International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 4
Article Received: *25 January 2022* ***Revised:*** *12 February 2022* ***Accepted:*** *30 March 2022*

Design-Centric Research on the Ethics of Generative AI and Manipulation

[1] Mrs. S. Menaka, [2] Dr. Chaitanya Krishnakumar, [3] Mrs. C. Meera Bai, [4] Mr. T. Thiyagarajan
Department of Computer Applications, Nehru Institute of Information Technology and Management, Coimbatore, Tamilnadu, India.

Abstract: Generative AI allows for large-scale, automated manipulation with significant efficacy. Although there is increasing ethical discourse on generative AI, the specific risks related to manipulation are not yet thoroughly explored. This article addresses crucial questions related to the conceptual, empirical, and design aspects of manipulation, which are essential for understanding and mitigating these risks. By focusing on these inquiries, the article emphasizes the importance of accurately defining manipulation to promote the responsible advancement of Generative AI technologies.

Keywords: *Generative AI, Large Language Models (LLMs) ,Manipulation ,Value sensitive design ,AI ethics ,Persuasion ,Deception*

Introduction

The field of generative AI is experiencing rapid growth, with impressive outcomes from recent implementations (refer to Floridi, 2023, however). According to The Economist (2023), these advancements provide "enormous promise and peril," particularly because they make it possible to exert large-scale, automated influence.

One the one hand, this capacity is encouraging because effective influence is necessary for many positive things to happen. For instance, better lifestyle interventions are necessary to enhance health outcomes, and this requires effective influence (see, for example, Tremblay et al., 2010). Additionally, it might enhance public policy by assisting governments in reaching out to the populace in the midst of misinformation, filter bubbles, and fake news (European Commission, upcoming).

Effective influence, on the other hand, encourages manipulation, a dubious moral practice. For example, generative AI might learn to produce outputs that successfully take advantage of users' cognitive biases to influence their conduct (Kenton et al., 2021) or "make email scams more effective by generating personalised and compelling text at scale" (Weidinger et al., 2022). More generally, there is a strong incentive to switch from legitimate forms of influence like rational persuasion to more effective but morally dubious forms of influence like manipulation whenever effective influence is rewarded, which is the case in almost any area of human interaction, such as social life, marketing, or politics. Therefore, generative AI "aggravates" current ethical worries regarding online manipulation (Klenk & Jongepier, 2022b).

But there's no consensus on how the (dis-)value of manipulation should be considered when creating new generative AI-based technology. Stated differently, how can generative AI—or, more accurately, the applications that employ it—be created in a way that prevents improper types of manipulation? The majority of the current research in AI ethics concentrates on the crucial but still early stage of highlighting relevant ethical problems, with very little attention paid to design-related issues (e.g. Weidinger et al., 2022). Furthermore, generative AI applications can already be made "helpful, honest, and harmless" in general according to certain technical work on AI alignment (Askell et al., 2021).

However, given that manipulation is a challenging notion to understand, it is not unexpected that there is not enough emphasis given to a suitable conceptualisation of manipulation that can guide design. One notable omission in the discussion of generative AI is the lack of focus on manipulation. The EU's upcoming AI Act, for example, specifically names manipulation as a disvalue and makes it the explicit goal of AI regulation (European Commission, 2021; European Commission et al., 2022). In general, manipulation is viewed as a danger to democracy and reliability, which implies that it poses a serious risk to the main objective of responsible, reliable AI (Faraoni, 2023). Furthermore, a substantial amount of research highlights concerns with manipulation in various other situations, such as advertising and nudging (see Sunstein, 2016).

Hence, the focus of this article is to outline a research plan for investigating manipulation in generative AI. I maintain that conducting comprehensive research on manipulation and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 4
***Article Received:** 25 January 2022 **Revised:** 12 February 2022 **Accepted:** 30 March 2022*

generative AI, which is a concern for everyone interested in reliable AI and the preservation of democratic values, is heavily reliant on how we define manipulation. This is crucial because our understanding of manipulation will bring different aspects to light. It also holds practical significance because varying definitions of manipulation will lead to different implications for design and regulation.

