2021) and by reshaping, completely changing or even overturning the theories of media communication, because the AI devices used in communication are different from the classical communication channels and require rethinking the theoretical framework in the field of communication sciences (Guzman & Lewis, 2020), bringing into question their definition from a procedural perspective, as Human-Machine Communication (HMC), despite the general assumption of communication as an eminently human function.

3.2. Research methodology

For pointing up biases generated by truth altering via AI, as long as the sources of error are not directly attributable to machines or to *machine learning* process, and as long as it is difficult to quantitatively establish whether the source of the bias is human or mixed, quantitative research is inadequate. In addition, qualitative research methods are suitable for the study of new phenomena, such as political biases of AI, and for cases where quantitative research methods cannot provide precise data regarding the foundations, reasons and ways through which the previously mentioned biases occur.

Qualitative research is recommended for locating the source of altering truth, starting from the quantitative, qualitative or mixed data as results of previous studies. Therefore, a meta-analysis as a qualitative research method was opted (as defined by Schreiber, 2008), i.e. interpretative meta-analysis, and for qualitative meta-synthesis (Thorne, 2008; Xu, 2008), starting from primary studies with distinct objectives. This meta-synthesis uses research results to be aggregated in a synthetic research, located on a higher level of generality compared to primary studies, able to define the patterns of those vaguely convergent studies (Levitt, 2018). The method is appropriate for this research because it exceeds the contextual boundaries of each primary study and extends conceptual trends to provide a synthesis image (Stall-Meadows & Hyle, 2010) or a big picture of a broad phenomenon, regardless of whether the primary case studies are quantitative, qualitative or mixed. In addition, qualitative meta-analysis subsumes trends of the answers to *who* and *what* questions specific to quantitative research to highlight (in summary) the answers to the *why* questions specific to qualitative research. The stage of theoretical bias analysis as it result from subchapter 1.3 of the current research was followed and a series of criteria regarding the primary empirical studies were established. To identify them, the following sampling criteria was used: (i) the temporal criterion, applied from the appearance of ChatGPT until now, that is, between November 30, 2022 and August 15, 2024; (ii) publication language criterion: English only and (iii) bias type criterion: political bias only. Five primary studies from a series of studies focused on AI biases, converging as results even if they are different in terms of objectives, methods or corpus, qualified for the meta-analysis.

3.3. Altering the truth through AI. ChatGPT's political biases

The main source of corrupting the information environment is the result of a social representation, which is a dynamic and open phenomenon defined by a dynamic concept (Marková, 2004). Social representations and the product of representation are like ideologies, a set of *social objects* (Deconchy, 1995). The differences are the need for contingent representation in the former case and the need for ideological alignment in the latter. In the AI-HI interaction, this source alters the decisional, cognitive, human or social level in the different models of information environment, contributing to the reduction of the *training examples* by simplification or by their ideologically motivated selection, under the circumstances of a generalization that differs from the classical one, specific to the human stereotypes. The classic *Ego-Alter-Object* triangle, fundamental in the design of theories of cognitive consistency, also fundamental in the field of social representations, changes itself by gaining a higher degree of objectivity in relation with the object, through generalization based on a corresponding statistical inventory, but maintaining the *Ego-Alter* relationship in the area of possibilities of inducing a certain perspective in small and imperceptible doses, unintentionally, as a simplified social representation, or intentionally, as an ideological alignment.

The most eloquent example is the very chatbot launched on the free market, from which the presentation of the perspective on corrupting the truth was started, ChatGPT. Its response tendency based on Political Compass, Stemwijzer test, Politieke Oriëntatie Test, is rather left (centre-left) and libertarian rather than right and authoritarian (Van der Broek, 2023; Rozado, 2023a; Rozado, 2023b). For example, Rozado administered to ChatGPT 15 sets of known political orientation tests, including Political Spectrum Quiz, Political Compass, Political Ideology Selector, but also Pew Political Typology Quiz, to which the chatbot refused to answer several questions. Even if to the direct questions AI declared that it is neutral from a political point of view, the result of all 15 tests was explicit and formulated as follows by Rozado (2023b): when administering several political orientation tests, ChatGPT provides answers that attest a left-wing political leaning.

