Topoi (2023) 42:833–842
https://doi.org/10.1007/s11245-023-09940-3

On Artificial Intelligence and Manipulation

Marcello Ienca[1,2]

Accepted: 31 May 2023 / Published online: 20 June 2023

Abstract
The increasing diffusion of novel digital and online sociotechnical systems for arational behavioral influence based on Artificial Intelligence (AI), such as social media, microtargeting advertising, and personalized search algorithms, has brought about new ways of engaging with users, collecting their data and potentially influencing their behavior. However, these technologies and techniques have also raised concerns about the potential for manipulation, as they offer unprecedented capabilities for targeting and influencing individuals on a large scale and in a more subtle, automated and pervasive manner than ever before. This paper, provides a narrative review of the existing literature on manipulation, with a particular focus on the role of AI and associated digital technologies. Furthermore, it outlines an account of manipulation based of four key requirements: intentionality, asymmetry of outcome, non-transparency and violation of autonomy. I argue that while manipulation is not a new phenomenon, the pervasiveness, automaticity, and opacity of certain digital technologies may raise a new type of manipulation, called "digital manipulation". I call "digital manipulation" 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. Drawing on insights from psychology, sociology, and computer science, I identify key factors that can make manipulation more or less effective, and highlight the potential risks and benefits of these technologies for individuals and society. I conclude that manipulation through AI and associated digital technologies is not qualitatively different from manipulation through human–human interaction in the physical world. However, some functional characteristics make it potentially more likely of evading the subject's cognitive defenses. This could increase the probability and severity of manipulation. Furthermore, it could violate some fundamental principles of freedom or entitlement related to a person's brain and mind domain, hence called *neurorights*. To this end, an account of digital manipulation as a violation of the neuroright to cognitive liberty is presented.

Keywords Digital manipulation · AI · Cognitive liberty · Ethics · Influence

1 Introduction

In everyday language, the word *"manipulation"* refers to the act of influencing or controlling someone or something in a skillful or devious way, often with an intent to deceive or gain an advantage. The etymology of this word offers interesting insights into its semantics. The term originates from the Latin word "*manipulare*," which means "to handle, control, or manipulate". *Manipulare* is derived in turn from the Latin word "manipulus", which means "maniple", that is "a handful," "a sheaf," or "a troop". In ancient Rome, a "manipulus" was a military unit of approximately 60–120 soldiers, that was employed in the Roman legions between the Samnite Wars and the Marian reforms, that is in the 3rd–2nd centuries BC (Armstrong 2019). The resulting verb *"manipulare"* was used to describe the actions of a commander who directed or controlled these soldiers.

The modern word "manipulation" first appeared in modern European languages in the mid-18th century. The trajectory of this word from Latin to modern English can be traced to the French word "*manipulation*", which was used in the 17th century to describe the act of handling, crafting or controlling something with one's hands, specifically, a method of digging ore. The corresponding word "manipulation"

✉ Marcello Ienca
marcello.ienca@tum.de

[1] Professorship of Ethics of Artificial Intelligence and Neuroscience, School of Medicine, Technical University of Munich (TUM), Munich, Germany

[2] College of Humanities, Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland

entered the English language in the mid-18th century, when it was used in medical literature to describe various manual techniques for adjusting the bones and soft tissues of the body (Onions et al. 1966). The term was later adopted more broadly to describe any act of skillful or devious control or influence over others, and became a common term in fields such as psychology, politics, and marketing.

Today, the word "manipulation" is used in many languages to describe a variety of behaviors related to influence, control, and deception, reflecting its continued relevance and importance in modern society.

Despite the relatively recent appearance of the term, antecedents of the concept of *"manipulation"* have been discussed by philosophers throughout history, often in the context of moral and political philosophy. In general, manipulation has been understood as a form of influence or control involving deception, coercion or exploitation and is therefore morally despicable.

2 Manipulation in the History of Philosophy

In ancient Greek philosophy, especially in Plato's work, manipulation was often associated with the idea of "sophistry," which referred to the use of clever or deceptive arguments to persuade others. In his dialogue *Sophist*, Plato was critical of the sophists—i.e., itinerant professional teachers and intellectuals in the fifth and fourth centuries BC—and their use of rhetoric to manipulate public opinion. Plato describes sophists as deceptive shadows of the true, in opposition to the true knowledge-seekers, namely the philosophers. As he famously wrote, "the art of contradiction making, descended from an insincere kind of conceited mimicry, of the semblance-making breed, derived from image making, distinguished as portion, not divine but human, of production, that presents, a shadow play of words—such are the blood and the lineage which can, with perfect truth, be assigned to the authentic sophist" (Rosen 1983). While Aristotle shared Plato's critique of Sophistry, his work on rhetoric also provided new conceptual tools for enabling manipulative rhetoric. According to Di Leo (2020), Aristotle actively engaged in teaching manipulative rhetoric and outlined a strategy for regime preservation to his students.

