

MARIA ZANZOTTO

**GENERATIVE AI, MANIPULATION, AND POLITICAL COMMUNICATION.
CAN GENERATIVE AI POSE A MANIPULATIVE THREAT IN SOCIAL MEDIA COMMUNICATION?**

1. Introduction

2. AI and political communication in social media: How technology (AI) changed the public discourse online 3. Generative AI tools enter the social media 4. Generative AI and manipulation

ABSTRACT: GENERATIVE AI, MANIPULATION, AND POLITICAL COMMUNICATION. CAN GENERATIVE AI POSE A MANIPULATIVE THREAT IN SOCIAL MEDIA COMMUNICATION?

In the last years concern over the impact of technology, specifically artificial intelligence (AI), on democratic institutions has been growing. In particular, concerning three phenomena on social media enabled by AI (recommendation algorithms) that are considered dangerous for democracy: disinformation, polarization, and manipulation. However, with the recent spread of generative AI tools that can produce visual (DALL-E, Midjourney, Stable Diffusion, etc.) and textual content (ChatGPT, Claude, Llama, etc.), the scope of the risk has broadened. This paper aims to explore how generative AI technologies impact democracy with regard to disinformation, polarization, and manipulation on social media. It is argued that generative AI technologies exacerbate the issues of disinformation (and partially polarization) because they allow the creation of false information at scale. However, the change is quantitative. Instead, when it comes to manipulation it seems that they pose a qualitative risk vis-à-vis recommendation systems. Firstly, they unlock new microtargeting possibilities and secondly, they can produce content almost indistinguishable from human-made one. The paper goes deeper into the manipulation discussion, addressing the issue on two levels: how non-transparency applies to generative AI systems and whether personal autonomy can be diminished through interaction with AI-generated content. The paper stresses the importance of expanding to generative AI technologies in the study of the impact of AI on social media and democracy.



1. Introduction

The rapid advancement of technology, particularly artificial intelligence (AI), has had profound effects on various aspects of society, including democratic institutions. Social media platforms, powered by AI recommendation systems, play a significant role in shaping the digital public sphere. These platforms, while initially celebrated for their potential to democratize communication, have increasingly become sources of

concern due to their facilitation of disinformation, polarization, and manipulation.

This paper aims to explore the impact of generative AI technologies on democracy, focusing specifically on their role in social media political discourse. The question this paper wants to address is: «Do generative AI tools, such as DALL-E, Midjourney, Stable Diffusion for visual content, and ChatGPT, Claude, and Llama for textual content, have expanded the scope of risks associated with AI on social media? Are they really changing the picture when it comes to disinformation, polarization, and manipulation? ». These tools not only amplify the existing issues of disinformation and polarization by enabling the mass production of false information but also introduce new qualitative risks in terms of manipulation. They allow for unprecedented microtargeting and the creation of content that is nearly indistinguishable from human-generated material, raising concerns about epistemic agency and personal autonomy.

The paper provides an in-depth analysis of how generative AI exacerbates disinformation, contributes to polarization, and poses novel challenges in terms of manipulation. The paper is structured as follows: firstly, it discusses the general role of AI in political communication on social media and its implications for democracy. This is followed by detailed sections on disinformation and polarization, highlighting the mechanisms through which AI recommendation systems influence these phenomena. Secondly, the paper provides some examples of generative AI tools (both visual and textual) that can have a real impact on disinformation, polarization, and manipulation. In the third section, the core of the paper, manipulation will be examined as a unique threat posed by generative AI technologies. Finally, the importance of expanding research on the impact of AI on social media and democracy to include generative AI technologies is emphasized.

2. AI and political communication in social media: How technology (AI) changed the public discourse online

Social media have radically changed the political domain: with the birth of a digital public sphere citizens can directly interact between themselves and with politicians even when physically far away. Social media sparked hope that a better democratic community would be achieved: after all, what would be a better opportunity to give a voice to each member of society? Of course, these hopes have only been met partially. On the one hand, people (especially minorities underrepresented in mainstream public discourse) finally had their public stage, research also has shown that social media has been a fertile ground for the spread of fake news and misinformation. On the other hand, concern about whether social media damages democracy has been growing¹. To complicate the picture, we have to consider the fact that artificial intelligence is present and shapes the digital public sphere. AI is a term that includes very different technologies, I now need to clarify what technology is involved in social media platforms. In this context, we are dealing mostly with AI recommendation algorithms, which are models based on machine learning. Contrary to other AI systems (like expert systems), that need explicit rules provided by the programmers, machine learning systems infer the rules by processing large quantities of data in the so-called "training". In this process - using algorithms and statistical models - the machine learning system identifies patterns in the data, which enables it to make inferences and predictions on new data. AI algorithms play a fundamental role on platforms like Facebook, X, Instagram, TikTok, and YouTube: they filter results, propose targeted ads and suggest content for users' feeds. These algorithms are based mainly on machine learning systems that use users' data to derive patterns of engagement with the aim of keeping the users inside the platform. For example, if I liked a

¹ N. Persily, J. A. Tucker (Eds.), *Social media and democracy: The state of the field, prospects for reform*, Cambridge University Press, Cambridge 2020.

video of a cat the algorithm will keep showing me videos of cats, and if users similar to me (some of those who like videos of cats) like videos of dogs, the algorithm might recommend me a video of a dog. These systems have been criticized for the radicalization of phenomena such as filter bubbles and echo chambers, as we will discuss. This is because if a user faces content and profiles that she likes, she will end up interacting only with “similar” users. This is not problematic per se – in real-offline life we look out for people with similar interests to us – however, when considering the political field and the broader consequences of AI, one might ask whether this technology is good or bad for democracy.

2.1. Three major phenomena online: disinformation, polarization and manipulation

This section is devoted to the study of those phenomena in social media that are enabled by AI and whose effects are detrimental to democratic institutions, namely disinformation, polarization, and manipulation. There is a variety of phenomena online that concern the political domain, I will focus on the three most discussed: disinformation – the relationship between AI and the spread of misinformation and fake news; polarization, more specifically in relation to AI-enabled Filter bubbles and Echo Chambers; and finally, manipulation – manipulation based on two phenomena: micro-targeting and interactions with non-human (bots) actors.

Firstly, disinformation. «A Lie Can Travel Halfway Around the World Before the Truth Puts On its Shoes» is a famous quote attributed to Mark Twain. Ironically, we have no proof that he ever said it and (as the quote goes) the lie spread. Along the same lines goes a more corroborated quote, older than Twain himself, the one written by Jonathan Swift in *The Art of Political Lying*² (1710):

² J. Swift, *The Art of Political Lying*. In «*The Examiner*», November 9, 1710. Available at:

Besides, as the vilest writer hath his readers, so the greatest liar hath his believers: and it often happens, that if a lie be believed only for an hour, it hath done its work, and there is no further occasion for it. Falsehood flies, and the truth comes limping after it so that when men come to be undeceived, it is too late; the jest is over, and the tale hath had its effect.

It seems that this general truth can be found also in social media platforms where fake news is extensively discussed. In this context, however, fake news seems to be just a part of a bigger problem. Therefore, a new term “informational disorder” has been coined by the Council of Europe to better describe the different phenomena around the circulation of false information online³. The informational disorder includes misinformation, disinformation, and malinformation. *Misinformation* happens when those sharing the false information are in good faith, for example when in the first days of COVID people shared voice records of alleged doctors via WhatsApp chats. *Disinformation* happens when false news is spread in bad faith with the intention of causing harm, for instance when the Notre Dame Cathedral in Paris went up in flames in 2019, a disinformation campaign initiated by right-wing extremists in Spain, France, Germany, and Italy blamed Islamic extremists for the fire to fuel anti-Muslim hatred in Europe⁴. And finally, *malinformation* is when true information is shared at the expense of the person involved without their consent, as it can happen when private information about a political adversary is disclosed to damage his/her reputation. I will leave the latter phenomenon aside, as it has nothing to do with false information, but rather with harming someone or a group with authentic, albeit private,

<https://www.fountainheadpress.com/expandingthearc/assets/swiftpoliticallying.pdf>.

