Essays

ChatGPT: Challenges and Legal Issues in Advanced Conversational AI

Amalia Diurni[*] and Giovanni Riccio[**]

Abstract

This paper analyzes ChatGPT, the first mass-deployed chatbot using Generative Pre-trained Transformer (GPT) technology. While there is unprecedented potential for application of this technology, many concerns with multidimensional value have arisen, all of which are inherent in pre-trained AI, generative AI, and communicative AI profiles. Investigation by the Italian Data Protection Authority and political debates have revealed the inadequacy of existing regulations, and call for more flexible regulation focusing on transparency, risk assessment and explainability. This article underscores the ontological vulnerability of individuals in their interaction with AI, drawing attention to a new legal category—digital vulnerability. As communicative AI transforms society, this paper, starting from the data protection regulations, emphasizes the need for an adaptive and comprehensive regulatory framework to safeguard against emerging challenges.

I. Introduction

ChatGPT is a chatbot, a '(ro)bot capable of simulating a conversation with a human being'.[1] The acronym GPT, which stands for 'Generative Pre-trained Transformer', is particularly auspicious: a 'transformer' (an algorithm), which allows the neural network to focus only on relevant data (not in terms of significance, but in terms of the number of hits), thus achieving more accurate results; a 'pre-trained' transformer (one fed with data sets selected by the programmer, or self-learned through information drawn from the web); functioning thanks to a complex 'generative' neural network which, as such, is able to autonomously generate original content in its interaction with the user.

Each of these features deserves to be explored in depth for the multiple possibilities that the use of this technology can offer. Practical applications, such

[*] Full Professor of Comparative Private Law, University of Rome Tor Vergata. Author of the text up to and including part IV. The research for this article was funded by the Italian Ministry of Universities and Research (MUR) under the PRIN 2020 project 'Digital Vulnerability in European Private Law' and by the Istituto Italiano di Studi Germanici under the research project 'Instabilità del diritto e vulnerabilità digitale'.

[**] Full Professor of Comparative Law, University of Salerno. Author of parts V-VIII.

[1] This definition is from the Grande dizionario italiano dell'uso (GRADIT), which dates the use of the expression to 2004. For further details see Lucia Francalanci of the Accademia della Crusca, *Una risposta col bot*, 16 November 2021, available at urly.it/3zwax (last visited 10 February 2024).

as enabling the deaf to hear and the blind to see,[2] can be as miraculous as they can be threatening to the rights of the parties involved and to the ordinary running of social, political and economic interactions. The potential for use is unprecedented and increases significantly with Socratic models (modular frameworks in which multiple pretrained models can be trained to exchange information among each other and capture new multimodal capabilities without requiring fine-tuning)[3] or with cross-technology models.[4] In cross-technology models, bots can mimic human intelligence and create textual content as well as audio, video, images and computer codes. They could be used as personal assistants to restore or improve the capabilities of people with disabilities or to increase the performing quantity, speed and quality of personal or work activities; they could be a valuable professional tool to support research, education and security in any field.

However, just as it is impossible to predict the many uses and applications of bots, it is impossible to predict the dangers associated with them. It is therefore all the more important to look closely at how a chatbot works in order to focus the analysis on the profiles of human vulnerability that might exist when interacting with such a form of artificial intelligence. Digital vulnerability appears to be a universally intrinsic feature of individuals interacting with AI and a situationally extrinsic feature of individuals interacting between themselves by means of AI. Thus defined, digital vulnerability may amount to a new macro-category of private law, from which all regulatory instruments to protect individuals and guarantee the role of the law and that of public and private institutions may be derived.

ChatGPT was developed and launched by OpenAI, but other chatbots are currently competing to improve their performance and diversify and broaden their applications.[5] In general, with respect to any currently available chatbot, there seem to be three main factors from which threats may arise: (i) training, (ii) text generation, and (iii) communicative power. This article analyzes these three aspects before turning to the ChatGPT case and to the measures taken by the Italian Data Protection Authority.

II. The Pretrained AI

[2] On the future frontiers of AI applications and the ethical issues involved, see F. Jotterand and M. Ienca, *The Routledge Handbook of the Ethics of Human Enhancement* (New York: Routledge, 2023), in particular Part IV on cognitive enhancement 187-250 and Part VI on human enhancement and medicine, 307-356.

[3] A. Zeng et al, 'Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language' *arXiv:2204.00598* (2022).

[4] On 25 September 2023, OpenAI announced new ChatGPT applications with voice and image capabilities, allowing users to use ChatGPT as an assistant, ready to be engaged in back-and-forth conversation in daily life: http://tinyurl.com/zsr4bjmz.

[5] Along with OpenAI's GPT-3 and 4, popular LLMs include open models such as Google's LaMDA and PaLM (the basis for Bard), Hugging Face's BLOOM and XLM-RoBERTa, Nvidia's NeMO, XLNet, Co:here, and GLM-130B.

Chatbots are an evolution of Large Language Models (LLMs) – algorithmic models capable of processing natural language inputs and predicting the next word, or completing an entire sentence, according to what probabilistically best fits the input (Natural Language Processing (NLP)). But chatbots go beyond this; not only are they able to process natural language, they are also able to generate it (Natural Language Generation (NLG)). To do this, NLGs are controlled by parameters that help the model choose between several possible responses: the higher the number of parameters, the better the performance. OpenAI's ChatGPT-4 has allegedly reached a trillion parameters, whilst GPT-3, only a few months earlier, had 175 billion parameters. These bots consist of neural networks that are 'educated' with data input/output sets. They are transformers that have been pretrained by means of machine learning. Machine learning can be self-supervised, semi-supervised or unsupervised, or even 'reinforced' by the environment in which it operates. Bots, therefore, learn to produce text by ingesting information.[6]

This new AI frontier is the outcome of an increase in computational power, requiring huge server farms[7] and ever-increasing amounts of data.[8] And it is precisely in relation to such data that problems arise. Pre-training of chatbots is achieved through enormous quantities of web content, regardless of either its quality or source. The quality of chatbot outputs is therefore affected by both the quality and quantity of its inputs (data sets and prompts).[9] The first threat, therefore, specifically concerns the unreliability of chatbot responses. Such unreliability may range from slight (ie responses that merely contain inaccurate or incomplete information) to serious (ie responses that are entirely wrong, if not utterly absurd – as in cases identified by data analysts as 'hallucinations').[10] The

[6] The conditions for chatbot technology are machine learning in the unsupervised and deep version together with the indispensable availability of big data, which is linked to the spread of the participatory web. Each web user produces data in a voluntary or automated way, through participation in social media or web browsing or by means of GPS tracking or Internet of Things: R. Kitchin, 'Big Data, New Epistemologies and Paradigm Shifts' *Big Data & Society*, 1-12 (2014); I. Goodfellow et al, *Deep Learning (Adaptive Computation and Machine Learning)* (Cambridge, *The MIT Press*, 2016).

[7] This is accompanied by all the questions regarding the economic and environmental sustainability of theuse of such server farms, as emerges from the Report 'The Digital Revolution and Sustainable Development: Opportunities and Challenges', prepared by The World in 2050 initiative and published by the IIASA, Laxenburg, Austria, 2019, available at urly.it/3zx0r (last visited 10 February 2024). In 2022 the UN Office of the Secretary-General's Envoy on Technology launched an Action Plan for a Sustainable Planet in the Digital Age: urly.it/3zx0s. See also P. Sacco et al, 'Sustainable Digitalization: A Systematic Literature Review to Identify How to Make Digitalization More Sustainable', in Y. Borgianni at al eds, *Creative Solutions for a Sustainable Development* (IFIP: Springer 2021), 14-30.

[8] One of the limits in the capacity and quality of chatbot responses to prompts is precisely the extent and accuracy of the data sets with which they are educated, so much so that a new chatbot market tailored to the needs of customers, professionals and businesses has already emerged.

[9] In addition to the market for 'professional' chatbots, a market for prompt engineering has also developed.