In the Enlightenment era, philosophers such as Immanuel Kant and Jean-Jacques Rousseau were concerned with the use of power and coercion in politics, and the ways in which individuals could be manipulated by those in authority. In his *Critique of Practical Reason* (1788), Kant notoriously argued that individuals should be treated as ends in themselves, and not merely as means to an end. Therefore, if manipulation entails the act of influencing or controlling someone in a devious way with a deceptive intent to gain an advantage, it thereby entails the treatment of the manipulated person as means to an end.

Albeit moving from radically different philosophical premises, Rousseau emphasized in the *The Social Contract* (1762) the importance of popular sovereignty and the need for a social contract to protect individuals from manipulation and exploitation.

In the 20th century, philosophers such as Hannah Arendt and Michel Foucault explored the relationship between power, knowledge, and manipulation with a pronounced political focus. Arendt (1951) argued that totalitarian regimes relied on the manipulation of language and propaganda to maintain control, while Foucault (1961) examined how disciplinary institutions such as prisons and hospitals could be used to manipulate individuals and exert power over them. While different in focus and philosophical underpinnings, their analysis converged in that they both highlighted how manipulation can be utilized as a mechanism of control and exercise of power.

3 Manipulation in Clinical Psychology

The notion of manipulation has evolved significantly with the advent of modern clinical psychology. In particular, the concept of manipulation has been linked to a range of personality disorders, including narcissistic personality disorder, borderline personality disorder, and antisocial personality disorder (Hamilton et al. 1986).

In clinical psychology literature, manipulation typically refers to a set of behaviors that are intended to control or influence others in a way that is selfish, harmful, and often without regard for the other person's well-being. These behaviors may include lying, guilt-tripping, gaslighting, or other forms of emotional manipulation.

In the current psychological literature, manipulation is seen as a significant problem in interpersonal relationships, particularly in cases where one person is attempting to control or exploit another. Manipulation is often associated with power imbalances and can lead to a range of negative outcomes, including decreased self-esteem, social isolation, and anxiety or depression. While manipulative behavior is not a diagnostic criterion for any specific mental disorder, it can be a feature or symptom of some personality disorders and other mental health conditions. In particular, individuals with borderline personality disorder (BPD) or narcissistic personality disorder (NPD) may display manipulative behaviors as a way of controlling their environment and others around them. Other conditions, such as antisocial personality disorder, histrionic personality disorder, and psychopathy, may also involve manipulative behavior as a symptom (Lay 2019). In many of these conditions, manipulation also

appears in the form of interpersonal exploitiveness, also called social manipulation. It is worth noting, however, that not everyone with these conditions will necessarily exhibit manipulative behaviors, and not all individuals who engage in manipulative behavior necessarily have a mental disorder (Hamilton et al. 1986). Manipulative behavior can be learned or situational, and may not be a sign of an underlying psychological condition (Lau 2022).

4 The Moral Significance of Manipulation

In all the works listed above, manipulation is not presented as a morally neutral enterprise. In contrast, it is described as a form of moral failure. In particular, it is a morally deplorable act executed by the manipulator at the expenses of the manipulated. Even in the psychological literature, despite programmatic attempts to take a value-free and non-judgmental stance on psychological traits involved in personality disorders, manipulation is often referred to as a morally despicable act and not merely as a symptom of certain disorders (McHoskey 1995; Poless et al. 2018).

The moral significance of manipulation is extensively discussed in the ethics and political philosophy literature. Across the whole philosophical tradition that goes from Plato to Foucault via Rousseau, Kant and Arendt, there is consensus that whoever (an individual, a government or other institution) engages in manipulating others is necessarily engaging in morally tainted behavior. In this tradition, manipulation is considered ethically wrong because it involves influencing someone's behavior or beliefs in a non-transparent way that (i) undermines their autonomy, freedom, or dignity, (ii) promotes the personal gain of the manipulator at the expense of the manipulated, and (iii) may result in direct or indirect harm for the manipulated.

The moral significance of manipulation has been further explored in the last six decades. In his 1980 book *"Manipulatory Politics"*, Robert Goodin argued that manipulation is inherently deceptive, and offered two criteria for assessing whether an influence is manipulative: "(1). Is the interference deceptive? (2). Is the interference contrary to the putative will of those subject to it?" (Goodin 1980, p. 35). Similarly, T. M. Scanlon morally condemned manipulation as a means of inducing false beliefs and expectations (Scanlon 1998, pp. 298–322), while Coons and Weber characterized manipulation as a subtler and more pervasive form of influence than coercion. In order to differentiate manipulation from benign influence, Noggle associated manipulation with a special form of influence that attempts to get the target to stray from ideals or rational standards of belief, desire and emotion (2018). In a similar fashion, Lau identified intentional concealment as the core feature of manipulative influence: "if a target of influence has no explicit knowledge of the influence or does not fully understand the way in which he or she is affected, the influence is probably (but not necessarily) manipulative" (Lau 2022).