³ C. Wardle, H. Derakshan, *Information Disorder: Toward an interdisciplinary framework for research and policy making*, in «Council of Europe DGI», September 27, 2017. Available at: <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html>.

⁴ C. Colliver, *Click Here For Outrage: Disinformation in the European Parliamentary Elections 2019*, in «Institute for Strategic Dialogue», 2020. Available at: <https://www.isdglobal.org/isd-publications/click-here-for-outrage-disinformation-in-the-european-parliamentary-elections-2019/>.

information. In the former two phenomena – misinformation and disinformation – AI plays an important role: it favors the spread of false information. The explanation lies in the fact that, as it has been shown, more sensationalistic content is actually liked more in comparison with more toned-down one⁵, and it is hence favored by the algorithm which pushes more the more liked/shared content. Fake news is usually sensationalistic and clickbait⁶ and has, thanks to the algorithms of the platforms, a fast track for visibility. It is the technology (the algorithm) that boosts the circulation of fake news. An example of a politically relevant case of disinformation through technology was a 2019 digitally altered video that showed Nancy Pelosi, the speaker of the US House of Representatives, stammer drunkenly through a speech at a news conference. While the video was soon debunked, it was shared on Facebook and afterward posted on Twitter by Donald Trump (then president), receiving millions of views.

Secondly, polarization: polarization refers to the phenomenon of dividing people into opposing groups. In the last decade, there has been a lively debate over the alleged surge of polarization worldwide, together with the rise of right-wing populism and the radicalization of the public sphere⁷. In social media political communication, polarization is usually discussed in connection with echo chambers and filter bubbles for their relation to the idea that people are divided into closed groups. Usually, the two are used as synonyms but actually are two connected but different phenomena: echo chambers are a radicalization of the filter bubbles experience. Filter bubbles refer to groups of people who

⁵ S. Khawar, M. Boukes, *Analyzing Sensationalism in News on Twitter (X): Clickbait Journalism by Legacy vs. Online-Native Outlets and the Consequences for User Engagement*, in «Digital Journalism», 2024, pp. 1–21.

⁶ R.R. Mourão, C.T. Robertson, *Fake news as discursive integration: An analysis of sites that publish false, misleading, hyperpartisan and sensational information*, in «Journalism studies», 20, 14, 2019, pp. 2077–2095.

⁷ P. Lorenz-Spreen, S. Lewandowsky, C.R. Sunstein, R. Hertwig, *How behavioural sciences can promote truth, autonomy and democratic discourse online*, in «Nat. Hum. Behav.», 4, 2020, pp. 1102–1109.

do not listen to those in a different group, whereas echo chambers refer to groups of people who *do not trust* those in a different position (who have a different opinion)⁸. Examples of echo chambers are anti-vax groups that gained relevance during the pandemic. Jiang and her colleagues found empirical evidence that «political echo chambers are prevalent, especially in the right-leaning community, which can exacerbate the exposure to information in line with pre-existing users' views»⁹. Members of no-vax echo chambers are deeply convinced that vaccines are harmful, and they would not trust who says otherwise¹⁰. To illustrate the difference between echo chambers and filter bubbles, consider as an example filter bubbles with COVID-19. In this context, people in filter bubbles are people who encounter some news outlets rather than others. For example, younger people would probably more likely have accessed news related to COVID-19 from media outlets targeting young people, like in Italy Vox Media or Will Media. Those outlets were more likely to publish news more appealing to their target, like news around mental health problems, and sacrifice topics less appealing to their target. The result is that people would have been ignorant about some topics. Still, echo chambers and filter bubbles are very different phenomena in magnitude: members of no-vax echo chambers actively refuse opposing views, whereas people inside filter bubbles are only partially informed and receive news confirming their beliefs if they do not actively seek out different news outlets and “exit the bubble”.

Filter Bubbles and Echo Chambers are not properly created by AI but are enabled by it. That is, filter bubbles and echo chambers would exist even without recommendation algorithms. People with

⁸ C.T. Nguyen, *Echo Chambers and Epistemic Bubbles*, in «Episteme», 17, 2, 2020, pp. 141-161. doi:10.1017/epi.2018.32.

⁹ J. Jiang, X. Ren, E. Ferrara, *Social Media Polarization and Echo Chambers in the Context of COVID-19: Case Study*, in «JMIRx Med», 2, 3, 2021. doi: 10.2196/29570.

¹⁰ W. Jennings et al., *Lack of Trust, Conspiracy Beliefs, and Social Media Use Predict COVID-19 Vaccine Hesitancy*, in «Vaccines», 9, 6, 593, 2021.

similar interests could find themselves even if no algorithm was in place. However, with AI it all becomes so easy and effective that it is almost inevitable for people with similar interests to cross. The algorithm is meant to recommend content you are or might be interested in so that you are kept engaged in the platform. This is visible also on the home page of different social media: each home page is personalized for each user. My YouTube page is different from yours. The YouTube algorithm "chooses" which videos are shown on my (and your) home page. The choice is based on my search history and my demographics, which are used to build my persona, which, in turn, is matched to those similar to me, so that I am recommended videos they liked. After some interactions, the algorithm has learned what content I am interested in so that it will tend not to propose videos very far from my persona. Moreover, when I am watching a video, or reading a post, I am interacting with people like me, with probably similar interests and demographics, and, depending on the type of content, also a similar political position. Following this mechanism, online everybody lives inside filter bubbles¹¹¹². Finally, if we combine the first point (disinformation) with the second (echo chambers and filter bubbles) we end up with closed groups that are not only exposed to content reinforcing their beliefs but also taking their position to a more radical one (more radical is more sensationalistic). This phenomenon is polarization magnified by AI. Not only are we invested of our ideas, but when we face those with a different view the conflict is exacerbated. And lastly, manipulation, which is the most complex of the three phenomena since we can understand it in many different ways. However, it is beyond the scope of this paper to provide a comprehensive and philosophical account of manipulation. I will

¹¹ S. Milano, M. Taddeo, L. Floridi, *Recommender systems and their ethical challenges*, in «AI & Soc», 35, pp. 957-967.

¹² C.R. Sunstein, *Republic.com*, Princeton University Press, Princeton 2002.

rather use practical accounts of manipulation using AI in social media that have a strong bearing on the political discourse.

The most discussed instance of political manipulation in social media is through political micro-targeting based on users' data. That is when highly personalized political messages are delivered to people based on their demographics, behaviors, or other data about them (like political leaning). This means that users' data are (unknowingly) used to maximize the impact of a campaign (i.e. win more votes).

The most popular example of political manipulation through microtargeting was the Cambridge Analytica scandal¹³¹⁴, when «Data from Facebook users were used by the Trump presidential campaign in 2016 to target voters based on their personality types. AI was used to categorize voters but also to automatically test thousands of variations of an ad before deciding which one to place»¹⁵. The scandal showed not only how AI can be used to manipulate voters, but also how effective it is. Users' behavioral data are constantly used to feed the AI recommendation algorithms as we saw before, but with Cambridge Analytica the mechanism of recommendation was brought to the political campaign, it was not about delivering entertaining content, it was about leveraging people's data to win votes and interfere with their beliefs' formation. In synthesis, Cambridge Analytica showed:

How powerful and effective AI is in grasping things we do not, namely the patterns between contents (machine learning);

How people do not perceive they are being manipulated;

¹³ M. Rosenberg, S. Frenkel, *Facebook's Role in Data Misuse Sets Off Storms on Two Continents*, in «The New York Times», March 18, 2018. Available at: <https://www.nytimes.com/2018/03/18/us/cambridge-analytica-facebook-privacy-data.html?hp&action=click&pgtype=Homepage&clickSource=story-heading&module=first-column-region®ion=top-news&WT.nav=top-news>.

¹⁴ E. Menietti, *Il caso Cambridge Analytica, spiegato bene*, in «Il Post», March 19, 2018. Available at: <https://www.ilpost.it/2018/03/19/facebook-cambridge-analytica/>.