[10] L. Arnaudo and R. Pardolesi, 'Ecce robot. Sulla responsabilità dei sistemi adulti di intelligenza artificiale' *Danno e responsabilità*, IV, 409-417 (2023).

risk is that of an exponential increase in dis- and misinformation, whether voluntary or involuntary. Moreover, the application does not provide for a human expert in the loop, capable of assessing the correctness, genuineness and truthfulness of responses or of correcting or eliminating erroneous answers. Nor is it possible to prevent dissemination of such responses across the web, or their disguised or fraudulent use.[11] The risk is further increased by the fact that chatbots answer user prompts in such an affable and linguistically precise manner that the average user can be misled as to the reliability of responses.[12]

Addressing the propagation of fake news is one of the most urgent issues on the political agendas of most countries in the world.[13] The uncontrollability of the content of user prompts and of chatbot responses has led China, Russia, North Korea, Cuba, Syria and Iran to block access to ChatGPT within their national borders.[14] In these countries, the reason for the ban is the protection of public order[15] (and not the protection of users, which – as shall soon be shown – was the primary concern of the Italian Data Protection Authority). Chinese companies have created and launched their own chatbots as an alternative to commercial Western ones,[16] likely having programmed them to flag and report 'politically

[11] In the OpenAI terms of use, it is made clear that the responsibility for the use of the chatbot and the results obtained from it lie with its users. OpenAI Terms of Use are available at http://tinyurl.com/mrxezyp7 (last visited 10 February 2024).

[12] M. Heikkilä, 'Here's how Microsoft could use ChatGPT' *MIT Tech Review*, 17 January 2023: 'Models like ChatGPT have a notorious tendency to spew biased, harmful, and factually incorrect content. They are great at generating slick language that reads as if a human wrote it. But they have no real understanding of what they are generating, and they state both facts and falsehoods with the same high level of confidence'.

[13] For an overview of literature, notions and theories on fake news see E. Aïmeur et al, 'Fake news, disinformation and misinformation in social media: a review' 13(1):30 *Social Network Analysis and Mining*, 1-36 (2023), available at urly.it/3zx0w (last visited 10 February 2024). For a deep analysis of the issues raised see L. G. Jacobs, 'Freedom of Speech and Regulation of Fake News' *The American Journal of Comparative Law*, 70, 1278-1311 (2022). A map of the different actions against misinformation may be found in the work of D. Funke and D. Famini, 'A guide to anti-misinformation actions around the world' Poynter.org, first edition 2018, last update 2019: http://tinyurl.com/bdfxwvpr (last visited 10 February 2024). In 2022 the UN Human Rights Council adopted a plan of action against fake news and the European Commission presented a revised version of the Code of Practice on Disinformation which, in its first version dated 2018, was the first-of-its kind tool with which major industry players agreed on self-regulatory standards to fight disinformation. The 2022 Code is part of a broader regulatory framework, consisting of legislation on Transparency and Targeting of Political Advertising and the Digital Services Act.

[14] NikkeiAsia in February 2023 stated that 'Tencent Holdings and Ant Group, the fintech affiliate of Alibaba Group Holding, have been instructed not to offer access to ChatGPT services on their platforms, either directly or via third parties': the article is available at urly.it/3zx0y (last visited 10 February 2024).

[15] It should be noted that the UK has also banned the use of ChatGPT for work purposes by civil servants on public order grounds: S. Trendall, 'Government guidance bans civil servants from using ChatGPT to write policy papers', 3 July 2023, available at urly.it/3zx0- (last visited 10 February 2024).

[16] After Ernie Bot (created by Baidu as the Chinese answer to ChatGPT and made publicly available in August 2023) and SenseChat, created by SenseTime, the creation of AndesGPT by Oppo is the latest news. Its integration in the new version of the ColorOS 14 operating system, currently

sensitive' or 'potentially dangerous' questions and to act as a disseminator of state 'truths'.[17]

III. The Generative AI

We have so far looked at vulnerability profiles posed by this type of AI and associated with algorithmic pre-training and the consequences of deep learning. However, chatbots are neural systems that continue their training autonomously, either by scraping the Internet directly or by prompting a response. Consequently, the algorithm constantly acquires and uses new data to generate information and improve its performance. The issues raised by this constant training, learning and generating mechanism are multifold, but some are of particular interest for the purposes of our analysis, namely: privacy violations that take place with the processing – for training purposes – of personal data entered into the web by third parties[18] or fed into the chatbot by users themselves with their prompts (an issue addressed in the second part of this paper and directly connected with the provisions of the Italian Data Protection Authority); issues of output authenticity and authorship; process opacity; and reflexive conditioning.

Starting with the last of these issues, the information produced by chatbots is generated with the intention of being used for other inquiries as well as by users themselves for their own purposes. A high percentage of the information and text thus produced goes towards enriching the web and generating public resonance.[19] The reflexive conditioning thus created is twofold: on the one hand data produced by bots – and which feed other bots – trigger a slow but relentless substitution mechanism whereby information and text of human origin is replaced with information and text of robotic or mixed origin,[20] and on the other hand, these hits consolidate the processed and disseminated content regardless of its quality, correctness or truthfulness. In addition to a concrete risk of spreading dis- and misinformation, there is a serious and real danger of perpetuating and reinforcing

only active in China, allows better dialogue with users via the Breeno virtual assistant: urly.it/3zx0_.

[17] Through its party newspaper, the China Daily, the Chinese government published a video accusing the United States of instrumentalizing ChatGPT for propaganda purposes by spreading disinformation. The video can be seen at http://tinyurl.com/hxppa4nz (last visited 10 February 2024).

[18] Privacy issues arise with respect to both Large Language Models and Image Diffusion Models, as shown by studies carried out by a group of Google researchers: N. Carlini et al, 'Extracting training data from large language models' *USENIX Security Symposium*, 2021, available at urly.it/3zx11 (last visited 10 February 2024), and N. Carlini et al, 'Extracting Training Data from Diffusion Models' *USENIX Security Symposium*, 2023, available at urly.it/3zx12 (last visited 10 February 2024).

[19] S. Fürst, 'Öffentlichkeitsresonanz als Nachrichtenfaktor-Zum Wandel der Nachrichtenselektion' 37(2) *MedienJournal*, 4-15 (2017).

[20] A very interesting study with questions raised by M. Lana, 'L'agency dei sistemi di intelligenza artificiale. Un punto di vista bibliografico' *DigitCult - Scientific Journal on Digital Cultures*, 1, 67-78 (2022), on the book Lithium-ion batteries, a machine-generated summary of current research, by Beta Writer.

prejudice and discrimination, and of generating mass manipulation.[21] This is because AI works in a binary fashion (action/reaction, input/output), arranging data according to a probabilistic assessment, with a tendency to simplify (and thus to reduce), the number of variables. Evidence of this is found in the widespread political polarization and the alteration of democratic dynamics that the public opinion in Western countries has slowly undergone since the launch of social media.[22] Therefore, beyond disinformation, a significant aspect of human vulnerability posed by AI interaction is the progressive reduction of pluralism and dissent.

It would be easier to curtail these threats if the process that takes AI from certain inputs to certain outputs were transparent. Originally, the lack of transparency was a direct consequence of the IP protection of the algorithmic codes used by different platforms to process user data. Nowadays, however, as a result of self-training and deep learning mechanisms, programmers are able to explain the functioning of AI agents but are not able to reconstruct, *a posteriori*, the logical-mathematical processes that take place in the black boxes of AI functioning nor are they able to predict the outputs produced by them.[23] The opacity of the processes and the unpredictability of the results increase human vulnerability in direct proportion to the increase in AI agency. This vulnerability persists even after the actions taken by OpenAI on ChatGPT to comply with the request of the Italian Data Protection Authority to set up tools capable of amending incorrect personal data or of deleting them at the request of the concerned party. ChatGPT training is automatic and continuous, so the exercise of the right to withdraw consent for personal data processing by the user cannot operate retroactively, and data ingested by the bot remain in its black box.[24]

In addition to being autonomous in their continuous learning, chatbots are, by definition, 'generative', which is to say they process the enormous amount of data they feed on to answer questions or carry out text composition tasks. In answering questions and composing text the whole original connection with the data (the processing of which has generated such answers and text) is lost. Chatbots do not quote their sources and the generation process remains obscure. Just as it is not possible to correct chatbots – other than through new training – it is likewise not possible to credit the authors of the data from which answers

[21] E. Falletti, *Algorithmic Discrimination: A Comparative Perspective*, (Torino: Giappichelli, 2022) passim; N. Gross, 'What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI' 435 (12) *Social Sciences*, 1-15 (2023).

[22] A. Tedeschi Toschi and G. Berni Ferretti, 'Il contrasto legislativo ai socialbot. Alcuni spunti per una riforma in Italia' *Rivista italiana di Informatica e Diritto*, 155-175 (2023).

[23] J. Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' *Big Data & Society*, 1-12 (2016).