Since the 1980s, political-philosophical discussions of manipulation were extended to the field of medical ethics. In their influential book, *"A History and Theory of Informed Consent"* (1986) Ruth Faden, Tom Beauchamp, and Nancy King argue that manipulation is morally wrong as it undermines the validity of consent. However, they acknowledge that it may not be always easy to determine whether and when a given form of influence is manipulative. In particular, manipulation needs to be distinguished from other forms of behavioral influence that do not share this inherently negative moral status. In fact, as Coons and Weber (2014) observe, people influence one another in all domains of human activity, from political organizations to corporations, friendships, parenting and romantic relationships. Much of this influence, they argue, is benign. As an example, giving advice to friends or serving as role models for our children and students. Therefore, it is important to distinguish manipulation from non-manipulative behavioral influence, in particular influence through persuasion.

It should be noted that the ethical concerns regarding manipulation have co-evolved with the evolution of technology, from older debates about the ethics of advertising to more recent debates about the ethics of artificial intelligence (AI).

For example, the rise of mass media in the early 20th century brought about new ethical concerns about the use of advertising to manipulate consumer behavior. Critics argued that advertising was often deceptive, manipulative, and exploitative, and that it violated the principles of autonomy, truthfulness, and respect for persons. The economist John Kenneth Galbraith famously called advertising "the manipulation of consumer desire" and compared being the target of advertising with being "assailed by demons which instilled in him a passion sometimes for silk shirts, sometimes for kitchenware, sometimes for chamber pots, and sometimes for orange squash" (Galbraith 1958).

These concerns led to the development of ethical guidelines and regulations for advertising, such as the Better Business Bureau's Advertising Code of Ethics and the Federal Trade Commission's Truth in Advertising guidelines. Similarly, the use of propaganda for political and ideological purposes during World War I and II raised new ethical concerns about the use of persuasive techniques to manipulate public opinion. Arendt argued that propaganda was often deceptive, manipulative, and unethical, and that it undermined the principles of democracy, freedom, and respect for persons (Arendt 1951). These concerns led to the development of ethical guidelines and principles for propaganda, such as the

Institute for Propaganda Analysis' 7 Propaganda Techniques and the Universal Declaration of Human Rights.

As we will outline later in the text, the advent of computing technologies, especially those based on Artificial Intelligence (AI), has further ignited ethical debates about the ethics of technology-driven or technology-mediated manipulation.

5 Manipulation vs Persuasion

To draw a distinction line between manipulative vs persuasive behavioral influence, a reference to classical philosophy can be useful. In ancient Greek philosophy, the terms "manipulation" and "persuasion" were often used to describe different forms of influence or control over others. Manipulation was typically understood as the use of clever or deceptive arguments to persuade others, often with the goal of achieving some self-interested end. The sophists were often accused of using manipulation to win debates and gain influence over others (Sidgwick 1872).

In contrast, persuasion (in Greek, Πειθώ "peitho") was seen as a more ethical and legitimate form of influence. Persuasion involves presenting arguments that are truthful and fair, and which aim to convince others through reason rather than deception. According to Aristotle's Rhetoric, persuasion was based on three modes of proof: logos (the use of logical arguments), ethos (the credibility and authority of the speaker), and pathos (the emotional appeal of the argument) (Cope and Sandys, 2010).

While the distinction between manipulation and persuasion was not always clear-cut in ancient Greek philosophy, it reflects a broader concern with the ethics of rhetoric and persuasion. Many philosophers, including Plato and Aristotle, believed that rhetoric and persuasion could be used for good or ill, depending on the intentions and character of the speaker. In general, the use of manipulation was seen as more morally suspect than persuasion, and was often associated with the sophists and their emphasis on winning debates at any cost.

Contemporary philosophers continued to debate the distinction between manipulation and persuasion, with little consensus. Still in 1978, Rudinow wrote (p. 338):

> *"I have encountered no one who has not immediately known what sort of thing I have in mind when I talk about manipulation between persons and who has not also had examples of it ready at hand. Strange, then, to find so little in the way of a systematic account of the concept of interpersonal manipulation, distinguishing it from other means of eliciting behavior; ' for I have not encountered such an account."*

He then conducted a case study analysis, at the end of which he proposes the following operational definition of the manipulative dynamics based on the notion of deception: "A attempts to manipulate S iff A attempts the complex motivation of S's behavior by means of deception or by playing on a supposed weakness of S." (p. 346). Rudinow's work also provides a Kantian account of the morality of manipulation as he writes:

> Finally, we can understand the typical moral reaction to manipulation. To attempt to manipulate someone is to attempt to elicit his behavior without regard for- and with a will to interfere with-his operative goals. Insofar as a person regards the selection of goals as rightfully within his sphere of autonomy and the freedom to pursue his goals as a prima facie right, it is little wonder that he finds attempts to manipulate him objection- able. So it is, in addition, that being manipulated is so frequently assimi-lated to being used, treated instrumentally-7.in the jargon of Kant-as a means rather than an end. (p.347).