¹⁵ M. Coeckelbergh, *Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence*, in «AI and Ethics», 3, 2023, pp. 1341-1350.

The fragility of democratic institutions towards this technology;
Power dynamics: Politicians paying Cambridge Analytica, and the role of Social Media platforms in harvesting users' data.

The phenomenon of using users' data harvested from their behavior to conduct micro-targeting ads has been extensively studied by Shoshana Zuboff in "The Age of Surveillance Capitalism"¹⁶ (2019). She argues that a new form of capitalism has emerged with internet companies, starting from Google which was the first to identify users' behavioral data as goods, what she calls "behavioral surplus". The aim of harvesting users' data was to sell targeted ad spaces to marketing companies. This way, they could maximize their return. The more data can be extracted from the user, the more profitable the platform can be. The surveillance capitalism business model was so efficient that it was soon embraced by most companies online, including Meta. The Cambridge Analytica project can be interpreted as a way to apply the model at the political level. Instead of being a company trying to sell its products, it was the political actors trying to win votes.

Another interesting instance of manipulation is not being able to discern if one is interacting with a human or not. Social media platforms are full of bots, and it is sometimes difficult to spot them straight away. Bots can influence users in two main ways: publishing posts and liking/sharing existing posts. In the first use, bots can amplify the importance of a topic/hashtag. In the second use, bots can have the effect of amplifying the influence of a post, making it more successful than it would be ("pumping" the algorithm), and this can lead people to find the content more legitimate because it has more likes/shares. Even if right now the problem is not particularly worrisome, it might become if interaction with bots becomes more and more human-like, for instance, if bots are enabled by generative AI technologies.

¹⁶ S. Zuboff, *The Age of Surveillance Capitalism*, Profile Books, London 2019.

To summarize, non-generative AI – mainly through platforms’ algorithms – changes disinformation, polarization, and manipulation¹⁷: social media algorithms tend to favor the spread of fake news, recommendation systems cluster people based on their personal preferences and demographics, and with microtargeting enabled by machine learning systems, the manipulation has been achieved at scale.

2.2. AI and political communication in social media: risks for democracy

Let me briefly express in what sense I am talking about democratic risks posed by AI technologies. Here I want to clarify two aspects. The first is that my account will deal with democracy in so far as political discourse is considered, more specifically political discourse on social media networks. I do not mean that the risks AI poses to democracy are confined to risks in the political discourse on social media of course, there are many instances where AI (both generative and not) would impact democracies outside social media discourse, for example in the algorithmic-based procedure of benefits distribution (like public housing). Unfortunately, the scope would broaden too much, hence I have to restrict the area of research. I could have simply referred to *political discourse* in social media networks without the *democratic* qualification, but here lays my second clarification: what I take as democracy. Let me preface this by saying that I do not intend to provide – nor is this – a proper account of democratic political theory. I will not discuss what democracy means nor will I enter political philosophy debates in depth. I will be greatly simplifying by saying there are two main conceptions of what a democracy is: a thin and a thick version¹⁸.

¹⁷ Although humans already spread false information, aggregate in groups, and manipulate others without AI, the introduction of AI algorithms worsens these three phenomena.

¹⁸ M. Coeckelbergh, *Why AI undermined democracy and what to do about it*, Polity Press, Cambridge 2024.

The former holds that a democracy is a political system according to which citizens have the right to vote for their representatives. The representatives change every set number of years, they have to conform to the democratic laws of the country, and they have to respond to citizens' needs/requests/etc. to be able to win their votes. On the other hand, citizens should make informed choices when they get to the ballot box. This is why I talk about “*democratic political discourse*” and not just “*political discourse*”, it is because the epistemic layer is very important as a function of citizens' right to vote. The thick account starts from the thin one but adds requirements that make the conception fuller: shared democratic principles guiding citizens and their representatives, and the idea of promoting the “*public good*”. This conception builds on top of the thin, which is not discarded but rather augmented. However, given that disinformation, polarization, and manipulation have an impact on the thin conception of democracy already, as I will discuss shortly, I take it as a good starting point for research. Hopefully, future work can target the thicker conception of democracy. The thin conception of democracy relies heavily on the epistemic environment where citizens form their political beliefs which are reflected in their votes. The epistemic environment in social media is particularly interesting as to how AI impacts it as we discussed before (disinformation, polarization, and manipulation). Now I will try to delineate in a very simplified and introductory way why disinformation, polarization, and manipulation are considered evils of a healthy public sphere.

First, the spread of misinformation and disinformation online can be problematic at a political level because citizens are supposed to make informed decisions especially when it comes to voting¹⁹. But when the epistemic environment in which they form their beliefs is polluted by fake news, circulating via popular actors,

¹⁹ C.R. Sunstein, *op. cit.*

common people, and friends, it becomes more difficult to navigate and reach reliable news.

Secondly, echo chambers and filter bubbles are problematic for democracy from a political philosophical perspective that finds the dialogue or disagreement among different people the starting point of a democratic deliberation process aimed at reaching a consensus. If people are sealed off, we lose the possibility of dialogue. Even worse when we consider echo chambers, where members refuse to come to terms with those outside of it. Moreover, since the content proposed is always similar to one's own beliefs, due to confirmation bias people reinforce those beliefs, and once the belief is formed, they are less likely to revise them and listen to others' points of view. Hence, people end up only partially informed but strongly firm in their beliefs. Instead, in democratic and pluralistic societies citizens should be encouraged to listen to those with different beliefs and to peacefully engage with them. Dialogue and ideas exchange is a fundamental epistemic aspect of political belief formation in theories of deliberative democracy²⁰. Polarization is argued to be detrimental to democracy because it can render a society less prone to find consensus and agreement. As our society becomes more pluralist, finding a consensus (or an overlapping consensus if we adopt Rawls' terminology²¹) should become a priority, but the effort can be hindered by polarization if citizens have so strongly opposing goals. Moreover, polarization can fuel social divisions and fragmentation, endangering a country's social fabric and social cohesion. Even in an agonistic conception of politics²², where citizens should be actively seeking conflict those with opposing views (hence exiting their own bubble or echo chamber, which is not easy), they are to be considered opponents and not *enemies*.

²⁰ J. Habermas, *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*, The MIT Press, Cambridge MA 1996.

²¹ J. Rawls, *Political Liberalism*, Columbia University Press, New York 1993.

²² C. Mouffe, *Agonistics: Thinking The World Politically*, Verso, London-New York 2013.

Agonistic conceptions see conflict – not consensus but neither war – as the basis of a vibrant political community (again, I am simplifying). Members of the political community should respect each other even in the context of pluralism, but in a polarized society respect is not easily granted. However, here we risk crossing to the thicker version of democracy, for ideals of social cohesion and fraternity are taken into consideration.

Finally, from a political philosophy point of view manipulation is problematic for the notion of autonomy. Autonomy has a long history in philosophical thought, and it is regarded as one of the most distinctive features of being human. According to Berlin, one of the pillars of the liberal theory, freedom should be conceived not just as a state without external interferences, but also as a state where one can practice his autonomy (what he calls positive liberty), as in acting by his intentions or his belief about what is best for him, being the master of his own life (Berlin, 1969, chapter "Two Concepts of Liberty"²³). Therefore, violations of autonomy amount to interferences at a deeper state, at the self-determination level²⁴. For example, paternalistic accounts have been criticized for violating positive liberty. Nudging – the practice of gently pushing individuals towards better societal choices (like getting vaccinated, donating blood and organs, eating healthier, etc.,) through the interference in their choice architecture – has been considered a violation of positive liberty. Coeckelbergh in the book *Political Philosophy of AI* considers the practice of micro-targeting as nudging and criticizes its paternalistic implications²⁵. But I will come back to Coeckelbergh's argument in section 3.

²³ I. Berlin, *Four Essays on Liberty*, Oxford University Press, London 1969.