My approach is as follows. The part titled "Design for values and conceptual engineering" provides an introduction to the general design for value approach. Following that, the section "Design for non-manipulation" covers relevant research inquiries about manipulation that pertain to the conceptual, empirical, and implementation stages of a design for value project, with an emphasis on the conceptual stage.

Design for values and conceptual Engineering

The design of new technologies is founded on the principles set by engineers (van de Poel, 2020; van den Hoven et al., 2015). Emphasized by the IEEE, the WHO, UNESCO, the EU, and other organizations, the design perspective stresses the significance of human values influencing and shaping appropriate design requirements. As a result, several critical questions for any value-based design project revolve around determining the nature of the values that should be integrated into the design.

The concept of design for values revolves around the idea that the desired values can be defined in a manner that allows for a systematic and dependable derivation of specific design requirements from a broad, abstract understanding of target values like 'trust,' 'democracy,' or 'non-manipulation' (van de Poel, 2013, 2020; Veluwenkamp & van den Hoven, 2023). It is widely recognized that there are often multiple, initially plausible conceptualizations of target values, and considerable attention has been focused on different methods of determining a value conceptualization (cf. Friedman & Hendry, 2019).

The question of how to judge between different, potentially conflicting conceptualizations of a specific value has recently gained attention (Himmelreich & Köhler, 2022; Veluwenkamp & van den Hoven, 2023). According to Veluwenkamp and van den Hoven (2023, p. 2), "it is not always clear which concepts used in breaking down requirements are the most suitable in the relevant context of use." When addressing how to determine which conceptualizations to utilize, we must recognize that conceptualizations have implications and carry significant weight. One reason why various conceptualizations are important for our comprehension is that they will highlight various phenomena. For instance, becoming rather than viewing manipulation as a form of social pressure that doesn't require any concealment from the user, researchers and designers will be prompted to consider entirely different phenomena when they view manipulation as an influence that is hidden from the user. Different interpretations of manipulation are thus like searchlights. When they are implemented for a specific goal value, they highlight some phenomena while blocking out others that might be just as significant or even more so (see also Barnhill, 2022).

Therefore, it is important for high-quality research on generative AI and manipulation that the conceptualization chosen captures or reflects the phenomena that initially raises concerns about manipulation. Moreover, conceptualizations have an impact on the actual technological innovations and interventions created to address the design challenge. For instance, conceptualizing trust in terms of morality, such as benevolence, will lead to very different technical solutions toward the goal of trustworthy AI than thinking of trust as epistemic reliability (cf. Veluwenkamp & van den Hoven, 2023).

Therefore, choosing a conceptualization is not "just about words." It's a significant, tangible decision. When we design for non-manipulation, we probably end up with two distinct technical artifacts or systems based on our initial conceptions of manipulation. Moreover, the design challenge addresses a fictitious problem if our conceptualization is flawed or inappropriate. Thus, our success in designing for values hinges on the types of conceptualizations we choose. Thus, a proper conceptualization of "manipulation" is essential for conducting high-quality research on manipulation and generative AI 5. In that regard, current talks about manipulation and generative AI fall short. Weidinger et al.'staxonomyof generative AI risks is the focus of their study from 2022. They don't do a good job of differentiating manipulation from deception when they talk about it. They don't address the questions raised by this omission. Is non-deception design the same as design against manipulation? Or is there something else? In the event that there is more, how would that conception appear? More in-depth discussion and a comprehensive conceptualization of manipulation are provided by Kenton et al. (2021). They argue from a safety standpoint that the more phenomena covered, the safer the final design.

However, they admit that their conceptualization might be "too wide-ranging" (Kenton et al., 2021, p. 11). An excessive number of phenomena will be perceived as examples of manipulation, distorting our understanding of what manipulation actually entails, and designs aimed at the phenomena may become overly preoccupied with requirements. In the future, studies on manipulation in generative artificial intelligence should concentrate on

developing more accurate and relevant theories of the target phenomenon.