Other philosophers shared the view that manipulation involves the use of deceptive tactics to influence someone's beliefs or behavior, whereas persuasion involves a more transparent form of influence that aims to convince someone through reasoned argument and evidence, rather than through deception or coercion.

This view is supported by several contemporary social psychologists, including Robert Cialdini and Cass Sunstein, who have argued that manipulation involves the exploitation of cognitive biases and heuristics to influence people's decisions without their knowledge or consent. In contrast, they suggest that persuasion involves transparent communication and an appeal to the individual's reason and interests. In particular, Cialdini has explored the ways in which individuals can be manipulated through the exploitation of cognitive biases and heuristics. His work has identified a range of tactics used in marketing and advertising, such as social proof, authority, and scarcity, which can be used to influence behavior and decision-making without the individual's knowledge or consent (Cialdini 1984).

However, other philosophers, such as Tamar Gendler, have argued that the distinction between manipulation and persuasion is not always clear-cut. Gendler suggests that some cases of persuasion may involve the use of indirect or subtle forms of influence that can still be seen as manipulative, while other cases of manipulation may involve forms of influence that are more transparent and straightforward (Gendler 2004). However, this criticism neglects that the lack of transparency is not the only determinant of manipulation. In contrast, additional factors seem to

be involved. The following section presents an account of manipulation based on four key features.

6 Key Features of Manipulation

Based on the review above, I propose an account of manipulation based on four key features. I posit that a certain behavioral influence should be regarded as manipulative if it displays the co-occurrence of the following characteristics:

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 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 obliterated.

Based on this account, manipulation differs from other forms of influencing other people's behavior in three fundamental ways.

Firstly, manipulation is associated with behavior that is intended to deceive and gain an advantage, while other forms of influence may be aimed at promoting the common good or achieving a mutualistic goal. While necessary, intentionality alone is not sufficient to identify manipulation. For example, advocacy and negotiation are involve some degree of intentional influence, but are typically seen as legitimate and necessary components of democratic decision-making.

Secondly, manipulation involves the use of deceptive or coercive tactics to control or influence someone's beliefs or behavior, often with the intention of benefiting oneself or others at the expense of the person being manipulated. This can include tactics such as lying, emotional manipulation, or threats, and often involves an imbalance of power or information between the manipulator and the person being manipulated. In contrast, other forms of influence, such as persuasion, involve presenting arguments that are truthful and fair, and which aim to convince others through reason rather than manipulation. Although the truthfulness and fairness of arguments, can be confounded using rhetorical strategies and informal logical fallacies, in a persuasive dynamic the agent retains the ability to inspect those arguments and scrutinize their empirical correctness and logical soundness. This ability can be strengthened through educational activities such as training in logic and science. Therefore, persuasion typically involves a more equal and reciprocal relationship between the persuader and the person being persuaded. Even when persuasion does not rely on a shared commitment to truth-seeking and open communication, the possibility of truth-seeking is not obliterated.

Finally, manipulation is characterized by a lack of respect for the other person's autonomy and agency, and can be seen as a violation of their basic human rights. In contrast, other forms of influence may respect and uphold the other person's autonomy, allowing them to make their own decisions based on the information and arguments presented to them.

While the first requirement is quite semantically unambiguous, the latter two require specification. In fact, autonomy is believed to be violated only when deception is at stake. However, what constitutes a "deceptive" vs "transparent" or "implicit" vs "explicit" influence depends on many factors including design features, contextual factors, and the cognitive abilities of the influenced individual. Therefore, a narrower characterization of deceptiveness and non-transparency is needed. Furthermore, it should be highlighted that violations of personal autonomy are not unique to manipulative behavior but characterize first and foremost coercion. Therefore, it should be clarified what form of violation of personal autonomy is peculiar to manipulation in contrast to coercion.

In the following, I will argue that the notion of cognitive liberty provides useful conceptual ground to specify these two requirements.

7 Manipulation as a Violation of Cognitive Liberty

The concept of cognitive liberty is a relatively new area of discussion in philosophy and ethics that deals with the freedom to control one's own consciousness and mental processes. According to this concept, individuals have the right to access and control their own thoughts, emotions, and perceptions, free from interference or coercion by others. Proponents of cognitive liberty such as Farahany (2019, 2023), Bublitz and Merkel (2014), Ienca & Andorno (2017), Ienca and Vayena (2021), have argued that cognitive liberty is an essential component of individual autonomy and human dignity, and a necessary prerequisite of several other freedoms such as freedom of speech and freedom of expression.