²⁴ For reasons of space I didn't consider different philosophical schools, like neo-republicans and the concept of non-domination (for a start: P. Pettit, *The Freedom of the City: A Republican Ideal*, in *The Good Polity*, eds. A. Hamlin, P. Pettit, Blackwell Publishers, Oxford 1989; P. Pettit, *Republicanism: A Theory of Freedom and Government*, Clarendon Press, Oxford 1997).

²⁵ M. Coeckelbergh, *The Political Philosophy of AI*, Polity, Cambridge 2022, p. 28.

3. *Generative AI tools enter the social media scene*

In the last section, I have discussed three main threats (disinformation, polarization, and manipulation) to democratic institutions in the context of social media, and I have outlined how AI is playing an important role. In this section, I will focus on my research question: do generative AI technologies, like ChatGPT or Midjourney, change the picture? And if yes, how? What is their impact on disinformation, polarization, and manipulation? To visualize and locate these questions the table here below can come to our aid: we are considering the column "Generative AI" on the right and we are asking ourselves how to fill the respective boxes.

TABLE 1

	Without AI	Non-generative AI	Generative AI
Disinformation	Yes	Yes	?
Polarization	Yes	Yes	?
Manipulation	Yes	Yes	?

Indeed, with the emergence and sudden spread of generative AI tools like OpenAI's ChatGPT the landscape of AI technologies has dramatically changed: generative AI technologies have taken the stage, and new tools to create images, music, and text are at everybody's hands.

AI-generated content has already flooded social media platforms, from synthetic pictures to posts written using text generators. The question is: given the impact AI algorithms already have on democratic institutions in social media (section 1), is there a new risk now with the emergence of these new technologies? What is their impact from a political point of view? What changes with generative AI systems from non-generative/non-generative AI when it comes to disinformation, polarization, and manipulation? Not only will I try to understand if and how generative AI can enter this debate, but I will also try to clarify if it entails a quantitative or qualitative shift. Let me briefly explain what I

mean by "quantitative" and "qualitative" shifts. The former represents an improvement in terms of efficiency, meaning when one using generative AI can do a task with less effort compared to using non-generative AI systems. The latter represents a different kind of jump, a new type of jump: when generative AI enables one to do something that cannot be done with non-generative AI tools alone. I will argue that when it comes to disinformation and polarization there is a quantitative change (generative AI technologies unlock new levels of efficiency), whereas for manipulation I will try to outline a qualitative change (generative AI can do not just more, but new things). The risk of manipulation when it comes to generative AI is particularly still in its infancy in the literature, hence I will try to address the issue in a section devoted to it (section 3). But since generative AI tools have a broader and relevant impact on democracy it is useful to discuss the quantitative changes as well. My research, being only referred to shifts from non-generative AI to generative AI, will leave aside shifts from humans to generative AI, which are broadly considered quantitative²⁶ (write text, create paintings/pictures faster, etc.).

Before addressing the issue, a better explanation of what is meant by generative AI is needed. Differently from the categorization of AI subfields based on how (technology-wise) the model works, the category "generative AI" is functional: it defines technologies based on the kind of output they produce. More specifically, those able to produce new content: texts, images, music, video, etc. Being a functional demarcation, generative AI systems encompass a variety of different systems technology-wise, even though the more sophisticated models are based on neural networks. Neural networks are a subcategory of machine learning that tries to model the

²⁶ Some argue that the quantitative and qualitative aspects of the shift from human to AI systems (both generative and non-generative) merge. This convergence occurs because the quantitative force - driven by vast amounts of data and immense computational capacity - is so powerful that it crosses into the qualitative realm.

system like a brain: the networks are formed by neurons (technically called nodes) interacting with each other. For example, ChatGPT is a chatbot based on a neural network, specifically ChatGPT is based on a Large Language Model²⁷. The power of these models is strongly given by the amount of data used and their complexity (the number of parameters, estimated to be billions or trillions). The more powerful they are the better the outputs can be, but also the more impactful they are on natural resources: energy consumption, water consumption, and minerals used for GPUs, storage sections, and general computer components.

3.1. *Generative AI: creating images and text*

Let me clarify how I will proceed in my analysis: I will distinguish two categories of generative AI tools and for each category I will propose actual applications or case studies that have a strong impact in the political sphere. The first category is AI-generated images; in this regard, I will discuss an AI-generated picture of Trump with a group of black people that went viral. The second category is AI-generated text/AI that understands texts. Here, I will explore 3 applications: AI-generated-fake news; AI-generated-targeted context; and finally, bots powered by generative AI systems.

3.1.1. *AI-generated images*

An AI-generated image is an image generated by an AI software, that following the user's textual description of the desired image

²⁷ Large Language Models (LLMs) are a category of foundational models (i.e. models that work on large amounts of data) capable of understanding and generating natural language, and hence are used for a wide range of tasks, like summaries, translations, and text predictions. LLMs work through highly complex architectures called transformers – models that were introduced by Google in a very influential 2017 paper titled *Attention is all you need*. A transformer is a «neural network that learns context and meaning by tracking relationships in sequential data, like the words in this sentence» (Nvidia, n.d.). For a technical, but easy to follow, explanation of how LLMs work I suggest reading this engaging blog post divided in levels from beginner to advanced: A. Stöffelbauer, *How Large Language Models Work*, in «Medium», October 24, 2023. Available at: medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f.

(i.e. the prompt) can produce the visual output. The most popular tools are Midjourney, DALL-E, and Stable Diffusion, which employ increasingly complex models (and increasing amounts of data) to achieve higher-quality results. This is unprecedented, right now we have at our hands tools that can produce visual content of quality that until some years ago could only be made by a human, whether it is a portrait, a picture, a cartoon, or a piece of conceptual art. AI-generated content can have as many uses as you can think of. In this paper, I am interested only in uses that can have a political implication, or that can be used for political aims, especially during an electoral campaign.

One case study that can be useful to understand the political implications of these tools is a group of AI-generated image of Trump pictured with a group a black people that went viral on X (picture below). The episodes have been explored by a BBC Panorama research²⁸. At the beginning of March 2024, the BBC Panorama investigation tracked down on X many AI-generated pictures of Trump with a group of black people. These pictures were generated or shared by Trump supporters. The BBC interviewed Cliff Albright – the co-founder of the campaign group Black Voters Matter – who said the fake images «were consistent with a ‘very strategic narrative’ pushed by conservatives – from the Trump campaign down to influencers online – designed to win over black voters », adding: «They are particularly targeting young black men, who are thought to be more open to voting for Mr. Trump than black women»²⁹. Can these pictures be effective? We would need a larger study on the matter that we, unfortunately, do not have, as it is now it will suffice that the BBC asked a black man likely to be a target of this kind of communication what he thought of the picture, and he admitted the picture «bolstered his view, shared

²⁸ M. Spring, *Trump supporters target black voters with faked AI images*, in «BBC Panorama», March 24, 2024. Available at <https://www.bbc.com/news/world-us-canada-68440150>.

²⁹ *Ibid.*

by some other black people he knows, that Trump is supportive of the community»³⁰, only later the BBC told him the image was fake, he said how easy it was to be fooled.



3.1.2. *AI-generated text*

AI tools that generate text have boomed after the release of OpenAI's ChatGPT at the end of 2022. Previous attempts to develop models that seem to be able to "comprehend" and generate texts did not produce sensational outputs. But with the increasing computational capacity and the increasing ability to process huge amounts of data, the new models (like those of OpenAI) have indeed produced impressive results. Tools like ChatGPT, but also Google's Gemini, Anthropic's Claude, and Meta's Llama, seem to "understand" and creatively come up with textual content following the user's prompts. Like AI images, AI text can be used for a lot of applications, but again, I will only focus on those having to do with political communication on social media. More specifically, I will explore three applications: firstly, AI-generated fake news, secondly AI targeted messages, and third bots powered with generative AI tools.

Firstly, the most obvious application: AI-generated fake news, meaning using AI tools to ask them to come up with new fake news. With tools like ChatGPT creating fake news is a matter of minutes. Moreover, with the right prompt fake news can be generated on the

³⁰ *Ibid.*

style and structure of past ones, so that one can be copied from another with ease. Hence, the technology makes the creation of fake news at a large scale possible.