[24] For more details on the mechanism introduced by OpenAI following the processing measures of the Italian Data Protection Authority, see L. Megale, 'Il Garante della privacy contro ChatGPT: quale ruolo per le autorità pubbliche nel bilanciare sostegno all'innovazione e tutela dei diritti?' *Giornale di diritto amministrativo*, 3, 409-410 (2023).

and text have been generated. This poses the problem of authorship, since each output (eg a scientific article, a poem, a play or film script) may well be the result of serial infringement of publishers 'copyright[25] and/or authors' intellectual property.[26] Moreover, the question of authorship also arises in relation to intellectual property and the right to exploit the text thus generated. The users – whose chatbot prompts have caused the generative act – may now use the result of this act at will, even attributing it to themselves.[27] This, however, raises the further question of authenticity.[28] At the time of writing, there is no technology able to ascertain what percentage of 'artificial assistance' is present in any written text. The issue of so-called 'paper mills' is particularly emblematic in terms of production in the scientific field.[29] The solution proposed in various fora is that of a watermark (ie an indelible imprint on the electronic format of a text) capable of identifying and separating artificial products from human ones. That said, bypassing any prohibition of use would be easy, as demonstrated by the interdiction order case involving the Italian Data Protection Authority.[30]

[25] The New York Times has questioned the future of publishing and journalism on a number of occasions in recent years (S. Podolny, 'If an Algorithm Wrote This, How Would You Even Know?' 7 March 2015; J. Peiser, 'The Rise of the Robot Reporter', 5 February 2019) and in August news came out on NPR that NYT's lawyers were exploring whether to sue OpenAI to protect the intellectual property rights associated with its reporting: urly.it/3zx19.

[26] In July 2023, Sarah Silverman sued artificial intelligence producers OpenAI and Meta Platforms for not having her permission to use her copyrighted works. Other authors joined Silverman in these suits to seek a class-action status. It is doubtful, however, whether the lawsuit will succeed, especially in consideration of the landmark case involving Google Books in 2016. In this case, the Second Circuit Court of Appeals in the United States ruled that Google Books practice of summarising texts did not violate copyright law and that Google's use of copyright protected works is a case of non-infringing fair use: *Authors Guild of America* v *Google* 721 F.3d 132 (2nd Cir. 2015). The United States Supreme Court subsequently rejected the Authors Guild's petition for appeal from the US Court of Appeals decision. Among the most recent comments on the topic of transformative and non-transformative use of copyright protected works is the article by C. Sandalow, 'I Did You a Favor By Taking Your Work: Reconsidering the Harm-Based Approach To the Fourth Fair Use Factor' 46 *Columbia Journal of Law & Arts*, 457-485 (2023).

[27] The OpenAI's Terms of Use of ChatGPT states that you (the user) '(a) retain your ownership rights in Input and (b) own the Output. We hereby assign to you all our right, title, and interest, if any, in and to Output'.

[28] For an examination of the issues related to the concept of author in the digital world, see R. Morriello, 'OpenAI e ChatGPT: funzionalità, evoluzione e questioni aperte' [S.l.] 8 (1) *DigitCult - Scientific Journal on Digital Cultures*, 59-76 (2023).

[29] N. Lucchi, 'ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems' *European Journal of Risk Regulation*, 1-23 (2023); R. Morriello, *Dalla pirateria dei libri all'editoria predatoria. Un percorso tra storia della stampa ed etica della comunicazione scientifica* (Milano: Ledizioni, 2022) 116-121, therein extensive literature. The alarm mainly concerns biomedical science: most recently, cf B.A. Sabel et al, 'Fake Publications in Biomedical Science: Red-flagging Method Indicates Mass Production' preprint 18 October 2023 doi: http://tinyurl.com/46ua3h2b.

[30] Instructions of how to circumvent the Italian Data Protection Authority's ban by means of VPN were easily found on the web, as were instructions on how to circumvent ChatGPT's Ethics Safeguards: Jon Christian, Amazing 'Jailbreak' Bypasses ChatGPT's Ethics Safeguards, in Futurism, 4 February 2023, available at http://tinyurl.com/3tpmmj2x (last visited 10 February 2024).

IV. The Communicative AI

Apart from the critical issues that have emerged in our observation of the technical functioning of chatbots, another major concern is the polished ability of these robots to mimic human conversation. Indeed, in their exchanges, chatbots are able to affect empathy and impressive erudition, mixed with arrogant assertiveness and variability of tone (friendly, excited, stable, serious) and style (professional, informative, educational, storytelling, benefit-focused or solution-oriented). The constant performance improvement occurs through complex attention mechanisms that enable the bots to focus on specific parts of the input text to generate more relevant and accurate outputs with respect to context, recipients' personalities and their wishes. Furthermore, the ability to store input information from the same user in memory modules makes questioning and answering exchanges more coherent and similar to those between humans. But whilst chatbot answers may seem 'sensible', they actually make 'no sense' to the machines whatsoever. And this is where the most insidious threat lies, because chatbots are modelled to imitate human conversation, thus making it hard to recognize responses as 'artificial'. It is for this very reason the chatbots must, by default, warn users of their nature. In addition to the problems of authorship, authenticity and reliability that have been addressed, the risk in the medium to long terms is the unpredictable consequences of human-machine interaction itself, at least in two respects: epistemological and sociological.

Let us start with the epistemological profile. Whilst the ability of chatbots to assimilate syntactic rules endows their responses with impressive linguistic consistency, the modest writing ability of the average user puts the latter in a position of inferiority *vis-à-vis* the machine. This condition of perceived or actual inferiority constitutes a prerequisite for vulnerability. Beyond any form of 'amusement',[31] 'intellectual' challenge[32] or professional use of chatbots (by experts capable of appraising the reliability of responses), most users who question a chatbot are technically inexperienced and not competent in the subject matter. For such users chatbots are neither entertainment nor work, but tools for understanding reality. So, in addition to the danger of spreading misinformation that has been addressed, the inferior subject's proneness to rely on those who are perceived to be more experienced, capable and educated alters the normal course of the interaction. Whilst this is common in interactions between humans – so much so that experts take responsibility for what they say – in interactions with chatbots users are warned of the nature and limits of the bot and, in accordance with the terms of use, are held solely responsible for their prompts, the ChatGPT

[31] H. Holden Thorp, 'ChatGPT is fun, but not an author' 379:6630 *Science*, 313 (2023).

[32] See the experiment on mathematical, semantic and ethical questions and the conclusions on the 'significant consequences of the industrialisation of automatic and cheap production of good, semantic artefacts' by L. Floridi and M. Chiriatti, 'GPT-3: Its Nature, Scope, Limits, and Consequences' *Minds and Machines*, 681-694 (2020).

responses that ensue, and the use that is made of such responses.[33] The manner in which warnings are given and terms of use are accepted does not prevent users from consciously or unconsciously perceiving epistemological value in the chatbot's responses. The syntactic accuracy of bot answers and the adjustment of tone and style to match those of the questions are not, of course, the result any consciousness and sensitivity of the bot, but interlocutors are, nonetheless, led to perceive the bot as being endowed with both. Scientists' warnings[34] about NLG's lack of empathy, semantic cognition or attribution of meaning have been to no avail. The mirror with which AI reflects their image back onto humans is both deceiving and beguiling.[35] Thus, the danger is to witness a human preference, at the micro level, for perpetually available, educated, accommodating and benevolent artificial conversation.[36] At the macro level, the threat lies in the deliberate or accidental manipulation of reality[37] and human knowledge as generated by 'meaningless' AI narratives.[38]

The analysis of the communicative potential of chatbots under an epistemological profile allows us to imagine their impact on the dynamics of social interaction under a sociological profile. Indeed, in its digital interactions, human vulnerability undoubtedly emerges as a characteristic of the individual. Since the individual is 'interdependent people in the singular', vulnerability also pertains to the whole of society, as it is composed of 'interdependent people in the plural'.[39] The ability

[33] In regard to content, it is said that you (the user) 'may provide input to the Services (Input) and receive output from the Services based on the Input (Output). Input and Output are collectively Content. You are responsible for Content, including ensuring that it does not violate any applicable law or these Terms. You represent and warrant that you have all rights, licenses, and permissions needed to provide Input to our Services'. See A. Malaschini, 'ChatGPT e simili: questioni giuridiche ed implicazioni sociali' *Consulta online*, II, 583-606, 597 (2023).

[34] D. McQuillan, 'Manifesto on algorithmic humanitarianism' *openDemocracy*, 4 April 2018; E.M. Bender and T. Gebru, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?' *Conference on Fairness, Accountability, and Transparency* (FAccT 21), 610-623 (2021); K. Arkoudas, 'ChatGPT is no stochastic parrot. But it also claims that 1 is greater than 1' 36 *Philosophy & Technology*, 54 (2023).

[35] Well described by R.W. Gehl and M. Bakardjieva, *Socialbots and Their Friends. Digital Media and the Automation of sociality* (New York: Routledge, 2016), 2: Social bots are 'intended to present a self, to pose as an alter-ego, as a subject with personal biography, stock of knowledge, emotions and body, as a social counterpart, as someone like me, the user, with whom I could build a social relationship'.