In particular, these authors have suggested that advances in neuroscience and related technologies have the potential to greatly enhance cognitive liberty, by allowing individuals to gain greater control over their own mental processes and to overcome cognitive barriers such as mental illness or cognitive disabilities. However, they have also noted that these same technologies could be used to manipulate or coerce individuals, potentially violating their cognitive liberty and autonomy. More recent work as highlighted that the notion of cognitive liberty is a useful tool to scrutinize the moral valence not only of technologies that intervene directly into the brain (so-called *neurotechnologies*) but also of technologies that have no direct access to the subject's brain but enable interactions with the subject's sensory and behavioral abilities (Ienca and Malgieri 2022)

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 own consciousness, and may undermine their sense of autonomy and dignity. This form of violation of personal autonomy is more subtle 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-determination.

Through these lenses not all forms of influence are necessarily violations of cognitive liberty. Persuasion that is open, honest, and based on transparent communication may be consistent with cognitive liberty, as it allows individuals to make informed decisions based on accurate information and reasoning. Therefore, some amount of moral valence seems to be attributable to the extent to which a certain behavioral influence engages with or evades from an individual's capacity to reason.

A useful conceptual tool to determine whether and when a certain form of behavioral influence is evading from an individual's capacity to reason, and thereby constitutes a violation of cognitive liberty (and in turn a manipulation), is Douglas' notion of "arational influence". According to Douglas (2018), arational influence is a type of influence that does not rely on explicit or rational arguments, but rather on more subtle and indirect means of persuasion. It contrasts with "rational influence", that is influence based on explicit and rational arguments.

According to this view, non-transparence influence is an influence that can "bypass reason". Consequently, this would construe manipulation as an influence that is intentionally designed to be covert in a way that allows it to bypass reason and to produce an asymmetry of outcome between the manipulator and the manipulated. As such, manipulation interferes with a core domain of personal autonomy, that is cognitive liberty.

This definition is consistent with another definition put forward by Daniel Susser, Beate Roessler, and Helen Nissenbaum, who wrote that: "manipulation is hidden influence.… Covertly influencing someone—imposing a hidden influence—means influencing them in a way they aren't consciously aware of, and in a way they couldn't easily become aware of" (Susser et al. 2019, p. 4).

8 Manipulation and AI

With this definition in mind, let us explore the impact of digital technologies on the notion of manipulation. The rise of digital media in the late 20th century and early 21st century brought about new ethical concerns about the use of online advertising, social media, and other digital platforms to manipulate consumer behavior and public opinion. This section will provide an overview of the use of digital technologies, especially artificial intelligence (AI), for exerting manipulative effects. Furthermore, it will assess the ethical status of various AI-driven activities based on the working definition above.

In the light of the definition above, among the sociotechnical trends based on digital technologies that have raised concerns about the risk of manipulation, the following require special attention: social media platforms, microtargeting advertising, personalized search algorithms and deepfake technology.

Social media refer to web-based platforms that allow individuals and organizations to create, share, and exchange user-generated content, such as text, images, videos, and links. These platforms provide a variety of features, such as profile creation, content publishing, sharing, commenting, and reacting, and are designed to facilitate communication, socialization, and networking among users. Social media have become ubiquitous in contemporary society and are used for various purposes, such as entertainment, information dissemination, political engagement, and business promotion. The use of social media has also raised important social, cultural, and ethical issues, such as privacy, cyberbullying, filter bubbles, and the spread of misinformation. In particular, social media platforms have been accused of using algorithms to manipulate what content users see in their feeds, in order to promote certain political or commercial interests (Ienca and Vayena 2018).

There are several aspects and functions of social media platforms that raise concerns about manipulation, including:

- *Filter bubbles*: Social media platforms use complex AI algorithms to determine the content that users see, and

advertisers can use these algorithms to target specific groups of users with tailored advertising. This can create filter bubbles, where users are only exposed to information and ideas that align with their existing beliefs and values.