Secondly, it is possible to create messages targeted for specific groups at a large scale, that is to apply a micro-targeting strategy for political ads, in the same way it can be done for general marketing advertising. An interesting but distressing frontier of micro-targeting is to target people at their personality level. Some preliminary studies³¹³² warn us that it would be possible to create messages targeting personality traits – a technique called Personalized Persuasion (PP) – at scale thanks to the use of generative AI tools. One study conducted by computational social scientist Sandra Martz from Columbia Business School and her colleagues tested the effectiveness of AI-generated targeted messages against general ones. They gained preliminary evidence that the targeted ones are more persuasive:

Of the 33 message instantiations we tested, 30 were directionally effective, and 20 were significantly so (61%;). This proportion of significant effects is higher than chance ($t = 8.30$, $p < .001$). When extrapolating this effect to the hundreds of advertisements people see daily, the ease with which AI can personalize persuasive message makes their potential influence unprecedented³³.

It is possible to take this application a step further: firstly, by letting the AI system infer the personality trait and secondly, by automatically creating a personalized message. It is possible to infer a personality trait with a “classic” approach by processing users’ activity online (profiles interacting with, pages followed, likes, main topic shared, etc.), but one study conducted by researchers from the Ben-Gurion University of the

³¹ S. Matz, J. Teeny, S.S. Vaid, H. Peters, G.M. Harari, M. Cerf, *The Potential of Generative AI for Personalized Persuasion at Scale*, in «PsyArXiv», 2023.

³² A. Simchon, M. Edwards, S. Lewandowsky, *The persuasive effects of political microtargeting in the age of generative artificial intelligence*, in «PNAS Nexus», 3, 2024, pp. 1-5.

³³ S. Matz, J. Teeny, S.S. Vaid, H. Peters, G.M. Harari, M. Cerf, *op. cit.*, p. 20.

Negev and the University of Bristol³⁴ developed a machine learning system able to infer personality trait both for content written by the users and for text consumed by them.

A following study by the same researchers puts the two aspects (the ability to infer a personality trait and the ability to create personalized content) together. In the paper, they argue the technologies at our disposal can «design messages that appeal to these traits – even when given very limited input. In other words, current technologies, which continue to innovate and improve, have the potential to allow message sources to fully “close the loop” on automating personalized persuasion»³⁵.

Thirdly, we can imagine bots enabled by LLMs: a 2017 article³⁶ about the interaction of humans with bots in social media spaces warned us of their impact when it comes to influencing purposes. The author – Professor of Artificial Intelligence and Data Science at the University of Toulouse – reports how easy it is to influence people on social media employing bots: «By simply selecting random popular words and parroting other users’ tweets, one researcher’s Twitter bot was able to reach influence scores close to celebrities and higher than many human users»³⁷³⁸. As discussed in section 1, bots can be used in two main ways: to publish posts and to like/share existing posts. For both uses, there is a crucial aspect to consider, if humans understand they are interacting with a non-human. In the experiment «this bot was intended to deceive human users into believing it was also human, and it appears to have succeeded. The difficulty of separating a bot, even a simple scripted one, from a human user on Twitter is

³⁴ A. Simchon, A. Sutton, M. Edwards, S. Lewandowsky, *Online reading habits can reveal personality traits: towards detecting psychological microtargeting*, in «PNAS nexus», 2, 6, 2023, Article pgad191.

³⁵ A. Simchon, M. Edwards, S. Lewandowsky, *op. cit.*, pp. 22-23.

³⁶ D.G. Wilson, *The ethics of automated behavioral microtargeting*, in «AI Matters», 3, 3, 2017, pp. 56-64.

³⁷ *Ibid.*, p. 59.

³⁸ The research cited is J. Messias, L. Schmidt, R. Oliveira, F. Benevenuto, *You followed my bot! transforming robots into influential users in twitter*, in «First Monday», 18, 7, 2013).

so difficult that modern AI has been utilized to perform the task»³⁹. What is interesting is that these bots are not even powered by LLMs, they use less sophisticated models and can produce unnatural outputs. However, since just five years after the study was published, we have been having at our hands very powerful tools that can seemingly create human-like texts. Therefore, we can imagine these discussed phenomena with bots even more hardly detectable by humans, and hence, more powerful in influencing people.

Before discussing how generative AI can impact disinformation, polarization, and manipulation, let me make a short premise of philosophy of technology. In this paper, I understand AI as a tool, primarily. Hence, intentionality is always placed in the person who uses it, in this way, it is quite linear to think of disinformation, polarization, and manipulation as activities in which the human being makes use of AI to achieve their purposes. However, it is also true that AI is a particular kind of tool: it has a higher level of autonomy (in terms of how it arrives at a particular result, not in terms of a will) than other technologies, raising the problem of the so-called "unintended consequences". A well-known instance in this sense are AI biases, whereby AI can have harmful consequences that the designer and user did not intend to cause. Generative AI tools are part of this family of technologies and can have these kinds of consequences. Therefore, in this paper, I am assuming that the person, when using a generative AI tool for disinformation, polarization, and manipulation aims, can monitor and change the results in a way that aligns with their purposes. A trivial example is when a person wants to create a fake picture of an event, they can edit the prompt until they get a result they find satisfactory.

³⁹ D.G. Wilson, *op. cit.*, p. 59.

3.2. *Generative AI and disinformation, polarization, and manipulation*

Now that we discussed the technologies at stake, it is time to discuss how these technologies impact the three phenomena problematic for democracy.

3.2.1. *Disinformation*

AI-generated pictures as well as AI-generated texts can be a vector for misinformation and disinformation. For example, in the case of Trump posing with a group of black people smiling, the fact that there is a picture makes the event more credible than just if there was a post describing it. Hence, on the one hand, people can start sharing this kind of content believing it is true when it is not (misinformation). On the other hand, it becomes easier to construct and post impactful and effective fake news (disinformation). It is not surprising that the picture of Trump with the group of men on the porch went viral.

Also, the same holds for AI-generated fake news even if only textual. Here, the powerful aspect is not the truthfulness of the post but the potential for creating fake news at scale. Using this technology, it becomes very easy to come up with stories that can resonate on the web and go viral. AI-generated fake news can clearly aggravate the problem of disinformation because it reduces the time and effort needed to create it. Therefore, I argue that the shift is quantitative: generative-AI tools enable more efficient production of both fake images (e.g. fake photorealistic pictures) and false narratives (e.g. writing fake news). It would also be possible to combine an AI-generated fake news with AI-generated pictures backing the story.

3.2.2. *Polarization*

I argued that members of filter bubbles and echo chambers will likely be faced with content that reinforces their existing beliefs, therefore when content is aligned with their beliefs,

they will less likely check if it is real or fake. This is problematic when we think at scale. If it becomes easier to produce targeted content (including fake ones) with generative AI tools, we can argue that people will be more likely exposed to it when navigating through social media. Given the high resemblance of AI-generated content to human one and given that it is possible to target specific groups appealing to their values and beliefs, they will more likely accept the content, and be convinced by it. For example, the Trump picture targeted the black community, and it was effective in confirming the belief of the Trump supporter that Trump is by the side of the black community. Moreover, these kinds of content targeting specific groups tend to spark a very polarized discussion from the two opposing views, for example, whether Trump is pro or against the black community, fueling a conflict that divides the country. We can imagine textual content targeting a specific group and expect similar results.

However, it seems that generative AI can only partially contribute to political polarization and the radicalization of echo chambers, it seems more that it is the environment that favors the acceptance of AI-generated fake news and therefore aggravates the problem of disinformation. The role of the environment in the formation of our knowledge and specifically the ability to discern good sources from bad sources has been studied by Neil Levy. He argues that we underestimate the power of the epistemic environment in which we form “good beliefs” and “bad beliefs”, and that the major cause of bad belief formation lies in the pollution of the epistemic environment⁴⁰.

Targeting groups with specific features is also an aspect of manipulation, which is the phenomenon I will turn to now.