[36] R.W. Gehl and M. Bakardjieva, n 35 above, 2.

[37] Actually, transformations arising from deep mediatisation ensue as a sort of re-figuration, namely a fundamental, structural shift of human relationships and practices: A. Hepp, *Deep Mediatisation* (New York: Rutledge, 2020), 106-112.

[38] L. Floridi, 'AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models' 36:15 *Philosophy & Technology*, 1-7 (2023); J.M. Bishop, 'Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It' 11 *Frontiers in Psychology*, 1-18 (2021); S. Amato, 'Tra silicio e carbonio: le macchine saranno sempre stupide?' *BioLaw Journal - Rivista di BioDiritto*, I, 295-302 (2023).

[39] N. Elias, The Civilizing Process. The Development of Manners (New York: Urizen Books, 1978), 125.

of chatbots to imitate human communication directly threatens the individuals with whom they interact and indirectly threatens the entire society.[40] The communication is not authentic but appears so. The answers to questions, which appear original and full of new content, are, in fact, merely syntactically ordered and stylistically elegant reformulations of existing content. The process of retrieving archived content takes place automatically according to the law of the most probable.[41] Chatbots do nothing more than perpetuate the most quantitively prevalent data (and not the most qualitatively or ethically superior data)42 through an iteration of logical models, with no critical capability or semantic awareness whatsoever. If communication is a form of symbolic construction of reality, then the massive proliferation of communicative AI chatbots cannot but affect the construction mechanism of the representation that individuals have of themselves, of the society they belong to, and of the environment they live in.[43] It is impossible to predict whether the entry of AI as a new individual and social interlocutor will disrupt or be beneficial to psychological and sociological dynamics. However, given this uncertainty, it would be advisable not to take risks, and to avoid exposing users to such risks it is necessary to regulate the phenomenon and exercise control over it vis-à-vis constitutional principles, fundamental human rights, and general public interest. But what rules? What control?

V. Agile Regulation and Prior Public Scrutiny

The free ChatGPT application launch in November 2022 may be likened to a mass experiment. The purpose of the launch was not ‘to ensure that artificial general intelligence benefits all of humanity’ (as claimed on the OpenAI website), but rather to fine-tune the product by exploiting the prompt training provided by users; a freemium-type marketing strategy[44] for the purpose of subsequent

[40] On October 2022, after a year-long process led by the US Office of Science and Technology Policy, the White House released the Blueprint for an AI Bill of Rights to inform policy decisions. The concept of community is integral to the scope of this Blueprint and affirms that, while AI and other data-driven automated systems most directly collect data, make inferences, and may cause harm to individuals, the overall magnitude of their impacts is most readily visible at the community level: urly.it/3zx1h.

[41] M. Bertolaso and A. Marcos, Umanesimo tecnologico. Una riflessione filosofica sull’intelligenza artificiale (Roma: Carocci editore, 2023), 49-52.

[42] L. Floridi, Etica dell’intelligenza artificiale. Sviluppi, opportunità, sfide (Milan: Raffaello Cortina Editore, 2022), passim.

[43] A. Hepp et al, ‘ChatGPT, LaMDA, and the Hype Around Communicative AI: The Automation of Communication as a Field of Research in Media and Communication Studies’ 6 *Human-Machine Communication*, 41-53 (2023).

[44] Freemium is a business model that consists in offering a basic version of a product free-of-charge. The marketing strategy aims to attract a large volume of potential customers (customer acquisition) and at the same time to test the product in order to simultaneously or later offer an updated or improved version of it for a fee. A business model already used since the 1980s, it was christened by J. Lukin in 2006 when commenting on an article by Fred Wilson entitled ‘My Favourite

commercial exploitation. Fine-tuning of the product was left to the massive training by users, who were warned (in the terms of use), that their prompts would go towards improving the service[45] and training the algorithms.[46] The analysis of how the algorithm works has revealed several critical concerns. Moreover, in recent months, operators, academics and legislators have all been compelled to consider the instruments already in place and the ones still needed to prevent risks and curb threats. This consideration is taking place on two levels: a specific one, with reference to chatbots, and a general one, with reference to any future AI product that might be launched without prior public scrutiny.

With regard to ChatGPT, we shall shortly look at the *a posteriori* actions taken by the Italian Data Protection Authority to demonstrate the uselessness of the ban instrument and the inadequacy of the GDPR (General Data Protection Regulation) to deal with threats posed by this new technology. In regard to both the enforcement action taken by the Italian Data Protection Authority and the inadequacy of GDPR, on 13 April 2023 European Data Protection Board (EDPB) members decided to launch a dedicated task force to foster cooperation on the matter. In general, the Fall of 2023 witnessed multiple political reactions to ChatGPT: the UK White Paper on AI (29 March 2023),[47] the Chinese Cyberspace Administration Draft Measures for managing generative AI (11 April 2023) and the EU Parliament amendments to the Commission's AI Act Proposal (11 May

Business Model'. The Freemium model was then studied and structured around four models by Chris Anderson (*Free: The Future of a Radical Price*, New York: Random House Business, 2009) and dissected by Eric Seufert's lucid analysis of its application to software products (*Freemium Economics: Leveraging Analytics and User Segmentation to Drive Revenue*, Waltam MA: Morgan Kaufmann, 2014). Seufert provides insight into how freemium products generate revenue, keep users engaged, and grow. Of particular relevance to the identification of vulnerability profiles following the free launch of ChatGPT are the reflections on some of the most important concepts in freemium design, namely lifetime customer value (Chapter 5) and virality (Chapter 7).

For an analysis of how the freemium model operates from an application point of view and with respect to empirical findings, see the Spotify case examined by C. Becagli et al, 'Il modello di business "Freemium" nel settore musicale e i fattori incentivanti del passaggio da utente free a premium: Evidenze empiriche dal caso Spotify' in F. Culasso and M. Pizzo eds, *Identità, innovazione e impatto dell'aziendalismo italiano. Dentro l'economia digitale* (Torino: Collane UniTo, 2019), 526-527, available at http://tinyurl.com/3r9e726j (last visited 10 February 2024).

[45] The ChatGPT's Terms of Use make explicit that OpenAI may use Content to provide, maintain, develop, and improve our Services, comply with applicable law, enforce our terms and policies, and keep our Services safe.

[46] Following the Italian Data Protection Authority's intervention, an opt-out option has been included in the terms of use, which, however, is not quick to implement, as it requires a form to be filled in, and could affect the efficiency of the service: 'If you do not want us to use your Content to train our models, you can opt out by following the instructions in this Help Centre article. Please note that in some cases this may limit the ability of our Services to better address your specific use case'.

[47] The UK Government launched an AI White Paper to guide the use of artificial intelligence and to drive responsible innovation. AI use will be guided by five principles: safety, transparency, fairness, accountability and contestability: Office for Artificial Intelligence, Department for Science, Innovation and Technology, 'Policy paper "A pro-innovation approach to AI regulation" ', Command Paper no 815, 2023 (updated 3 August 2023).

and 14 June 2023).[48]

The debate on ChatGPT so ignited the German Bundestag on 29 March 2023[49] that it led to the publication, in November, of an AI Action Plan by the Federal Ministry of Education and Research.[50] On 21 March 2023, the French Assemblée Nationale debated and rejected an amendment drafted by ChatGPT[51] and on 19 September Premier Élisabeth Borne set up a special *Comité de l'intelligence artificielle generative*.[52] Concerns about the spread of chatbots and the possible threats associated with them have also emerged at the international level, triggering a number of initiatives already underway and several others that are in the pipeline. UNESCO has published several papers dealing with ChatGPT, ranging from a Guidance for generative AI in education and research to a policy paper[53] containing analyses of new AI technologies through the lens of UNESCO's Recommendation on AI Ethics. At the annual Group of Seven (G7) Summit hosted by Japan and held in May 2023, the leaders of the G7 countries expressed concern over the

[48] On 11 May 2023 the Internal Market Committee and the Civil Liberties Committee adopted a draft negotiating mandate on the AI Act proposal (with many amendments thereto), stating that generative foundation models such as GPT would have to comply with additional transparency requirements. Such additional requirements include the requirement of disclosing that content was generated by AI, designing the model to prevent it from generating illegal content, and publishing summaries of copyrighted data used for training. The European Parliament adopted the amendments on 14 June, introducing Recital 60 g (Generative foundation models should ensure transparency about the fact that the content is generated by an AI system, not by humans. These specific requirements and obligations do not amount to considering foundation models as high-risk AI systems, but should guarantee that the objectives of this Regulation to ensure a high level of protection of fundamental rights, health and safety, environment, democracy and rule of law are achieved. Pre-trained models developed for a narrower, less general, more limited set of applications that cannot be adapted for a wide range of tasks such as simple multi-purpose AI systems should not be considered foundation models for the purposes of this Regulation, because of their greater interpretability which makes their behaviour less unpredictable), 60 h (As foundation models are a new and fast-evolving development in the field of artificial intelligence, it is appropriate for the Commission and the AI Office to monitor and periodically asses the legislative and governance framework of such models and in particular of generative AI systems based on such models, which raise significant questions related to the generation of content in breach of Union law, copyright rules, and potential misuse), Art 28b (4. Providers of foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video ('generative AI') and providers who specialise a foundation model into a generative AI system, shall in addition a) comply with the transparency obligations outlined in Art 52 (1), b) train, and where applicable, design and develop the foundation model in such a way as to ensure adequate safeguards against the generation of content in breach of Union law in line with the generally-acknowledged state of the art, and without prejudice to fundamental rights, including the freedom of expression, c) without prejudice to Union or national or Union legislation on copyright, document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law).