Another study confirming the political leaning of ChatGPT (Rutinowski et al., 2024) illustrates the prevalence of pro-libertarian over authoritarian responses and interrogates the contents to place them on the progressive-conservative axis. The result is also eloquent: Chat GPT is a form of intelligence that moves away from its imagined neutrality, being located in the vicinity of progressive values. The study of Rutinowski et al. proposes an analysis based on a methodology that involves querying the March 2023 Version of chatbot (ChatGPT-3.5) based on the political compass test, made by 62 items with answers scaled based on a four-point Likert scale and on iSideWith questionnaires corresponding the seven-member countries of the G7, as well as other tests in addition to the political affiliation instruments. In the case of political biases analysis, the questionnaires were applied to ChatGPT ten times, and the average score was -6.48 out of 10 (standard deviation: 0.95) on the progressive/conservative axis, respectively -5.99 out of 10 (standard deviation: 0.73) on the authoritarian/libertarian axis.

Studies regarding the political biases have multiplied considerably in the last months. The ChatGPT political biases were analysed and presented at CRIFST conference at the Romanian Academy (Lesenciuc, 2023), and resulted that the subsequent results are largely similar, with values in generally more moderate (Fujimoto & Takemoto, 2023) than those of the first study (Rozado, 2023a). The gain in accuracy of this study, assuming the questioning of the same version of ChatGPT, involved the use of seven tests of political orientation and responses scaled on a five-point Likert scale, with the questions repeated 20 times. Even though the average scores on the seven sets of tests varied, even if there are political where the tests demonstrated the neutrality of the chatbot, the results indicate a leftward orientation within the political spectrum, and this orientation is relevant as long as it results from the same test applied in all previous studies, The Political Compass. However, the results obtained indicate a lower degree of political biases than in the case of Rozado's studies, as follows: the IDRLabs test on political coordinates indicated a political quasi-neutrality (2,8% right-wing, 11,1% liberal), the Eysenck test highlighted the tendency of 12,5% of the responses towards the radical area, the political spectrum quiz indicated 16,9% left-wing responses and 4,9% authoritarian, the IDRLabs test on ideologies was largely irrelevant, indicating no trends in responses very far from the centre of political spectrum, the eight

values political test showed diplomatic, civil and social neutrality, but the most relevant one, regarding the political compass, revealed a left-libertarian inclination, more precisely 30,0% toward the political left, and 48,2% towards the libertarian perspective (Fujimoto & Takemoto, 2023).

Numerous articles highlight particularities of biases depending on the geographical area or other indicators (Motoki et al., 2024) – some of them emphasizing biases in foreign policy, or directly regarding the implication of the war in Ukraine (Urman & Makhortych, 2023) – and others analyse ideological, social, or economic patterns.

To foreground the trends regarding political biases from the primary studies, the trends were pre-coded on the three axes that are subject of the current research, through a Five-Point Likert scale, as follows: (i) right-left: +2 right-wing, +1 centre-right, 0 centre, -1 centre-left, -2 left-wing; (ii) libertarian-authoritarian: +2 libertarian, +1 moderate libertarian, 0 moderate, -1 moderate authoritarian, -2 authoritarian, and (iii) progressive-conservative: +2 progressive, +1 moderate progressive, 0 moderate, -1 moderate conservative, -2 conservative. This scaling only partially corresponds to distinct political systems in Europe or in the whole world, but it can be applied to all such systems. The meta-analysis shows the following (see in Table 1):

Table 1. Scaling of primary study trends

	right-left axis	libertarian-authoritarian axis	progressive-conservative axis
Rozado (2023a)	-2		
Rozado (2023b)	-2		
Van der Broek (2023)	-1	+1	+1
Fujimoto & Takemoto (2023)	-1	+1	
Rutinowski et al. (2024)	-1	+1	+2

The meta-analysis of the results of the primary studies is a qualitative one, which is why calculation equations cannot be applied to get average values. The studies assumed a greater or lesser number of applied tests, with a lower or higher repeatability, the latter two being more accurate. Any other empirical study could only reveal similar data, varying according to the language used, the number of repetitions, and the calibration of political tests in the language used. Instead, the trends are eloquent: ChatGPT shows a shift to the left, and towards libertarian and progressive values. These results are very important since, taking into account the liberal bias of ChatGPT from the perspective of algorithmic political bias (Choudhary, 2024), the left leaning can be due to any phase of AI machine learning process. This certainty as an orientation - all the studies, including those already mentioned, but which are not the subject of the meta-analysis -, doubled by the uncertainty regarding the source of truth altering and displacement from the neutral zone of the political spectrum, calls for the discussion of aspects related to AI ethics, which concern to a lesser extent principles or practices or methods of translating principles into practices.