- *Fake accounts and bots*: Social media platforms can be manipulated through the creation of fake accounts and AI-powered bots, which can be used to spread misinformation, amplify certain voices, or create the illusion of widespread support for a particular idea or movement.
- *Content moderation*: Social media platforms often rely on automated or human content moderation to identify and remove harmful or inappropriate content. However, the effectiveness of these moderation systems is often limited, and there is a risk that they can be manipulated or abused to censor certain voices or ideas.
- *Amplification of extreme or sensational content*: Social media platforms tend to amplify content that is extreme or sensational, as this is often the type of content that generates the most engagement and attention. This can create an environment where misinformation, conspiracy theories, and other harmful content spread rapidly.
- *Addictive potential*: Functions of social media such as endless scrolling, notifications, and autoplay plugins have been found to have addictive and manipulative potential on users (Sun & Zhang 2021; Hou et al. 2019). These functions are designed to keep users engaged on the platform for longer periods of time, increasing the likelihood that they will be exposed to more advertisements, generate more data, and ultimately increase the platform's profits. Endless scrolling refers to the practice of automatically loading new content as a user scrolls down a page, allowing them to endlessly consume content without having to actively search for it. This feature can be addictive as it creates a sense of "infinite" content, which can be difficult for users to pull away from. Notifications are another feature that can be addictive, as they are designed to grab a user's attention and encourage them to check the platform. Notifications are often personalized and designed to trigger a specific emotional response, such as fear of missing out or excitement about a new message or like. Autoplay plugins, such as autoplay videos or suggested content, are designed to automatically play new content after the user has finished watching or reading the current content. This feature can be addictive as it creates a sense of continuity and encourages users to stay on the platform for longer periods of time.

Microtargeting advertising is 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. The aim of microtargeting is 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 network.

Microtargeting can be used for a wide range of purposes, including political campaigning, product promotion, and social advocacy. While microtargeting has the potential to be highly effective, it has also been criticized for its manipulative potential. 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 (Wilson 2017).

Personalized search algorithms are computer algorithms (often based on machine learning) used by search engines, such as Google, to tailor search results to individual users based on their previous search history, browsing behavior, and other personal data. These algorithms analyze large amounts of data about each user, such as their location, search history, click-through rate, and other behavioral signals, in order to predict what types of search results they are likely to find most relevant and useful.

The aim of personalized search algorithms is to provide a more efficient and effective search experience for users, by prioritizing results that are more likely to be relevant to their interests and needs. However, personalized search algorithms have been criticized for their potential to reinforce existing biases, limit diversity of information, and create filter bubbles.

One of the challenges with personalized search algorithms is that they rely on large amounts of data about each user in order to function effectively. This data can be used to build a detailed profile of each user, including their interests, preferences, and behavior patterns. While this can be helpful for providing more relevant search results, it can also be used to target users with advertising or other forms of manipulation.

Finally, deepfake technology is a type of AI that is used to create realistic fake images, videos, and audio recordings that appear to be genuine (Westerlund 2019; Yu et al. 2021). Deepfake technology works by using machine learning algorithms to analyze and manipulate data, such as photographs, videos, and audio recordings, in order to create new, synthetic content.

The process of creating a deepfake typically involves training a machine learning algorithm on a large dataset of real images or videos, and then using that algorithm to generate new images or videos that have been modified in some way. For example, a deepfake algorithm might be used to

manipulate a video of a person's face in order to create a new video that shows them saying or doing something that they did not actually say or do.

Deepfake technology has been used for a variety of purposes, both benign and malicious. Some of the more benign applications of deepfake technology include creating digital avatars for video games or virtual assistants, while some of the more malicious uses include creating fake news stories, manipulating public opinion, and committing financial fraud or other types of cybercrime.

The potential harms of deepfake technology include the spread of misinformation, the creation of fake evidence for legal or political purposes, and the invasion of privacy. Additionally, deepfake technology has the potential to exacerbate existing social and political divisions by creating false narratives or spreading propaganda as it occurred during the Russian invasion of Ukraine. As a result, there is growing concern about the potential misuse of deepfake technology and calls for greater regulation and oversight to prevent its malicious use.

The common denominator of all these approaches is the utilization of machine learning to extract relevant information about target users with the intention of subsequently influencing their behavior in a manner that bypasses their rational defenses and their ability to reject that influence (see Fig. 1).

All the functions described above create a feedback loop that encourages users to stay on the platforms, generating more engagement, data, and revenue for the platforms. This can also be manipulative, as users may not be aware of how much time they are spending on the platform, or may feel compelled to keep using the platform even when it is not in their best interests. Additionally, the constant exposure to personalized and targeted content can reinforce existing biases and create filter bubbles, furthering social and political polarization. These risks are amplified by the significant level of opacity of AI due to the lack of transparency and explainability of most algorithms. Furthermore, it is intensified whenever consumers lack technical literacy on AI's shortcomings.

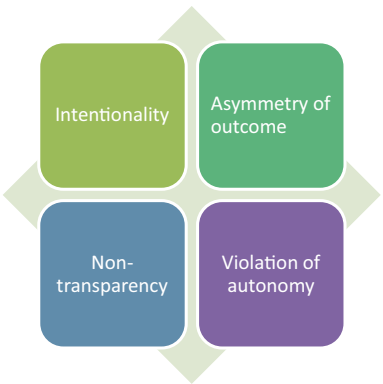

Fig. 1 Key features of manipulation

I call "digital manipulation" 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.