[49] On 29 March 2023, the Bundestag's Digital Committee debated the status of negotiations on the legal regulation of generative artificial intelligence (AI) at EU level: urly.it/3zx2b.

[50] Available for download at urly.it/3zx2m.

[51] The amendment is available for download at urly.it/3zx2p.

[52] For more information: urly.it/3zx2q.

[53] The policy paper deals with 'Foundation models such as ChatGPT through the prism of the UNESCO Recommendation on the Ethics of Artificial Intelligence': urly.it/3zx2t.

disruptive potential of rapidly expanding generative AI and agreed on the need for governance to ensure a human-centric and trustworthy development. The agreement triggered the so called 'Hiroshima Process' which – as recently as October 2023 – gave birth to the G7 Leaders' Statement, the International Guiding Principles, and the International Code of Conduct for Organizations Developing Advanced AI Systems. The concerns of world leaders have significantly escalated, leading – most recently, in November 2023 – to the twenty-eight signatures on the Bletchley Declaration[54] that closed the AI Safety Summit hosted in the UK by Premier Rishi Sunak.

An analysis of government strategies on the topic of artificial intelligence reveals a number of different positions on the matter, some preferring mild guidelines and others strict regulatory laws.[55] That said, recent debates and actions at national, regional and global levels all seem to converge towards the use of new regulatory forms.[56] The difficulties experienced with introducing rules into a hard law proposal such as the AI Act (now being debated by the European trilogue) amidst the hype surrounding ChatGPT and, conversely, the speed with which the Bletchley Declaration was signed, have taught us that the best strategy is that of a legal process with variable geometry and force; one that falls between intersectoral guidelines and hard laws, between sectoral codes of conduct and public authority controls, between business lobbying and democratic empowerment. The OECD has put together a regulatory policy with agile and innovative approaches, which describes tools to address digital era challenges such as regulatory sandboxes, behavioral insights, and risk-based and outcome-based regulations.[57]

There seem to be two directions along which political action is moving to try to regain control over fast-emerging technology. These are, on the one hand, to require that AI producers establish, implement, document and maintain a risk management system with third-party verification and comply with duties of

[54] Among the 28 signatories of the Bletchley Declaration are the UK, the US, the EU, China and Australia. Surprisingly, it was signed on the first day of the summit, even though the content is very daring and the goals are very challenging: from the assertion that 'there is potential for serious, even catastrophic, harm' to recognition that 'the protection of human rights, transparency and explainability, fairness, accountability, regulation, safety, appropriate human oversight, ethics, bias mitigation, privacy and data protection all need to be addressed'.

[55] In France, *La stratégie nationale pour l'intelligence artificielle* was launched in 2018 and is now in its second phase (urly.it/3zx2-). In Germany, the Bundesregierung adopted its *Strategie Künstliche Intelligenz* in 2018, now in its new version 2020: urly.it/3zx32. The Italian *Programma Strategico per l'Intelligenza Artificiale* was approved by the Council of Ministers on 24 November 2021 and has a scope for the two-year period 2022-2024: urly.it/3zx36.

US, UK, Australia and Japan tend to prefer overseeing AI with mild guidelines, while the EU and its member states have opted for strict regulatory laws.

[56] O. Pollicino, 'I codici di condotta tra self-regulation and hard law: esiste davvero una terza via per la regolazione digitale? Il caso della strategia europea contro la disinformazione online' *Rivista trimestrale di diritto pubblico*, 4, 1051-1068 (2022).

[57] See OECD, *Regulatory Policy Outlook*, 2021, in particular 'Regulatory policy 2.0, available at urly.it/3zx38 (last visited 10 February 2024).

transparency, explanation and provision of information to users,[58] and on the other hand to introduce procedures and authorities to control new technologies before their mass distribution. The latter solution has been adopted by the AI regulatory sandboxes of the European AI Act Proposal, the UK White Paper, and President Biden's Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.[59]

As for obligations of transparency, risk assessment and explainability,[60] the problem lies in the difficulty of predicting risks. This is because the more the AI is advanced, the more unpredictable its logical-mathematical processes, decisions and outputs are. To require that only inherently controllable algorithmic models (ie ones that are predictable *ex-ante* or re-constructible *ex post*)[61] be used is unreasonable. Trying to get inside the prompt-response process of ChatGPT is tantamount to trying to access a human mind to understand its decision-making processes. There is a specific branch of science on explainable AI. Initially, it focused on developing procedures to explain the operations of self-learning algorithms,[62] but the processes of complex algorithmic models are incomprehensible and will remain so. Research has therefore shifted to a different approach: one that specifically exploits the interactive ability of bots. In short, it is a matter of inducing the AI, through prompt/response interlocution, to give, itself, an explanation of its logic-deductive and logic-generative processes and to provide a record of the data used for this in a human-friendly post-hoc fashion.[63] This would allow for *a posteriori* control over the AI and a way to correct biases and eliminate discrimination. The solution implies deep and continuous human-bot interaction, with all the unknowns that go with the radical paradigm shift induced by this new communicative AI and unprecedented in human history.[64] In the nineties, Lawrence B. Solum concluded his essay on legal personhood for artificial

[58] See Art 9 and Art 13 AI Act Proposal of the EU Commission and No 47 UK White Paper.

[59] The public body involved for the development of standards and the provision of testing environment is the National Institute of Standards and Technology (NIST): urly.it/3zx3b.

[60] In the UK White Paper 'explainability refers to the extent to which it is possible for relevant parties to access, interpret and understand the decision-making processes of an AI system'. Prior to this, the GDPR's 'right to explanation' has been the tool for promoting fairness, accountability, and transparency and for granting investigatory powers to data authorities: for literature and comments see B. Casey et al, 'Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise' 34 *Berkley Technology Law Journal*, 143-188 (2019).

[61] S. Robbins, 'A Misdirected Principle with a Catch: Explicability for AI' 29 *Minds and Machines*, 495-514 (2019); C. Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' *Nature Machine Intelligence*, 1, 206-215 (2019).

[62] L. H. Gilpin et al, 'Explaining Explanation: An Approach to Evaluating Interpretability of Machine Learning' *ArXiv* (Cornell University), arXiv:1806.00069 (2018); T. Miller, 'Explanation in Artificial Intelligence: Insights from the Social Sciences' 267 *Artificial Intelligence*, 1-38 (2019).

[63] E. Esposito, 'Dall'Intelligenza artificiale alla comunicazione artificiale' 392 *Aut Aut*, 20-35 (2021).

[64] A catalogue of issues concerning the relationship between the concepts of personhood and humanity may be found in the interesting essay by L. Solum, 'Legal Personhood for Artificial Intelligences' 70 *North Carolina Law Review*, 1231-1287 (1992), in particular 1284-1285.

intelligences by stating that,

> ‘an answer to the question whether artificial intelligences should be granted some form of legal personhood cannot be given until our form of life gives the question urgency’.[65]

The time has come!

VI. Data Protection Issues. Territorial Scope and Application of the GDPR

The most urgent issues, apart from those already highlighted in the first part of the paper, are probably those related to data protection, as many data protection authorities (DPAs) of the single Member States of the European Union have been concerned with these aspects.[66]

How do data protection rules apply to ChatGPT and, in general, to chatbots and other generative artificial intelligence models? The first question to be considered is which rules apply to ChatGPT, and specifically, whether the GDPR is applicable and in which cases.