⁴⁰ N. Levy, *Bad Beliefs: Why They Happen to Good People*, Oxford University Press, Oxford 2021.

3.2.3. *Manipulation*

In section 1.1 I explored two different aspects of manipulation namely micro-targeting and human-like interactions or indistinguishability more broadly. Now let us understand and decline both with the applications discussed in 2.1.

Microtargeting with AI-generated images can be considered as to target a specific group of people with a particularly appealing image for the group in question i.e. black men in the example, however this system would substantially rely on platforms' algorithm to distribute the content. With AI-generated texts it would be possible to target people with personalized content and, more interestingly, use generative AI for personalized persuasion. The indistinguishability aspect is present both with images and text, but it takes different forms: with images it is difficult to ascertain if the content is authentic or synthetic, with text since I am referring to text produced by bots the indistinguishability lies in not being able to identify if the interacting user is a person or not.

Therefore, to sum up the candidate for manipulation with generative AI systems are:

1) targeting and micro-targeting: generative AI tools - I am thinking about systems that understand texts - can extrapolate users' traits and use them to produce content (text and images) suited for them.

2) indistinguishability - the quality of the AI-generated content is very high: it looks true or truthfully made by a human. With that I mean that the tools to create it are so advanced that humans are not able to detect they are made with AI.

To better visualize the new area of research let's create a new table, with more details for generative AI, and for the two elements of manipulation - 1) microtargeting and 2) being indistinguishable by humans. Generative AI can be distinguished in images and textual outputs (with images I mean both artistic

pictures and photorealistic ones, with text I refer to the application outlined in section 2.1.2, meaning the possibility to achieve personalized persuasion (PP) and bots enabled by LLMs).

TABLE 2

		Generative AI	
		Images	Text (PP and bots)
Manipulation	Microtargeting	?	?
	Indistinguishability	?	?

The two forms are both good candidate for manipulation, but it remains unclear on what basis the manipulation relies on. This will be the focus on the first part of the next session, and the second part will answer the qualitative or quantitative question regarding manipulation. Differently from disinformation, which is easily interpreted as a quantitative shift, the shift in manipulation remains less clear-cut, and still underdeveloped in the literature.

Let’s summarize the findings so far. In the table below it is possible to see how generative AI impacts disinformation, polarization, and manipulation compared with non-generative AI.

TABLE 3

	Without AI	Non-generative AI	Generative AI
Disinformation	Yes	Yes	Yes (quantitative)
Polarization	Yes	Yes	Indirectly (quantitative)
Manipulation	Yes	Yes	Probably (quantitative or qualitative?)

4. *Generative AI and manipulation*

In the first section, we argued that non-generative AI such as recommendation algorithms can have negative effects on democracy,

namely the spread of disinformation can hinder citizens' ability to form informed beliefs, polarization fuels social division, and manipulation conflicts with one's autonomy.

In the second section, we introduced generative AI tools and possible uses that can impact disinformation, polarization, and manipulation. Disinformation and polarization have already been discussed, it is now time to discuss manipulation with generative AI systems.

I initially stated that I would not write a paper on the definition of manipulation but would instead focus on cases of manipulation occurring on social media. This still is true, but to establish a clear reference point for this issue, I will rely on a study that specifically examined manipulation through microtargeting on social media, namely, I will use Ienca's work⁴¹ as a basis to expand the concept of manipulation through microtargeting with non-generative AI to also include microtargeting with generative AI. Furthermore, I will attempt to broaden his understanding of manipulation to encompass the notion of indistinguishability.

Ienca defines digital manipulation as «any influence exerted through the use of digital technology that is intentionally designed to bypass reason and to produce an asymmetry of outcome between the data processor (or a third party that benefits thereof) and the data subject»⁴².

He identifies 4 conditions that need to be satisfied for an act to be considered manipulatory:

A. *Intentionality*: Manipulation involves the intention of the manipulator to exert an influence on someone else's behavior and or system of beliefs.

B. *Asymmetry of outcome*: Manipulation involves behaviors that result in positive outcomes for the manipulator (e.g., personal gain) and negative outcomes for the manipulated (e.g. physical or psychological harm, performance of actions that are not in the best interest of the

⁴¹ M. Ienca, *On Artificial Intelligence and Manipulation*, in «Topoi», 42, 2023, pp. 833-842.

⁴² *Ibid.*, p. 840.

victim). After the manipulation, the manipulator is better off while the manipulated is worse off.

C. *Non-transparency*: Manipulation is inherently non-transparent as it involves a form of influence that is generally covert and hard to detect for the victim.

D. *Violation of autonomy*: due to its non-transparent character, manipulation involves a violation of the personal autonomy of the manipulated individual or group, as their ability to make free and competent decisions is diminished or even obliterated⁴³.

These conditions can give us an orientation to distinguish manipulation from other forms of influence like persuasion, that is not as ethically deplorable⁴⁴. I distinguish between the first two and the last two: the first are “internal” and the last “external” from the manipulator. The internal conditions A and B are difficult to ascertain, as we would need to know the manipulator’s intentions and gains and therefore, we would need to get into the details of each case, which is not always possible⁴⁵. Conversely, condition C has to do with the modality of the manipulation (i.e. in a hidden and non-aware way), and condition D has to do with what the manipulation occurs through, manipulation involves a violation of autonomy. Condition C and D can be achieved through the behavior of the manipulator but also with the use of a specific technology. In the case at stake, I am considering manipulation through the aid of generative AI. Therefore, to argue that microtargeting and indistinguishability with generative AI can be seen as manipulation I will focus on the conditions of non-transparency and violation of personal autonomy.

1) Let’s start with the first aspect, which is microtargeting.

⁴³ *Ibid.*, p. 837.

⁴⁴ *Ibid.*, p. 836

⁴⁵ Since I am considering humans as carriers of intentions, personal preferences, gains and losses, and AI as a tool used by them to achieve their goals, with the focus of understanding how effective AI is as an ally, I am not focused on the psychological and personal aspects of the manipulator, i.e. I am not ascertaining whether the person has the intention to manipulate and has gained by manipulating as this would be a very different kind of inquiry from the one I want to conduct here. On the other hand, I am not considering AI systems as bearers of intentions or personal gains, to argue that generative AI has intentions, autonomy, and the like would be, again, a rather different inquiry.

As Ienca points out: «In particular, since microtargeting advertising uses personal data to tailor ads to individual users and thereby to influence their choices, this has raised concerns about the manipulation of political and consumer behavior»⁴⁶.

Ienca defines micro-targeting advertising as:

A technique used by advertisers to deliver personalized and highly targeted messages to specific individuals or groups based on their demographic, behavioral, or psychographic characteristics. This technique involves collecting and analyzing large amounts of data about individuals from various sources, such as social media platforms, search engines, and third-party data brokers, and using this data to create highly customized advertising campaigns. Micro-targeting aims to deliver messages that are highly relevant and appealing to individual users, increasing the likelihood that they will engage with the advertisement or take a desired action, such as making a purchase or sharing the message with their social network⁴⁷.

From this quote it could seem that micro-targeting has only to do with advertising, so one might be asking what it has to do with politics, but as Ienca rightly points out «Microtargeting can be used for a wide range of purposes, including political campaigning, product promotion, and social advocacy»⁴⁸. Hence, political communication is one of the possible areas of application of micro-targeting.

He argues that microtargeting counts as manipulation because it satisfies all four conditions, specifically, non-transparency and violation of personal autonomy. He criticized the (hidden) process of collecting personal data for which users are mostly unaware of the significance.

He considers microtargeting as a violation of personal autonomy as it pushes people to make decision (also political) in a way that undermine the capacity of people to act by their own will. Ienca argues that by receiving continuous targeted messages the consequence is an interference in personal cognitive liberty, a prerequisite of personal autonomy. He talks about cognitive liberty and manipulation saying that:

⁴⁶ M. Ienca, *op. cit.*, p. 839.