The conclusion of one of these studies is logical and, unfortunately, more than worrying. Rozado (2023b) believes that regardless of the AI's political leaning, it is not the percentage of distance from the median values that is significant, but the implications for society, the answers that will be considered by users to be politically neutral. The AI claim of political neutrality should be a source of human concern, especially regarding the normative aspects, "given their potential for shaping human perceptions and thereby exerting societal control. (Rozado, 2023b)."

This is only the result of an analysis in relation to political orientation. Less interesting to the general public and, implicitly, to researchers, the cultural ideologies can be even more dangerous in terms of accurately rendering factuality, excepting any interference of any kind, within the limits of some neutral language from all perspectives.

4. Conclusions

Studies regarding the ideological corruption of artificial intelligence require a double interpretation, which can be summed up in a reference question: Are ChatGPT biases due to a *machine learning* process based on *training examples* altered in relation to the centre of political spectrum and the central area of political management practices, or is the sum of these training examples consistent with the values identified by David Rozado, Merel Van der Broek and Jérôme Rutinowski? Unfortunately, there are no studies conducted in this regard. From a technical perspective, these studies cannot even be carried out using the large mass of information on the Internet from which artificial intelligence learns.

To rephrase, the issue boils down to a simple question: is ChatGPT left-leaning, libertarian and progressive, or is the Internet itself oriented towards these values? If the corrupting error belongs to the human intelligence that selected *training examples* predominantly left, libertarian and progressive, ChatGPT and the other forms of artificial intelligence will calibrate over time, once the number of examples offered for learning expands. If, instead, the enormous mass of information hosted on the Internet is deviated from the neutral zone on each of the three axes considered, the problem does not involve solutions. In this case, on the one hand AI provides a real picture of the political biases of HI, of the world in totality, on the other hand it multiplies, in a virtual mirror, an image of a decentralized world that continues to contribute to decentring. The awareness of ideological nuances of this world through artificial intelligence that reproduces and multiplies it becomes in this case the main alert signal, not in relation to the ethics of AI, or, in particular, to the ethics of ChatGPT, but to the very ethics of humankind surprised and overtaken by information and knowledge. AI offers a simplified and generalized representation of the world. Such a hypothesis would come to provide an answer to how AI can overcome HI from a certain perspective, that of identifying its tendencies or predispositions, impossible to be aware of the reverse.

REFERENCES

Allport, G. W. (1954) *The Nature of Prejudice*. Boston, Addison-Wesley.

Choudhary, T. (2024) Political Bias in AI-Language Models: A Comparative Analysis of ChatGPT-4, Perplexity, Google Gemini, and Claude. *Preprints*. 2024071274. doi:10.20944/preprints202407.1274.v1.

Ciupercă, E. M., Donnelly, N., Gartland, A. & Stanciu, A. (2022) The Digital Divide in Education - Macrocultural Comparative Analysis between Ireland and Romania. *IFAC - PapersOnLine.* 55(39), 99-104. doi:10.1016/j.ifacol.2022.12.018.

Cîrnu, C. E., Vasiloiu, I.-C. & Rotună, C.-I. (2023) Comparative analysis of the main machine learning algorithms for the automatic recognition of fake news. *Romanian Journal of Information Technology and Automatic Control.* 33(1), 57-66. doi:10.33436/v33i1y202305.

Cordray III, R. & Romanych, M. J. (2005) Mapping the Information Environment. *IO Sphere*. Summer, 7-10.

Deconchy, J.-P. (1995) *Credințe și ideologii*. Iași, Editura Polirom.

de Lima-Santos, M.-F. & Ceron, W. (2022) Artificial Intelligence in New Media: Current Perceptions and Future Outlook. *Journal. Media.* 3(1), 13-26. doi:10.3390/journalmedia3010002.

Dumitrache, M., Stănescu, A. C. & Paraschiv, E.-A. (2023) Digitalizarea și inteligența artificială în aplicațiile de e-Guvernare. *Romanian Journal of Information Technology and Automatic Control.* 33(3), 43-54. doi:10.33436/v33i3y202304.