The effectiveness of digital manipulation can be influenced by a variety of factors, including:

- Personalization: Personalized content and advertising can be more effective at manipulating individuals, as it is more likely to align with their existing beliefs and values.
- Emotional appeal: Content that triggers strong emotional responses, such as fear, anger, or excitement, can be more effective at manipulating individuals, as it can bypass their critical thinking and reasoning abilities.
- Social influence: Content that appears to be popular or endorsed by others on digital platforms can be more effective at manipulating individuals, as it can create a sense of social proof or social pressure.
- Repetition: Repeated exposure to a message or piece of content can be more effective at manipulating individuals, as it can create a sense of familiarity or make the message more memorable.
- Trustworthiness: Manipulation is less effective when the source of the content or message is perceived to be untrustworthy or unreliable.
- User awareness: Individuals who are aware of the potential for manipulation and are able to identify and resist it are less likely to be manipulated.
- Time constraints: Manipulation can be more effective when individuals are pressed for time or lack the cognitive resources to fully process the information presented to them.

These factors highlight the importance of ethical considerations and responsible use of digital technologies. It is important for individuals, organizations, and policymakers to be aware of the potential for manipulation and to take steps to mitigate its effects. This may include measures such as improved digital literacy and education, increased transparency and accountability for digital technologies, and greater user control over their personal data and the content they are exposed to.

9 How to Mitigate Digital Manipulation

To mitigate the risk of digital manipulation, a range of strategies can be employed, including technical, regulatory, and ethical approaches. One focal area is certainly user education and digital literacy. Educating users about the risks and

harms associated with digital manipulation can help them to identify and resist manipulation attempts. This could include providing information on common manipulative tactics, such as fake news and phishing scams, and promoting critical thinking skills. On the technology side, a key factor is transparency. Increasing transparency around the use of digital technologies, especially concerning algorithms and data processing practices, can help users understand how their data is being used and how decisions are being made. This could include providing clear disclosures and explanations of how data is collected, used, and shared. Ceteris paribus, explainable AI approaches are less manipulative than opaque black box approaches. A third factor is regulatory oversight. Governments and regulatory bodies can play a role in mitigating the risk of digital manipulation by enforcing laws and regulations that protect user privacy, prevent misinformation and disinformation, and promote transparency and accountability. An important step in this direction is the EU AI Act (2021). As highlighted in a recent report of the The Future of Life Institute (2022), the Act directs its attention to manipulation in two main ways:

(A) By identifying the practice, target population and harm. In fact, the Act "covers practices that have a significant potential to manipulate persons through subliminal techniques beyond their consciousness or exploit vulnerabilities of specific vulnerable groups such as children or persons with disabilities in order to materially distort their behavior in a manner that is likely to cause them or another person psychological or physical harm."
(B) By acknowledging that other regulations might cover manipulation. In fact, the act states that "other manipulative or exploitative practices affecting adults that might be facilitated by AI systems could be covered by the existing data protection, consumer protection and digital service legislation that guarantee that natural persons are properly informed and have free choice not to be subject to profiling or other practices that might affect their behavior".

In conjunction with regulatory interventions, ethical design may play a key role. Digital technologies can be designed in an ethical and responsible manner, with a focus on reducing the potential for manipulation. This could involve designing algorithms and user interfaces that prioritize user cognitive liberty, privacy, provide clear explanations of how decisions are made, and avoid the use of manipulative tactics such as dark patterns. The combination of regulatory intervention and ethical design can provide users with greater control over their personal data and the content they are exposed to. This can help to preserve their personal autonomy (especially their cognitive liberty) and reduce the potential for manipulation. This could involve providing tools and features that allow users to control the content they see, and replacing opt-out models with affirmative ("opt-in") consent. Addressing the risk of digital manipulation will require collaboration between multiple stakeholders, including tech companies, governments, civil society organizations, and academic institutions. By working together, these stakeholders can develop and implement strategies to promote a safer and more trustworthy digital environment. Finally, policy proposals aimed at enshrining a neuroright to cognitive liberty and mental integrity (Ienca & Andorno, 2017; Ienca, 2021) may provide the necessary normative framework for targeted legal interventions that minimize the right to privacy. Overall, addressing the risk of digital manipulation will require a multi-pronged approach that involves technical, regulatory, and ethical strategies. Eliminating digital manipulation is a practically unattainable aim. The reason for that stems from the fact that manipulation is a phenomenon that largely outlasts digital technology, and several forms of manipulation have been occurring in the pre-digital world. Furthermore, since a digital environment is always a pre-designed environment, it will necessarily attribute to the designers the ability and power to influence the users (Ienca and Vayena 2021). However, prioritizing user autonomy, privacy, transparency, and control, and promoting ethical and responsible design practices, it may be possible to mitigate the risk of digital manipulation in social media and AI technology.