⁴⁷ *Ibid.*

⁴⁸ *Ibid.*

Manipulation can be seen as a potential violation of cognitive liberty because it undermines people's right to self-determination. When individuals are manipulated, their thoughts, emotions, and perceptions are (to a variable extent) being controlled or influenced by others, often without their knowledge or consent. This can interfere with their ability to exercise control over their consciousness and may undermine their sense of autonomy and dignity. This form of violation of personal autonomy is subtler than coercion because it does not limit to controlling or limiting another person's behavior, but interferes with an underlying and antecedent level, i.e., that person's mental self-determination⁴⁹.

Violations of autonomy are violations of positive liberty in Berlin's sense. That is, nothing is imposed on me, there's no clear interference that limits my freedom. Contrary, positive liberty has to do with being able to decide following one's own reasons and mind, being one's own master, and in the formulation of Ienca, self-determined. Given that users receive continuous ads and recommendations on what they should watch, buy, like, comment on, etc., their ability to exercise control over their mental self-determination is at risk. There could also be a paternalist reading of this mechanism: recommendation systems make you buy products/watch content that *you should want to buy/watch* as if you do not really know what you should buy/watch.

Now let's turn to the problem of microtargeting with generative AI. As described, there's the technology to detect information about ourselves either by looking at what we write or what we consume, and based on that information the system targets us with personalized messages. The novel aspect is only partly the inferring information - harvesting data from our social profiles has been done with precise results even without generative AI. What is new is the ability to create high-quality personalized messages at scale with minimal human intervention. How does this technology do in terms of non-transparency and violation of autonomy?

Starting with non-transparency, people are not aware of how much companies know about them, and how much they leverage that

⁴⁹ *Ibid.*, p. 838.

knowledge to maximize the impact of their campaigns. Moreover, the whole process is non-transparent: the way data are collected is non-transparent, and the way they are elaborated is non-transparent. And finally, it is not transparent how data are used to create the targeted content. More deeply, the technology used to generate content is non-transparent itself. Text generation tools like ChatGPT are based on neural networks, systems referred to as "black boxes": we know the data that come in and the output, but we have no access to why the systems elaborated the data the way they did⁵⁰. Therefore, I argue this condition is satisfied.

Turning to personal autonomy, with generative AI the personalized nature of microtargeted content remains, but it is further amplified by dynamically crafted messages tailored to individual traits, including personality. Furthermore, since the same data collection practices of traditional microtargeting apply, the considerations regarding both sides of traditional microtargeting (targeting and data collection) hold.

Therefore, since Ienca argues that traditional microtargeting meets the condition of violating personal autonomy, and generative AI does not significantly change this practice, we can conclude that personal autonomy is also violated when using generative AI. By combining the considerations of non-transparency and personal autonomy, there are good reasons to believe that microtargeting with generative AI qualifies as digital manipulation – provided the manipulator has the intention to influence and benefits from it.

2) Now we can turn to the second aspect of manipulation, that is humans are not able to detect

when the content is made by an AI. This is a partially new level of manipulation in social media platforms, but with the advent of generative AI tools the problem has been magnified. Ienca includes bots and fake accounts in the list of aspects and functions of

⁵⁰ M.A. Boden, *L'intelligenza Artificiale* (2018), Il Mulino, Bologna 2019.

social media that raise concerns about manipulation, but doesn't discuss them at length. Similarly, he considers deepfake technology for the disinformation risk but only briefly mentioning of its manipulation potential. My attempt is to provide a more structured argument, suggesting that trying to pass covertly an AI for a human is a form of manipulation for it satisfies the conditions of non-transparency and violation of autonomy. In this context, I am both thinking of AI-generated pictures and AI-powered bots. In both cases the manipulator wants the manipulated to form false beliefs to his advantage.

Starting from the condition of non-transparency according to Ienca it «involves a form of influence that is generally covert and hard-to-detect for the victim» we can say it is met as long as the content or profile is not properly flagged, or when the user fails to understand who or what they are interacting with. The non-transparency corresponds to the fact that the manipulation is hidden.

Personal autonomy is far more complex matter. As said the violation of human autonomy lies in the fact that there is a will to circumvent the manipulated person's ability to form competent and free decisions, or even before that, the ability to make up his own mind. For microtargeting Ienca's argument reasonably relied on cognitive liberty, still, for the peculiar epistemic issue of indistinguishability a suitable route is epistemic agency, as it precisely tackles beliefs formation and the ability to judge content.

Epistemic agency concerns the control exercised over one's beliefs⁵¹ and how these beliefs are formed and revised⁵². If our political beliefs are formed in an environment where we are not sure who we are interacting with and we are not able to say if

⁵¹ M. Schlosser, *Agency*. In *The Stanford encyclopedia of philosophy*, E. Zalta (Ed.), Stanford University 2019. Retrieved 13 Apr 2024, from <https://plato.stanford.edu/entries/agency/>.

⁵² M. Coeckelbergh, *Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence*, cit., p. 1342.

what we see is true our epistemic agency becomes unstable, as Coeckelbergh rightly points out:

The problem is also that citizens can no longer believe their own eyes and hence start doubting and mistrusting not only others but also their epistemic capacities. If AI fakes increasingly more 'believable', then I start questioning my own capacities as an epistemic agent to distinguish truth from falsehood. In addition, there is a basis for this mistrust: if AI fakes the news (or might fake the news: the point is that I never know if AI was used or not), then I have effectively less epistemic agency: I have less control over the formation of my knowledge⁵³.

This relates to the work of Levy on the epistemic environment on how epistemic pollutants such as predatory journals can instill doubts in people not just about the single journal or the individual writers but about the whole scientific community⁵⁴.

Even though Coeckelbergh does not refer to generative AI technologies (and Levy not even to technology altogether), I believe his reasoning can be extended also to those systems.

How epistemic agency and autonomy are connected? Coeckelbergh draws a parallel with nudging. Nudging is the process of changing a choice architecture to incentivize a particular behavior, and even if it may not be considered a proper violation of one's negative liberty, it nonetheless circumvents one's autonomy, hence violating one's positive liberty⁵⁵. He argues that whether intentionally or unintentionally social media dynamics enabled by AI change the epistemic architecture and the epistemic environment of people and influence their belief formation and belief revision. This does not render them unfree when it comes to changing their beliefs. Indeed «Nobody forces them to change or keep their beliefs [...]. However, citizens' belief formation and belief revision processes are manipulated by changing the epistemic architecture and environment in such a way that it

⁵³ *Ibid*, p. 1344.

⁵⁴ N. Levy, *op. cit.*

⁵⁵ We should say more about the relationship between AI recommendation systems and nudging. On the one hand, nudging – even if taken as a paternalistic technique – was born to be used to nudge people toward better outcomes for society. On the other hand, AI recommendation systems seem to lack this "positive attitude" and seem only to push for platforms' interests.

becomes *more difficult* for them to do this»⁵⁶. Coeckelbergh uses an example firstly proposed by Bondy of Claire, a white lady living in a white supremacist environment (online and offline)⁵⁷. She learns that there is no scientific evidence of the superiority of the whites. Will she revise her beliefs? Coeckelbergh notes that she is free to adopt non-racist beliefs, however, her being part of an environment reinforcing her initial belief makes it more difficult for her to revise it. Therefore, he concludes that in this way people's autonomy is diminished, which «in turn diminishes her political agency in a democratic society» (ibidem). Epistemic architecture in this context relates to social media algorithms, but it is not limited to it: the different content one faces is very important as it constitutes the options on which the person makes up her beliefs. Coeckelbergh seems to be primarily interested in the mechanism distributing the options, whereas I am considering the options themselves, but both (make up) the epistemic architecture that can influence "belief formation" and also "belief revision".

There is also another relevant aspect that connects epistemic agency and autonomy. When one starts mistrusting their own epistemic agency because they are unable to tell truth from falsehood, or distinguish a human from a bot, they will probably delegate the activity to someone else, or something else, hence losing control even more. For example, in detecting false content algorithms can be used, but if the user has no control over the mechanism they have lost their autonomy in this regard.