Franzke, A. S. (2022) An exploratory qualitative analysis of AI ethics guidelines. *Journal of Information, Communication and Ethics in Society.* 20(4), 401-423. doi:10.1108/JICES-12-2020-0125.

Fujimoto, S. & Takemoto, K. (2023) Revisiting the political biases of ChatGPT. *Frontiers in Artificial Intelligence.* 6, 1-6. doi:10.3389/frai.2023.1232003.

Gheorghe-Moisii, M., Gheorghe, C.-G. & Soviany, S. (2024) Ethical considerations on the use of AI technology in eHealth applications for neurodegenerative diseases. *Romanian Journal of Information Technology and Automatic Control.* 34(1), 97-108. doi:10.33436/v34i1y202409.

Gilbert, S, Kather, J. N. & Hogan, A. (2024) Augmented non-hallucinating large language models as medical information curators. n*pj Digital Medicine.* 7, 1-5. doi:10.1038/s41746-024-01081-0.

Guzman, A. L. & Lewis, S. C. (2020) Artificial intelligence and communication: A Human-Machine Communication research agenda. *New media & society.* 22(1), 70-86. doi:10.1177/1461444819858691.

Hermannn, E. (2021) Artificial intelligence and mass personalization of communication content – An ethical and literacy perspective. *New media & society*. 24(5). doi:10.1177/14614448211022702.

Heider, F. (1958) *The Psychology of Interpersonal Relations.* Hoboken, NJ, United States, John Wiley & Sons, Inc.

High Level Expert Group on Artificial Intelligence (AI HLEG) (2018, 18 December) *A definition of AI: Main capabilities and scientific disciplines*. Brussels, European Commission.

Howell, M. D., Corrado, G. & DeSalvo, K. B. (2024) Three Epochs of Artificial Intelligence in Health Care. *JAMA*. 331(3), 242-244. doi:10.1001/jama.2023.25057.

Joint Chiefs of Staff (2018) *Joint Concept for Operating in the Information Environment (JCOIE).* Washington, DC, United States Department of Defense (DoD).

Katz, Y. (2020) *Artificial Whiteness. Politics and Ideology in Artificial Intelligence.* New York, Columbia University Press.

Leng, J., Zhu, X., Huang, Z., Li, X., Zheng, P., Zhou, X., Mourtzis, D., Wang, B., Qi, Q., Shao, H., Wan, J., Chen, X., Wang, L. & Liu, Q. (2024) Unlocking the power of industrial artificial intelligence toward Industry 5.0: Insights, pathways, and challenges. *Journal of Manufacturing Systems.* 73, 349-363. doi:10.1016/j.jmsy.2024.02.010.

Lesenciuc, A. (2023) Adevăruri viciate. Amprente ideologice ale AI. *Conferința CRIFST (Comitetul Român de Istoria şi Filosofia Ştiinţei şi Tehnicii) Inteligența Artificială (IA) și noile științe bazate pe tehnologiile IA, 21 November, 2023, Bucharest, The Romanian Academy.* https://www.crifst.ro/adevaruri-viciate-amprente-ideologice-ale-ai/.

Levitt, H. M. (2018) How to conduct a qualitative meta-analysis: Tailoring methods to enhance methodological integrity. *Psychoterapy Research*. 28(3), 367-378. doi:10.1080/10503307.2018.1447708.

Li, Y.-F., Wang, H. & Sun, M. (2013) Chat GPT-like large-scale foundation models for prognostics and health management: A survey and roadmaps. *Reliability Engineering & System Safety.* 243, 109850. doi:org/10.1016/j.ress.2023.109850.

Lippmann, W. (1922) *Public Opinion*. New York, Hartcourt, Brace and Company.

Marková, I. (2004) *Dialogistica și reprezentările sociale.* Iași, Polirom.

Mitchel, T. M. (1997) *Machine Learning.* New York, McGraw-Hill.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016) The ethics of algorithms. Mapping the debate. *Big Data & Society*. 3(2). doi:10.1177/20539517116679679.

Morley, J., Floridi, L., Kinsey, L. & Elhalal, A. (2020) From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics.* 26, 2141-2168. doi:10.1007/s11948-019-00165-5.