The fundamental yet evident question is what standards should be used when selecting a conceptualization. What distinguishes one interpretation of "manipulation," for instance, from another? Conventionally, conceptualizations make sense when they align with the intended phenomenon. According to that perspective, a conceptualization of manipulation is appropriate if it encompasses all instances of manipulation.

This shall serve as the specific appropriateness criterion.Six Crucially, a conceptualization of manipulation that meets the narrow criteria is appropriate regardless of how well it "works" in actual applications, like policy or design work. The narrow criterion primarily seeks to clarify the components of an idea and pays little to no attention to whether conceptualization is useful for design projects. However, there may also be pragmatic and moral considerations that legitimately influence our choice of conceptualization, according to the recent debate on "conceptual engineering" in philosophy and the ethics of technology. Some suggestions have been made regarding how to systematically evaluate these considerations (cf. Veluwenkamp & van den Hoven, 2023).

According to this viewpoint, in addition to evaluating whether the conceptualization satisfies the narrow criterion by capturing all instances of the target phenomenon, moral and pragmatic considerations regarding the causal effects of applying a specific conceptualization or its practicality also influence the decision of whether it is an appropriate conceptualization. This should serve as the general standard for appropriateness when choosing a conceptualization. Since that design decisions should ultimately be informed by a conceptualization of manipulation in the context of generative AI, the broad appropriateness criterion may be particularly pertinent from a design standpoint. But the question of how much broad considerations should take precedence over narrow considerations is a difficult and open metaphilosophical one.

My purpose in writing this is not to address the metaphilosophical debate over whether or not we should favor the strict or loose appropriateness criteria.7. Instead, I will highlight the unanswered issues that still prevent me from supporting either strategy in the paragraphs that follow: What exactly is manipulation (as understood in folklore), and how should it be used, given our willingness to stray from it in the interest of truth or for other practical and ethical reasons?

Design for non-manipulation

According to Buijsman et al., forthcoming; Friedman & Hendry, 2019, design for value approaches generally involve the following stages: a phase wherein the appropriate conceptualization of a value using conceptual means (e.g., reasoning), an empirical stage wherein stakeholder input is solicited to contribute to the conceptualization, and a design or implementation stage. Conceptual, empirical, and design are the three phases of a design for value project that should be repeated at various points in the specification of the target value (i.e., from value identification to conceptualization, association with norms, etc.), until specific design requirements are met (cf. Veluwenkamp & van den Hoven, 2023). I limit my attention mostly to the conceptualization phase.

When disagreements regarding proper conceptualization are settled, we should anticipate that the discussion will shift to the operationalizing process that follows, leading to specific design specifications. I concentrate on non-manipulation, or the absence of manipulation, as a target value since manipulation is typically viewed as a dis-value. It follows that a non-manipulative design that succeeds will likely overlook a great deal of other ethically important issues. From the standpoint of manipulation, a generative AI application that does not manipulate might be morally acceptable, but it might still have other ethical problems (such explainability, privacy, etc.) in general.

Therefore, it might be necessary to integrate design for non-manipulation into more general design goals, like design for democracy or reliable AI (EGE, 2023).

Conceptual stage

In order to develop for non-manipulative generative AI, the following inquiries must be addressed at minimum:
1. What are the trustworthy standards to recognize manipulation and set it apart from other (typically less dubious) forms of influence?
2. How can applications of generative AI be in line with non-manipulation criteria?
3. When and why is it morally wrong to manipulate?
The first query is fundamentally related to understanding manipulation in the right way. By responding to it, we will be able to determine whether a particular influence—like the output of a generative AI application—is being manipulated. Let's say that a generative AI-powered personal digital health assistant uses the user's recent purchase history to generate the message, "You should be ashamed of yourself for ordering that meal." We need trustworthy criteria to recognize manipulation in order to determine whether that prompt or any other output produced by the system qualifies

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 4
***Article Received:** 25 January 2022 **Revised:** 12 February 2022 **Accepted:** 30 March 2022*

as manipulation. I'll quickly go over the most important manipulation criteria in this section. I shall propose—in Sect. "The indifference criterion"—that the indifference criterion is the most suitable to conceptualize manipulation after examining and discarding a number of alternative criteria.