Therefore, I argue that also indistinguishability can effectively diminish one's personal autonomy, satisfying Ienca's last condition.

⁵⁶ M. Coeckelbergh, *Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence*, cit., p. 1348.

⁵⁷ P. Bondy, *Epistemic deontology and strong doxastic voluntarism: a defense*, in «Dialogue», 54, 4, 2015, pp. 747-768.

Having discussed manipulating practices using generative AI (namely microtargeting and indistinguishability) with reference to Ienca's definition of digital manipulation, if finally, time to answer the question "Are we witnessing a quantitative or qualitative change?". I will firstly discuss microtargeting and secondly indistinguishability.

As we have seen, textual Generative AI technologies can enter the process at two stages: the first one is profile-inference based on the content published by the user, the second stage is content generation, where both textual and visual software could be used to generate personalized messages. Content generation now requires a great deal of human work, but with the use of generative AI technologies it can become way more efficient. So, in this sense it would be a quantitative shift. However, compared with non-generative AI the shift qualifies as qualitative, because generative AI unlocks a new kind of micro-targeting, i.e. automated personalized persuasion (PP). If we consider AI-generated images, we might say that micro-targeting can be attained (it is possible to target a specific group of people i.e. black men in the example), but not at scale. In fact, differently from recommendation systems that enable micro-targeting at scale, micro-targeting with generative AI images (or AI-generated text) would regard the content produced, not the way it is distributed inside the platform. Therefore, the microtargeting would be the result of the recommendation algorithms together with the AI-generated content. Nonetheless, the change qualifies as qualitative, as having highly personalized images can be more effective in manipulating the user, it not just a matter of rendering the targeting process more efficient. Still, we would need more research to actually study the impact of highly personalized generated images.

In a similar vein one could ask if generative AI will make manipulation more effective (i.e. it will achieve its goals at

higher rates). Again, unfortunately, no research has been conducted to compare the effectiveness of AI-generated ads compared with the traditional procedure. However, the research targeting personalities using generative AI showed that the effect was significant, meaning the AI-generated content was effective in persuading the subjects of the study⁵⁸.

Now, turning to indistinguishability, the level of resemblance that generative AI tools can achieve represents a qualitative shift and unlocks a new way of conduct digital manipulation on social media by inducing false beliefs and hindering epistemic agency.

On the one hand, with AI-generated images, the manipulation lies not only in the fact that the artificial output can look “human-made”, but also that it can “seem to be true”, hence the false belief can take two different meanings (human-made when it is artificially made, and true when it is fake). On the other hand, it doesn’t make sense to consider bots as true or false, it is only what they convey that can be true or false, therefore the manipulation here lies in the fact that the user is manipulated into thinking the bot is not what it actually is. And this way, the advantage is that people are more easily influenced.

Regarding bots, this shift is important because, as we said in previous sections, they already have a big role in influencing the impact and reputation of a post. Think again about the experiment discussed previously, the bot was successful in building a follow just by parroting human style. Now imagine the same bot but powered with an LLM, that is imagine for example if the posts and the messages with other users are written using ChatGPT or the like. These bots would be able to 1) create high-quality content, and by high quality I mean it resembles human content but also 2) more easily create the content and hence post more. Hence, in an environment/context where individuals are constantly risking

⁵⁸ S. Matz, J. Teeny, S.S. Vaid, H. Peters, G.M. Harari, M. Cerf, *op. cit.*

forming false beliefs like social media, the inclusion of human-like bots complicates the picture and erodes epistemic agency and autonomy even more.

To sum up, the findings are placed in the table below, where it is possible to see a qualitative shift in what concerns microtargeting, especially for text with the possibility of achieving personalized persuasion, and a qualitative shift concerning the indistinguishability aspect, both for images and for text (in this sense when talking about bots enabled by LLMs).

TABLE 4

		Generative AI	
		Images	Text
Manipulation	Microtargeting	Qualitative (Partial)	Qualitative - PP
	Indistinguishability	Qualitative	Qualitative - bots

As I have discussed epistemic agency functions as a filtering mechanism that evaluates the information we encounter, but certain types of content can put pressure on this capacity. The consequences may include a lack of trust in the epistemic environment and even in one's own cognitive abilities. A loss of trust in the epistemic environment is particularly problematic from a democratic perspective, as it distances individuals from the public sphere. Addressing these challenges requires interventions at different levels: political, ethical, cultural, and technological. A detailed analysis is beyond the scope of this article, but I aim to provide a brief overview to suggest possible directions.

From an ethical perspective, it is not necessarily the case that we should adopt a pessimistic view of epistemic agency. On the contrary, this transformation might help sharpen critical thinking and encourage a degree of skepticism. Developing a more critical attitude toward information consumption can strengthen individual

epistemic agency. Similarly, the decision of many users to leave platforms like Facebook and Instagram following controversial changes reflects a form of ethical engagement: rather than passively accepting an epistemically polluted environment, they seek out alternative, higher-quality spaces. Others, while remaining on these platforms, take an active role in clarifying and navigating misinformation, contributing to a healthier digital discourse.

From a technological standpoint, one simple and necessary solution to address the problem of indistinguishability is the use of “flags” or “watermarks,” similar to fact-checking banners. These tools can help users discern the credibility of online content. However, this approach is limited, as it does not fully consider how people interact with technology. Research in Human-Computer Interaction highlights that users’ attitude towards computers has dramatically changed in the span of decades⁵⁹: we are increasingly comfortable engaging with computers, and the same is visible with voice assistants and more recently chatbots. Yet, if AI interfaces become excessively human-like, it could lead users to overestimate their reliability or misunderstand how they actually function. The key challenge is ensuring that design choices promote usability without fostering misplaced trust. Technological solution should therefore include also the design of the technologies.

Beyond individual and technological responses, political and cultural interventions are necessary to reshape the broader epistemic landscape. Regulatory policies can enforce transparency in algorithmic decision-making and platform accountability. Culturally, media literacy and public education initiatives can equip individuals with the tools needed to critically engage with digital information and with new tools, reducing their

⁵⁹ S. Turkle, *Life on the Screen: Identity in the Age of the Internet*, Weidenfeld and Nicolson, New York 1995.

vulnerability to misinformation and misplaced trust, as just mentioned.

5. Conclusion

The aim of this paper was to explore how generative AI tools impact democratic political communication, specifically looking at the phenomena of disinformation, polarization, and manipulation on social media. The main question was to examine whether generative AI technologies really change the picture in comparison to non-generative AI ones (like recommendation algorithms). The analysis revealed that generative AI exacerbates issues of disinformation and partially polarization by enabling the large-scale creation of false information, which indicates a quantitative shift in the scope of these problems. More crucially, the paper argued that generative AI poses a qualitative risk in terms of manipulation, by unlocking new micro-targeting possibilities and producing content nearly indistinguishable from human-made one.

Initially, the paper provided a comprehensive overview of the role of AI in political communication on social media, highlighting how AI recommendation systems contribute to disinformation, polarization, and manipulation. This was followed by an examination of generative AI tools and their potential to intensify these issues. Detailed examples of implementation of the technology were given to illustrate how AI-generated images – with an AI-generated picture of Trump with a group of black supporters – and text – looking at technologies able to create fake news, or to undertake personalized persuasion, or to power bots – can influence public perception and political discourse.

The core of the argument focused on manipulation, discussing the non-transparency of generative AI systems and their potential to diminish personal autonomy. By targeting individuals with personalized, AI-generated content, these technologies can subtly

influence political opinions and behaviors without the user's awareness.

Future research should develop further the analysis on manipulation: both on the role of the epistemic environment for human autonomy, and on a wider account of manipulation, one that looks also the intentional stance reflected by the agent. Future work should also reflect on the combination of disinformation (and polarization) with manipulation, for example, researching if and why disinformation could become manipulating with generative AI (as could be with deep fakes). Finally, it would also be interesting to look at the consequences of a diminished agency on trust. Understanding these aspects would be useful to better mitigate the risks posed by generative AI and ensure the functioning of democratic processes in the digital age.