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M. & Floridi, L. (2023) Operationalising AI ethics: barriers, enablers and next steps. *AI & Society.* 38. 411-423. doi:10.1007/s00146-021-01308-8.

Motoki, F., Pinho Neto, V. P. & Rodrigue, V. (2024) More human than human: measuring ChatGPT political bias. *Public Choice*. 198, 3-23. doi:10.1007/s11127-023-01097-2.

Pan, Y. (2016) Heading toward Artificial Intelligence 2.0 *Engineering.* 2(4), 409-413. doi:10.1016/J.ENG.2016.04.018.

Peters, U. (2022) Algorithmic Political Bias in Artificial Intelligence Systems. *Philosophy & Technology*. 35 (25), 1-23. doi:10.1007/s13347-022-00512-8.

Pokholkova, M., Boch, A., Hohma, E. & Lütge, C. (2024) Measuring adherence to AI ethics: a methodology for assessing adherence to ethical principles in the use case of AI-enabled credit scoring application. *AI and Ethics.* doi:10.1007/s43681-024-00468-9.

Rotună, C.-I., Dumitrache, M. & Sandu, I.-E. (2022) Assessment of Machine Learning algorithms for automated monitoring. *Romanian Journal of Information Technology and Automatic Control.* 32(3), 73-84. doi:10.33436/v32i3y202206.

Rozado, D. (2023a) The Political Bias of ChatGPT – Extended Analysis. *Rozado's Visual Analytics.* https://davidrozado.substack.com/p/political-bias-chatgpt [Accesed 19th November, 2023].

Rozado, D. (2023b) The Political Biases of ChatGPT. *Social Sciences.* 12(3), 148. doi:10.3390/socsci12030148.

Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., Roidl. M. & Pauly, M. (2024) The Self-Perception and Political Biases of ChatGPT. *Human Behavior and Emerging Technologies*. 2024(7115633), 1-9. doi:10.1155/2024/7115633.

Samoili, S., Lopez Cobo, M., Delipetrev, B., Martinez-Plumed, F., Gomez Gutierrez, E. & De Prato, G. (2021) AI Watch. Defining Artificial Intelligence 2.0. *JRC Publications Repository*. doi:10.2760/019901, JRC126426.

Samuel, J., Kashyap, R., Samuel, Y. & Pelaez, A. (2022) Adaptive cognitive fit: Artificial intelligence augmented management of information facets and representations. *International Journal of Information Management.* 65(102505), 1-19. doi:10.1016/j.ijinfomgt.2022.102505.

Schreiber, J. B. (2008) Meta-Analysis. In Given, L.M. (ed.), *The SAGE Encyclopedia of Qualitative Research Methods*, vol. 1&2. Los Angeles, CA, SAGE. pp. 506-507.

Stall-Meadows, C. & Hyle, A. (2010) Procedural methodology for a grounded meta-analysis of qualitative case studies. *International Journal of Consumer Studies.* 34(4), 412-418. doi:10.1111/j.1470-6431.2010.00882.x.

Tajfel, H. & Wilkes, A. L. (1963) Classification and quantitative judgement. *British Journal of Psychology.* 54(2), 101–114. doi:10.1111/j.2044-8295.1963.tb00865.x

Taddeo, M., Blanchard, A. & Thomas, C. (2024). From AI Ethics Principles to Practices: A Teleological Methodology to Apply AI Ethics Principles in the Defence Domain. *Philosophy & Technology.* 37 (42), 1-21. doi:10.1007/s13347-024-00710-6.

Thorne, S. E. (2008) Meta-Synthesis. In Given, L.M. (ed.), *The SAGE Encyclopedia of Qualitative Research Methods*, vol. 1&2. Los Angeles, CA, SAGE. pp.510-513.

Urman, A. & Makhortykh, M. (2023) The Silence of the LLMs: Cross-Lingual Analysis of Political Bias and False Information Prevalence in ChatGPT, Google Bard, and Bing Chat. *OSFPreprints.* 1–11. doi:10.31219/osf.io/q9v8f.

Van der Broek, M. (2023) *ChatGPT's left-leaning liberal bias.*https://www.universiteitleiden.nl/binaries/content/assets/algemeen/bb-scm/nieuws/political_bias_in_chatgpt.pdf [Accesed 19th November, 2023].