The continuum model of influence

One type of influence is manipulation (Coons & Weber, 2014b). Humans are social animals that have a wide range of effects on one another. A speech act is an example of an intentional influence. Inadvertent influences include the intimidating effect of a very tall person on others. Not all deliberate influences, though, raise moral questions. For instance, you are not breaking the law if you shout for the driver to stop in the event of an accident while you are a passenger in a car (see Sunstein, 2016). Because manipulation is a morally dubious form of influence, the first question asks us to distinguish it from other forms of influence that are typically accepted as legitimate.

One can derive criteria for recognizing the manipulation that is implied by a particular conceptualization by comparing it to alternative forms of influence. Indeed, some scholars have proposed that manipulation lies somewhere along a continuum of influence, in between coercion and reasonable persuasion (Beauchamp, 1984; Beauchamp & Childress, 2019). The concept that there are some benign forms of influence, like reasoned persuasion, and other forms of influence that are obviously problematic, like coercion, is conceptualized and helped to be drawn upon by this continuum model.

But as of yet, the continuum model is unable to give us trustworthy manipulation criteria. It appears that there are non-coercive and non-persuasive influence techniques that are not manipulation (Noggle, 1996). For instance, dressing professionally for a job interview does not appear to be manipulation, nor is it a form of coercion or rational persuasion (Noggle, 1996). Depending on how we define the terms "persuasion" and "coercion," the continuum model may provide us with far too broad manipulation criteria, leading to unduly strict design specifications for generative artificial intelligence. Consequently, turning to philosophical theories of manipulation that provide more precise criteria for recognizing manipulation is more promising.

A number of widely accepted concepts in manipulation are clear-cut, instinctive, and appear to be simple to put into reality. The criteria for hidden influence The belief that manipulation is inherently a type of covert influence is arguably the most pervasive (see Faraoni, 2023, and its adoption and consideration in policy documents). Manipulation, according to Susser et al. (2019a, 2019b), is an influence that the victim is not aware of or could not readily recognize. It is essential to define precisely what is kept secret from the manipulation victim for this idea to be applicable in generative artificial intelligence.

Is it necessary, for instance, to conceal from the user the influence's intended result? or the specific psychological process that the influence is meant to operate through? or how the power was produced? According to the latter, for instance, any influence produced by generative AI that isn't identified as such would be considered manipulative under the concept of hidden influence. In any case, on the continuum model, the hidden influence theory aids in differentiating manipulation from persuasion and coercion because these types of influence are inherently overt (cf. Klenk, 2021c).

Nevertheless, it is unlikely that the hidden influence conceptualization of manipulation will offer trustworthy standards to fully or even partially describe the phenomenon of manipulation. On the one hand, it is difficult to manipulate a lot of hidden influences. For example, the psychology research program on heuristics and biases proposes that many of our decisions are the product of unconscious processes rather than conscious thought (Kahneman, 2012). Nevertheless, these procedures frequently appear to be valid and devoid of manipulation (see Sunstein, 2016). Because it labels an excessive number of cases as manipulation, the hidden influence criterion runs the risk of being overly inclusive and producing false positives. To explain hidden influence in a way that makes it a reliable standard for manipulation, more work would need to be done.

However, the hidden influence concept does not apply to all significant forms of manipulation (cf. Klenk, 2021c). During a house viewing, for instance, a cunning real estate agent might use the comforting aroma of freshly baked cookies to entice prospective buyers, who would know all along that they are being duped (Barnhill, 2014). Similarly, by making a service difficult and exhausting to cancel, the dark pattern referred to as a "roach motel" frequently keeps users from doing so (Brignull, 2023). Even though roach motel victims frequently have full awareness of the influence, they are nevertheless being used. Because it produces insufficient cases as manipulation, the hidden influence criterion thus runs the risk of being under-inclusive and producing false negatives. The stringent appropriateness standard Remember from my discussion in Section "Design for values and conceptual engineering" that a criterion is appropriate if it encompasses all instances of manipulation, according to the narrow criterion.