Funding Open Access funding enabled and organized by Projekt DEAL. This work has been supported by the ERA-NET NEURON project HYBRIDMIND (Swiss National Science Foundation 32NE30_199436).

Declarations

Conflict of interest The author has no conflict of interest to declare.

References

Arendt H (1951) The origins of totalitarianism. Harcourt Brace Inc, New York

Armstrong J (2019) Organized chaos: manipuli, socii, and the Roman army c. 300. Romans at war. Routledge, Oxford, pp 76–98
Bublitz JC, Merkel R (2014) Crimes against minds: On mental manipulations, harms and a human right to mental self-determination. Crim Law Philos 8(1):51–77
Cialdini (1984) Influence: science and practice. University of Arizona Press, Tucson
Coons C, Weber M (eds) (2014) Manipulation: theory and practice. Oxford University Press, Oxford
Cope EM, Sandys JE (eds) (2010) Aristotle: rhetoric, vol 2. Cambridge University Press, Cambridge
DiLeo D (2020) Aristotle's manipulative maxims. Rev Pol 82(3):371–392
Douglas T (2018) Modulation of motivation. Treatment for crime: philosophical essays on neurointerventions in criminal justice. Oxford University Press, Oxford
Farahany NA (2019) The costs of changing our minds. Emory LJ 69:75
Farahany NA (2023) The battle for your brain: defending the right to think freely in the age of neurotechnology. St. Martin's Press, New York
Foucault M (1961) Folie et Déraison: Histoire de la folie à l'âge classique. Librairie Plon, Paris
Galbraith JK (1958) The affluent society. Houghton Mifflin, New York
Gendler TS (2004) Thought experiments rethought—and reperceived. Phil Sci 71(5):1152–1163
Gendler TS (2008) Alief and belief. J Philos 105(10):634–663
Goodin RE (1980) Manipulatory Politics, New Haven: Yale University Press
Hamilton JD, Decker N, Rumbaut RD (1986) The manipulative patient. Am J Psychother 40(2):189–200
Hou Y, Xiong D, Jiang T, Song L, Wang Q (2019) Social media addiction: its impact, mediation, and intervention. Cyberpsychology: J Psychosoc Res Cyberspace. https://doi.org/10.5817/CP2019-1-4
Ienca M (2021) On neurorights. Front Human Neurosci 15:701258
Ienca M, Andorno R (2017) Towards new human rights in the age of neuroscience and neurotechnology. Life Sci Soc Policy 13(1):1–27
Ienca M, Malgieri G (2022) Mental data protection and the GDPR. J Law Biosci 9(1):Isac006
Ienca M, Vayena E (2018) Cambridge analytica and online manipulation, vol 30. Scientific American, New York
Ienca M, Vayena E (2021) Digital nudging exploring the ethical boundaries. In: Veliz (ed) The oxford handbook of digital ethics. Oxford University Press, Oxford
Kant I (1788) Critique of practical reason.
Lau S (2022) The good, the bad, and the tradecraft: HUMINT and the ethics of psychological manipulation. Intell Natl Secur 38(4):592–610
Lay G (2019) Understanding relational dysfunction in borderline, narcissistic, and antisocial personality disorders: clinical considerations, presentation of three case studies, and implications for therapeutic intervention. Psychol Res 9(8):303–318
McHoskey J (1995) Narcissism and machiavellianism. Psychol Rep 77(3):755–759
Onions CT, Friedrichsen GWS, Burchfield RW (eds) (1966) The Oxford dictionary of English etymology (Vol. 178). Clarendon Press, Oxford
Poless PG, Torstveit L, Lugo RG, Andreassen M, Sütterlin S (2018) Guilt and proneness to shame: unethical behaviour in vulnerable and grandiose narcissism. Eur J Psychol 14(1):28
Rosen S (1983) Plato's sophist: the drama of original and image. Yale University Press, New Haven, p 12
Rousseau JJ (2018) The social contract. Reprinted as Rousseau: the social contract and other later political writings. Cambridge University Press, Cambridge
Rudinow J (1978) Manipulation. Ethics 88(4):338–347. https://doi.org/10.1086/292086
Scanlon TM (1998) What We Owe to Each Other. Cambridge, MA: Harvard University/Belknap Press
Sidgwick H (1872) The sophists. J Philol 4(7):288
Sun Y, Zhang Y (2021) A review of theories and models applied in studies of social media addiction and implications for future research. Addict Behav 114:106699
Susser D, Beate R, Helen N (2019) Technology, Autonomy, and Manipulation. Internet Policy Rev 5:35
Westerlund M (2019) The emergence of deepfake technology: a review. Technol Innov Manag Rev 9(11):39
Wilson DG (2017) The ethics of automated behavioral microtargeting. Ai Matters 3(3):56–64
Yu P, Xia Z, Fei J, Lu Y (2021) A survey on deepfake video detection. Iet Biometrics 10(6):607–624