As already mentioned, ChatGPT is an AI tool, developed by OpenAI, a US company, and accessible worldwide, via any electronic device. The OpenAI privacy policy available on the landing page, relating to registration for the ChatGPT service, also specifies that the service can be used by ‘international users’, informing readers that the users’ personal data will be transferred in the United States or where the servers of the company are located. So, if a user located in the European territory uses ChatGPT, their data are processed by the platform and transferred to servers located outside the European territory.

Art 3 of the GDPR regulates this case, holding the territorial scope of application of the Regulation based on two criteria. The first criterion, established in Art 3, para 1 of the GDPR, decrees that the Regulation applies to the processing of personal data carried out in the context of the activities of an establishment by the data controller or processor in the European Union (EU), regardless of whether the processing is carried out in the EU. This criterion is also called ‘establishment’ since what is important for the application of European legislation is that the carrying out of the activities takes place within the framework of a stable

[65] L. Solum, n 64 above, 1287.

[66] For further details see J. Meszaros et al, ‘ChatGPT: how many data protection principles do you comply with?’, available at SSRN: https://ssrn.com/abstract=4647569, 7-9. Investigations have been opened by the Spanish Data Protection Authority, on 13 April 2023, as well as several German regional authorities, by sending a questionnaire to OpenAI, asking if a data protection impact assessment was made by the company and requesting further information on data subjects’ rights. Also CNIL (the French DPA) launched an action plan, scheduling investigations on several generative AI providers.

organization, including through a branch or an affiliate, within the EU.[67]

Taking into account the circumstances that characterize the service provided by OpenAI, according to the establishment criterion, the regulation referred to in the GDPR would not be applicable. The data controller, identified as Open AI, has its main office in the United States and there are no other units, companies or entities within European territory, nor entities operating on behalf of Open AI, which may be considered 'data controllers'.

However, according to Art 3, para 2 of the GDPR, the Regulation provides a second criterion according to which, if neither the owner nor the data processor is established within EU territory, the GDPR still applies. In fact, one of the purposes of the Regulation is to protect natural persons living in the EU, when the processing of their data is carried out by owners and managers established outside its borders.

In this case, reference is made to the so-called 'targeting' criterion according to which the GDPR applies when the processing is carried out by subjects established outside the EU, but offer of goods and services is addressed to European citizens, regardless of the existence of a payment of a monetary amount or (in the case of monitoring) of the behavior of the interested parties.

It should be noted that Art 3, para 2 of the GDPR refers to the 'processing of personal data of data subjects who are in the Union'. Therefore, the application of this measure is not limited by citizenship, residence or other elements of the legal condition of the interested party whose personal data are being processed, but to all subjects who are located within the EU borders.[68]

However, it is not sufficient for an interested party to be within the EU for the GDPR to apply; it is also necessary that certain characteristics connected to the processing are met. Of the two cases listed above by Art 3, para 2 of the GDPR, the one relating to the 'offering of goods or services, irrespective of whether a payment of the data subject is required, to such data subjects in the Union' is the one that finds recognition in relation with the activity carried out by OpenAI. The latter offers chatbot service, ChatGPT, available upon registration of the user, who has the right – but not the obligation – to purchase a paid package to obtain greater functionality compared to the service standard. Furthermore, since this service is accessible online, it is intended to be provided without distinction to anyone in the EU who has an electronic device. This would support the application of the GDPR to the instant case, bearing in mind what is supported

[67] See Recital 22 GDPR. For the notion of 'establishment', see also A. Spangaro, 'L'ambito di applicazione materiale della disciplina del regolamento europeo 679/2016', in G. Finocchiaro ed, *La protezione dei dati personali in Italia. Regolamento UE n. 2016/679 e D. Lgs. 10 agosto 2018, n. 101* (Bologna-Roma: Zanichelli, 2019), 422, as well as the examples provided by the European Data Protection Board, Guidelines 3/2018 on the territorial scope of Art 3 GDPR, available at urly.it/3zx44 (last visited 10 February 2024).

[68] See the Guidelines 3/2018, completing what is already provided for in Recital 14 of the GDPR.

by Recital 23 of the GDPR.[69]

VII. The Investigation of the Italian Data Protection Authority

Data Protection Authorities (DPAs) of several member States have investigated the data processing made by OpenAI through ChatGPT. The first intervention was that of the Italian DPA when a provisional ban measure was adopted against ChatGPT in March 2023, requiring interruption of the service in Italian territory.[70] In particular, the Italian Authority issued an urgent and interim provision, based on three main issues:[71] the lack of a privacy policy; the presence of improper personal data in the texts provided by the chatbot;[72] and, finally, the lack of age verification of users.[73]

However, chronologically speaking, the measure related to ChatGPT is the second provision of the Italian DPA concerning an artificial intelligence system. In fact, on February 2, 2023, a decision held according to Art 58 para 2 (f) of the GDPR was issued ordering a temporary limitation

> 'on the processing of personal data relating to users in the Italian territory as performed by Luka Inc., the US-based developer and operator of Replika, in its capacity as controller of the processing of personal data that is carried out via the said app'.

Replika is a chatbot which creates virtual replicants with a text and voice interface, that can be configured by the user to be a friend, a mentor or a partner. The peculiarity of this technology is to be designed to replicate human behaviors, learning from interactions with humans to provide interlocutors with psychological support, empathetic engagement, and relief from anxiety. Thus, unlike ChatGPT, which provides answers to user questions, Replika supplies a virtual assistant programmed according to user-defined metrics, replicating a human being with

[69] See, for instance, CJEU, C-352/85, *Bond van Adverteerders and others* v *Dutch State*, 26th April 1988, § 16, available at urly.it/3zx47 (last visited 10 February 2024).

[70] Garante per la protezione dei dati personali, 30th March 2023, (doc. web n. 9870832), available here: urly.it/3zx49.

[71] L. Scudiero and S. Di Benedetto, 'Artificial Intelligence (AI) and Data Protection: Lessons Learned from the ChatGPT case (Italy)' *Practical Law Global*, 2 (15th December 2023).

[72] M. Santana Fernandes and J.R. Goldim, 'Artificial Intelligence and Decision Making in Health: Risks and Opportunities', in H. Sousa Antunes et al eds, *Multidisciplinary Perspectives on Artificial Intelligence and the Law* (New York: Springer, 2024), 192: 'These new systems not only label or classify pre-existing data, but generate new content, by aggregation and combination, from the available elements'.

[73] According to Art 8 GDPR, in the case of minors, the consent given is valid if the child is at least 16 years old. The GDPR allows member States to reduce the age limit to 13 years, even if this is not the case of the Italian regulation. See C. Yakışır, 'An Evaluation of the ChatGPT Decision, Which Italy Blocked Access on the Grounds of Violation of the GDPR' (19 April 2023). available at SSRN: https://ssrn.com/abstract=4423779.

whom to engage in human-like interactions.

During the inspection of the DPA it was proven that the privacy policy provided by Luka Inc to its users stated

> 'that personal data relating to below-13 children are not collected knowingly, whereas parents and legal guardians are encouraged to monitor use of the Internet by their children, comply with the privacy policy by instructing children to never provide personal data on the service without their authorisation, and contact the platform in case they have reason to believe that a below-13 child has provided personal data so that such data be removed from databases'.

Furthermore, in the app stores, Replika was listed as devoted to individuals aged above 17 and the terms and conditions released by the platform advised that below-13 children are banned from using the app and below-18 users must be authorized beforehand by their parents or legal guardians.

Despite these measures, the Italian DPA found that no technical limitations were put in place if the user declared to be a minor, considering that, during the subscription, the only information required were names, email addresses, and gender. Furthermore, as reported by several newspapers, the responses provided by Replika to minors or other vulnerable subjects were not suitable for the condition of such individuals, especially concerning sexually inappropriate content. Furthermore, the Italian DPA argued that the legal basis for the processing of personal data could not be found in the contractual performance, as no age verification was set up by the system and, according to Italian law, minors do not have the legal capacity to enter into contracts for the supply of implying the processing of a substantial amount of one's personal data.[74]

During its investigation into OpenAI, the Italian DPA ascertained that the company did not provide users with a privacy policy, thus not explaining the identity of the data subjects, from whom the data are collected, the data that were collected and processed, or the purposes of these processing activities. In fact, users were allowed to use the chatbot merely by logging in through a subscription to the platform or by using their e-mail credentials. Therefore, the platform was able to recognize the identity of the user, also in connection to the queries posed by the user itself to the platform.