Wang, B., Zheng, P., Yin, Y., Shih, A. & Wang, L. (2022) Toward human-centric smart manufacturing: A human-cyber-physical systems (HCPS) perspective. *Journal of Manufacturing Systems*. 63, 471-490. doi:10.1016/j.jmsy.2022.05.005.

Xu, Y. (2008) Methodological Issues and Challenges in Data Collection and Analysis of Qualitative Meta-Synthesis. *Asian Nursing Research*. 2(3), 173-183. doi:10.1016/S1976-1317(08)60041-9.

Yzerbit, V. & Schadron, G. (2002) *Cunoașterea și judecarea celuilalt.* Iași, Polirom.

Zeng, J.; Yang, L. T.; Lin, M.; Ning, H. & Ma, J. (2020) A survey: Cyber-physical-social systems and their system-level design methodology. *Future Generation Computer Systems*. 105, 1028-1042. doi:10.1016/j.future.2016.06.034.

Zhou, Z.-H. (2021) *Machine Learning.* Singapore, Springer.

Adrian LESENCIUC is a specialist with multidisciplinary training, holding a bachelor's degree in artillery and anti-aircraft missiles (electromechanical engineering) obtained at "General Bungescu" Military Institute of Artillery and Anti-aircraft Missiles in Brasov. He completed post-graduate studies, a master's degree in communication at the National University of Political Studies and Administration (NUSPA). Adrian Lesenciuc also completed two doctoral programs: one in military and information sciences at "Carol I" National Defense University and the second in communication sciences at the National University of Political Studies and Administration. He is habilitated in intelligence and national security at "Mihai Viteazul" National Intelligence Academy. He is also a professor at "Henri Coanda" Air Force Academy in Brasov, where he teaches Communication Theory, Information Warfare and Media Influence on Security Systems. In the past he held the position of dean of the Faculty of Air Security Systems, and he is currently the vice-rector for science at "Henri Coanda" Air Force Academy. Adrian Lesenciuc is the author of several scientific works, the most recent being *Hybrid warfare or the return through doctrinal dissimulation to absolute war* (CTEA, 2023). In parallel, he is a member and specialist in numerous project teams, being also the coordinator of the research project *Measuring the parameters of mental preparation for the battlefield (Combat Mindset).* His areas of interest include: intercultural communication, semiotics, information operations, history of military thought, security studies and cultural intelligence. His academic and research contributions have a significant impact on the understanding and development of security strategies in the information age.

Adrian LESENCIUC este un specialist cu pregătire multidisciplinară, deținând o diplomă de licență în artilerie și rachete antiaeriene (inginerie electromecanică) obținută la Institutul Militar de Artilerie și Rachete Antiaeriene „General Bungescu" din Brașov. A urmat studii postuniversitare, finalizând un master în comunicare la Școala Națională de Studii Politice și Administrative (SNSPA). Adrian Lesenciuc a parcurs, de asemenea, două programe de doctorat: unul în științe militare și informații la Universitatea Națională de Apărare „Carol I" și cel de-al doilea în științele

comunicării la Școala Națională de Studii Politice și Administrative. Este cadru didactic abilitat în informații și securitate națională la Academia Națională de Informații „Mihai Viteazul”. De asemenea, este profesor universitar la Academia Forțelor Aeriene „Henri Coandă” din Brașov, unde predă Teoria comunicării, Război informațional și Influența mass-media asupra sistemelor de securitate. În trecut a deținut funcția de decan al Facultății de Sisteme de Securitate Aeriană, iar în prezent este prorector pentru cercetare științifică la aceeași instituție. Adrian Lesenciuc este autorul mai multor lucrări științifice, cea mai recentă fiind *Războiul hibrid sau întoarcerea prin disimulare doctrinară la războiul absolut* (CTEA, 2023). În paralel, este membru și specialist în numeroase echipe de proiect, fiind și coordonator al proiectului de cercetare *Măsurarea parametrilor pregătirii mintale pentru câmpul de luptă* (Combat Mindset). Domeniile sale de interes includ: comunicare interculturală, semiotică, operații informaționale, istoria gândirii militare, studii de securitate și intelligence cultural. Contribuțiile sale academice și de cercetare au un impact semnificativ asupra înțelegerii și dezvoltării strategiilor de securitate în era informațională.