Accordingly, the first critical issue was related to the collection of these data (those used for the subscription to the platform and the e-mail addresses used for accessing the services), which were processed without a prior privacy notice. Lacking this informative obligation, users were not in the condition to know, for instance, the identity and the contact details of the controller (or of its representative

[74] The Italian Data Protection Authority had already issued a temporary blocking measure against TikTok, as the social network platform did not provide proper methods for verifying the age of platform users, see Italian Data Protection Authority, 22 January 2021, doc. web [9524194].

in the EU territory, as in the instant case), the potential transfer of personal data outside the EU territory, the storage period of these data and the legitimation for this storage, the rights granted to the users and how to exercise these rights.

The privacy policy is the legal instrument through which information asymmetries are rebalanced between those who collect the data and the subjects to whom the data refers, who, as in the instance, may not know the uses that will be made of their data, nor the methods through which these data will be processed. One of the most controversial aspects of the use of artificial intelligence technologies is represented indeed by the opacity of the functioning of these tools, as well as the use that may be made of the data collected by the data controller after the first processing.[75]

After the first intervention of the DPA regulating its service, ChatGPT implemented the requested modifications to its landing page by including a privacy policy from the sign-up page before registration and allowing users, both located in Europe or in other territories, to opt-out from processing of their own personal data. Additionally, even if pointing out the impossibility of a full rectification of the inaccurate information provided by the answer of the machine, ChatGPT adopted a mechanism in order to authorize users to erase the incorrect information. However, at least at the time of writing, measures for age verification have not been implemented, nor has there been aninformation campaign aimed at empowering users with further information on the functioning of the artificial intelligence system.

VIII. Legal Basis and Right of Rectification

The second violation, according to the Italian DPA, concerned the absence of a legal basis for training the machine. Generative models are trained by scraping freely accessible contents on the internet, collecting and selecting among these contents, and allowing the machine to learn, similar to what a human brain would do.[76] As a part of this process, machine training can generate – as defined in technical jargon – machine hallucinations, where the machine, responding to a prompt from the user, can provide wrong answers, and may provide an incorrect

[75] It is also worth underlining that artificial intelligence systems learn, continuously and incessantly, like the human mind, from the information that is gradually provided. This aspect is clarified by OpenAI itself which, however, grants the user the possibility of setting the system in such a way as to disable learning at the time of its use; cf. urly.it/3zx4j.

[76] As already outlined, ChatGPT is trained using a method called unsupervised learning, where it learns from a diverse range of internet text. The training process involves exposing the model to a vast dataset containing parts of the internet, including websites, articles, and forums. The model learns to generate human-like text by predicting the next word in a sentence, given the context of the preceding words. This process is known as language modeling. According to the information provided by ChatGPT, Open AI uses a combination of human reviewers and automated filtering to curate and fine-tune the training data, aiming to create a model that is useful, unbiased, and safe.

reconstruction of the person's profile. These stem in part from the fact that ChatGPT is trained on information that is only periodically updated and all the information or contents created after this date are unknown to the machine: it means that ChatGPT may answer exclusively on facts that happened before 2022 and that the answers on specific persons (mainly public figures) could be imprecise, not taking into account events that happened in the last two years.

From a legal and data protection perspective, this can constitute a problem, since it could affect personal identity, providing users with partial or incorrect reconstructions of a specific person or context. For example, if we ask ChatGPT who the members of the current body of the Italian Personal Data Protection Authority are, the answer is correct, but the machine inserts, in addition to the existing ones, a fifth member, who in reality does not exist. Similarly, if ChatGPT is questioned about the identity of the members of the Authority, it states, for example, that Agostino Ghiglia, one of the four, is a former lawyer, when before being appointed to the DPA, he practiced as a journalist and was a politician, even though he also holds a law degree.

In addition to issues related to personal identity, the fact that AI machines are able to create texts that appear precise but in fact lack grounding in the real world could encourage misinformation. Frequently, users may not understand the difference between a search engine, which searches the information on the web and refers to a source of information, and a natural language processing tool (chatbot) like ChatGPT, which processes information scraped on the web, with the risk, already highlighted, of providing inaccurate data. This is the reason why, after the investigation of the DPA, ChatGPT (in its basic version) provides a warning to users, explaining that its training stops at 2022 and inviting them to consult other sources of information.

In addition to the issues described above, in the current operation of ChatGPT there could be a potential infringement of the principle of minimization, referred to in Art 5, para 1 (c) GDPR, which holds that it is necessary to process the least amount of personal data possible and these data must be 'adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed'.[77] However, compliance with the minimization principle could lead to a paradox. The possibility that ChatGPT returns complete results on an individual depends on the data used during the machine-learning process: the more data, the greater the precision in the biographical reconstruction of a person.

Furthermore, different issues arise if considered from different perspectives, and notably those of the users (data subjects) and of providers of the AI systems.

The first case perhaps is a non-issue, in the sense that, from a legal perspective,

[77] As explained by the UK Information Commissioner's Office, personal data processing must be: 'adequate – sufficient to properly fulfil your stated purpose; relevant – has a rational link to that purpose; and limited to what is necessary – you do not hold more than you need for that purpose'. On this principle see C. de Terwagne, in C. Kuner et al eds, *The EU General Data Protection Regulation (GDPR), A Commentary* (Oxford: Oxford University Press, 2020), 317.

the problem of the correct reconstruction of the identity of an individual may find a solution in the information requirements provided by the machine to the user. In other words, the DPA's provision does not require the result of the query to be exact, but that, on the one hand, the user is warned (as ChatGPT now does) that the information provided may not be complete, accurate, or, up to date; and that on the other hand, the data subject has the right to request the rectification of the personal information which is incorrect, pursuant to Art 16 GDPR.

Today, ChatGPT, when questioned about a person's identity, simply replies that their information is updated as of January 2022, inviting the user to check other and more recent sources of information. Probably, it would be appropriate for the system to add that some of the information provided may not be correct, in order to warn the user of the unreliability of the contents provided.

The second requirement, that data subjects be given the opportunity to request rectification of incorrect information, could mean that the system manager could be overwhelmed by millions of requests and be forced to correct the outputs of the machine. Rectification is a first-generation data protection right, included in Directive 46/95/EC, which allows data subjects to obtain 'without undue delay the rectification of inaccurate personal data concerning him or her'.[78] The boundaries of the right to rectification and of the right of erasure, contained in Art 17 GDPR, are unclear, in the sense that in the latter case the data controller is expected not to modify the information, but to delete it, while, in the case of the rectification, the information should be modified or updated following the request of the data subject.[79]

In the case of ChatGPT which is the right that the data subject may exercise? May I have the right of not being mentioned by an AI machine in its answers or may I only ask for the rectification of the incorrect information? If limited to ChatGPT, the answer is partially contained in the system itself, in the sense that the chatbot does not provide answers related to non-public figures.[80] In contrast, in cases of public figures, as already mentioned, the sentence starts with a warning 'As of my last knowledge update in January 2022'.

The exercise of the right to rectification, after the intervention of the Italian DPA, which required ChatGPT to inform users about its potential mistakes, seems to be a minor concern at the moment. In particular, as argued by some technology scholars, the exact deletion of personal data, especially in real-time, is hard to

[78] V. Mayer-Schönberger, *Delete: The Virtue of Forgetting in the Digital Age* (Princeton: Princeton University Press, 2009).

[79] J. Ausloos, The Right to Erasure in EU Data Protection Law. From Individuals Rights to Effective Protection (Oxford: Oxford University Press), 97.

[80] Generally, ChatGPT answers as follows: 'If John Doe is a private individual or someone not widely covered in publicly available sources, it might be challenging to provide detailed information. If there are specific details or context you can provide, it might help in giving more accurate information. Alternatively, you can check the latest online sources or databases for any recent developments related to John Doe'.

achieve and the suggested solution is that of the approximate data deletion from machine learning models,[81] which should be less time-consuming even if not totally complying with the provisions of Arts 16 and 17 GDPR.[82]

Furthermore, it is necessary to consider the profile of the falsity of the information provided and the possible liability of the platform operator.

In other words, it is necessary to separate two levels, that of the administrative sanction for unlawful processing of personal data from that of liability for false information. As noted above, an individual could ask ChatGPT to rectify information concerning him or her, pursuant to Art 16 GDPR and, according to the requirements of the Italian DPA, ChatGPT or another chatbot operator would be obliged to rectify the information provided as the output of a user's query. If, for example, I ask ChatGPT about who a public figure is, the latter may request rectification, in case of inaccurate information, or integration, in case of incomplete information. If ChatGPT does not comply with the data subject's request to rectify or erase some data, then the data protection authorities could sanction it with administrative fines.

In the latter case, the chatbot operator cannot be held responsible if the information is defamatory. In a scenario where ChatGPT claims that a public figure has committed a certain act, and that act represents a circumstance injurious to the honor or reputation of the public figure, intent on the part of the chatbot operator would be required. In the past, courts have dealt with cases involving suggestions from search engines. Despite some uncertainties, in most cases, the defamatory nature of the association between a person's name and certain criminal offenses has been excluded when facilitated by artificial intelligence, given the lack of an intention to harm the honor and reputation of that person. The same reasoning might be applied in the case of ChatGPT, where, not considering the cases of machine hallucination, information is provided to users automatically, without any human selection, such that it is impossible to attribute direct and malicious liability to the software programmers or to the managers of the chatbot.

IX. Age Verification and ChatGPT

The last aspect considered by the Italian DPA concerned age verification and the possibility for ChatGPT to limit its usage to individuals older than the age of sixteen years. As mentioned above, the Italian DPA had already taken action in this regard with respect to TikTok, by imposing more severe measures aimed at

[81] Z. Izzo et al, Approximate Data Deletion from Machine Learning Models, in Proceedings of the 24 the International Conference on Artificial Intelligence andS tatistics (AISTATS) 20210, San Diego, PMLR:Volume130.

[82] It is questionable whether AI based on machine learning is capable of erasing personal data (moreover, the data used for training is not recorded by the machine), see on this aspect 'We Forgot To Give Neural Networks The Ability To Forget', Forbes, 25 January 2023.

checking the age of the users.

Currently, Italian regulations do not explicitly require service providers to conduct specific age verification investigations, and generally websites rely on disclaimers through which users declare themselves to be of legal age.

In a comparative perspective, the English experience could show the difficulties in regulating these aspects. In UK, recently a controversial legislative provision has been adopted, named the Online Safety Bill,[83] which has introduced new obligations for tech platforms to prevent minors from accessing pornographic contents. The English law has garnered significant criticism, which can be summarized into three distinct strands: freedom of speech; privacy; and punitive measures.

Regarding the first aspect, critics have highlighted the fact that the Secretary of State and Ofcom[84] will have unprecedented powers to define and limit speech, without scrutiny by legislative bodies, potentially leading to a chilling effect on the quality of content transmitted online.

Additionally, the law imposes byzantine requirements, especially in the light of the guidance approved by Ofcom for the implementation of the specific measures to be adopted.[85] These costly measures, coupled with very high (including criminal) sanctions, could particularly deter startups, creating a competitive advantage for big tech firms, which are the only ones in a position to bear the transactional costs related to the implementation of these measures.[86]

Finally, the most challenging aspect concerning user privacy revolves around the potential obligation for platforms and internet service providers to monitor content exchanged among individuals. In fact, the Online Safety Bill undermines the core principle of the e-commerce directive (specifically, Art 15, Directive 2000/ 31/EC), stating that internet service providers are not liable if they have a merely neutral and technical role. For instance, according to some scholars, instant messaging services like WhatsApp or Telegram might be required to monitor user conversations to ensure there are no violations concerning minors.[87]

Similarly, following Ofcom's guidelines, it seems that a user might be compelled to prove their legal age by registering their identification document (eg driver's license, ID card, passport) to access an online service. This choice raises several questions about data collection, especially when it pertains to sensitive information such as sexual preferences (as in the case of adult websites), potentially leading

[83] On the long and widely discussed legislative process of the draft bill see V. Nash and L. Felton, 'Treating the Symptoms or the Disease? Analysing the UK Online Safety Bill's Approach to Digital Regulation', 2023, available at SSRN: https://ssrn.com/abstract=4467382

[84] Ofcom is the UK's communications regulator, with competences on TV, radio and video on demand sectors, fixed line telecoms, mobiles, postal services, online services.

[85] See Ofcom, Implementing the Online Safety Act: Protecting children from online pornography, 5 December 2023.

[86] See M. Lesh and V. Hewson, 'An Unsafe Bill: How the Online Safety Bill Threatens Free Speech, Innovation and Privacy, Institute of Economic Affairs Monographs' *IEA Briefing Paper*, 22 (2022) available at SSRN: https://ssrn.com/abstract=4172955.

[87] M. Lesh and V. Hewson, n 86 above, 10.

to discrimination, extortion, or other illicit behaviors. Similarly, the recourse to payment tools like credit cards does not seem more convincing either, as other documents would still be supplemented to provide age verification.

These major criticisms, however, seem less relevant in the context of chatbots. Firstly, at the moment, there is a limited number of operators like ChatGPT, and they have significant financial resources. Obligations could be tailored to the size of the provider, as held in the DSA with VLOPs (very large online platforms), imposing identification requirements only on the largest ones and preventing barriers to entry for smaller operators. Secondly, biometric authentication systems that do not store user data could be used: there are operators in the market who identify the user's age using artificial intelligence systems without storing the facial points used for biometric identification on servers.[88]

X. Final Remarks

In conclusion, aspects related to data protection seem ancillary in the debate concerning the development of generative communicative artificial intelligence. However, when ChatGPT was launched, GDPR was the only common legal source applicable to it. Currently, as also evidenced by the provisional text of the Artificial Intelligence Act, the focus is on issues involving social control over citizens through mass surveillance systems, behavioral manipulation or emotion recognition, leveraging AI and biometric technologies. While the EU's AI Act is still in the works and its latest version is not yet available to the public, it seems generative models such as chatbots are considered of limited risk. Thus, they would be subject to very light transparency obligations, such as the duty to advise users about any content generated by AI, as is already done by ChatGPT.

On the contrary, the threats arising from this new technology are multifold for individuals and societies, related to AI's pre- and self-training, text generation, and communicative power. The main concerns arise from the increasing agency demonstrated by AI and the interaction ability with humans. Regrettably, the AI Act considers artificial intelligence systems just as products and not as agents, stressing the risk-based approach.[89]

Moreover, analyzing the latest agreed version of the AI ACT, only non-European companies are obliged to monitor the development of their AI models and their impact on fundamental rights. The legislation thus sounds as a kind of protectionist policy promoted by European institutions, disguised as protection of fundamental rights. The renowned Brussels effect has raised a wall of protection of the values

[88] T. Sica, 'Dati biometrici, tutela del singolo e opportunità di mercato' *Diritti comparati*, 965 (2022).

[89] A. Mantelero and F. Fanucci, 'Great Ambitions. The International Debate on AI Regulation and Human Rights in the Prism of the Council of Europe's Cahai', in P. Czech et al eds, *European Yearbook on Human Rights* (Cambridge: Intersentia, 2022) 225.

enshrined in the Charter of Nice,[90] including data protection, by creating a safer digital place for all EU citizens who are users of digital services. However, in this sector, there seems to be a general feeling is that Europe, instead of fostering innovation and competitiveness, is confronting its lagging behind US and Asian countries (especially China and Korea) by preventing the invasion of technologies produced (and controlled) by third countries.[91]

However, it is necessary to be careful not to fall into the competing narrative. The interventions of the DPAs have raised many criticisms, as it has been argued that these provisions would hamper innovation and amplify disadvantages faced by European companies, who will be unable to use these artificial intelligence systems for further development. Theoretically speaking, law is aimed at selecting the interests that a given society is willing to protect. Europe has chosen to prioritize the protection of personal data over the uncontrolled development of technologies based on the liberal dogma that the market sets the rules. In the European constitutional tradition, human dignity takes precedence over liberty, consumers are protected from aggressive business practices, and finally, while the AI industry has so far built the datasets for AI models by indiscriminately scraping the web, it will now be forced to limit the use of personal data and process it by filtering the information on which the machines are trained.[92]

It is time to try to find a way to educate people so that they are aware of the risk involved not only in using AI, but also in interacting with it. The law should be given credit for this educational function. Like the consumer-professional imbalance recognized in consumer law, there is an imbalance of power and capabilities between humans and machines. Therefore, declaring the ontological vulnerability of humans in any interaction with AI, making the concept of digital vulnerability a new macro-category in private law, and interpreting existing norms or drafting future ones on its basis could be the right legal tool to lay the foundation for a global digital law.[93]

[90] A. Bradford, The Brussels Effect: How the European Union Rules the World (Oxford: Oxford University Press, 2020).

[91] On this aspect, see G. Greenleaf, 'The Brussels Effect of the EU's AI Act on Data Privacy Outside Europe' 171 *Privacy Laws & Business International Report* 1, 3-7 (2021).

[92] O. Pollicino, Judicial Protection of Fundamental Rights on the Internet. A Road Towards Digital Constitutionalism? (Hart Publishing, 2021).

[93] On the relationship between data protection and vulnerability see G. Malgieri, *Vulnerability and Data Protection Law* (Oxford: Oxford University Press, 2023).