The hidden influence theory's suitability on a broad appropriateness criterion is also debatable. Putting aside the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 4
***Article Received:** 25 January 2022 **Revised:** 12 February 2022 **Accepted:** 30 March 2022*

significant concerns brought up at the outset of this section, the criterion appears to be fairly simple to apply, which could work to its advantage given a broad criterion (though see Klenk, 2023). Nevertheless, it might have the unethical consequence of shifting some of the responsibility for thwarting manipulation from the aggressor to the victim (cf. Klenk, 2021c).

Given that manipulation is by definition concealed, bringing it into the open would imply its abolition. This encourages a straightforward but ineffective strategy for countering manipulation, which calls for potential victims of manipulation to become more adept at spotting manipulation when a more sensible strategy would concentrate on controlling the manipulator's actions. Therefore, moral considerations lead one to reevaluate how manipulation is conceptualized in light of the broad appropriateness criterion, even in the event that the hidden influence conception's overstuffing and understuffing are appropriately addressed.

The bypassing rationality criterion

Another widely accepted theory is that influences that subvert reason can be used to detect manipulation (Sunstein, 2016; Wilkinson, 2013). Once more, in order for the criterion to be useful, the idea of circumventing rationality needs to be further defined (see Gorin, 2014a for discussion). The bypassing rationality conception correlates with many classic cases of manipulation and, like the hidden influence conception, should aid in distinguishing manipulation from coercion and persuasion. For instance, using generative AI to guilt-trip a target into making a charitable donation is manipulative since it appeals to the victim's emotions rather than their reason. Nonetheless, there are still significant issues with the conceptualization of "bypassing rationality." It has been heavily criticized for producing false negatives, even though it appears to be fairly accurate and accounts for many classic cases of manipulation (Gorin, 2014a, 2014b). Peer pressure and charm are two examples of manipulation techniques that don't appear to evade reason (Baron, 2003; Noggle, 2022). Therefore, not all cases of manipulation can be accurately identified by the conceptualization of manipulation that avoids bypassing.

Furthermore, a great deal of extremely significant influences, like testimony or influences that "activate heuristics," circumvent reason but do not constitute manipulation. Because of this, the bypassing criterion also produces false positives and is overly inclusive. Testimony, for instance, avoids rationality since it is frequently taken at face value and given a favorable assessment of the source's credibility. Although testimony is not likely to be a tool of manipulation, this is not a conscious, rational process. Similar to this, people can make economical decisions instinctively when using the availability or recognition heuristic.When recognition and the criterion are correlated, it makes sense to use the heuristic (Gigerenzer & Goldstein, 1996). This implies that even though "activating" the availability heuristic entails eschewing rationality in the sense of conscious thought, it need not be manipulative. In conclusion, there are issues with both over- and under-inclusivity with the bypassing rationality criterion. It is also less relevant for the goal of designing for non-manipulation because it lacks the benefit of being relatively simple—insofar as eschewing reason is harder to operationalize than covert influence.

Disjunctive conceptions of manipulation

The notions of hidden influence and circumventing it are unsuccessful since manipulation is a multifaceted and diverse phenomenon. Neither the concept of hidden influence nor the concept of bypassing rationality can effectively capture all instances of manipulation or just instances of manipulation. This made some question whether there is any satisfactory way to conceptualize manipulation given the strict appropriateness criteria (see, for example, Coons & Weber, 2014a; Klenk & Jongepier, 2022b). Disjunctive conceptions, on the other hand, might offer a way to spot manipulation. For instance, Kenton et al. (2021) consider the variety of philosophical explanations of manipulation in their discussion of the ethical alignment of language agents and choose a disjunctive conception that incorporates a number of standards that are covered in the philosophical literature.

As a result, they propose that pressure, deceit, or reason are avoided in order to achieve manipulation.12 Similar generalizations are seen in recent work on AI ethics manipulation, which combines several criteria, such as "being hidden," which is correlated with many manipulation cases in an attempt to include the phenomenon in a broad conceptualization. Disjunctive conceptualizations of manipulation, however, present issues when evaluated using a limited appropriateness standard (see Noggle, 2020, 2022). When a disjunctive conception includes manipulation criteria that are excessively inclusive on their own, the resulting disjunctive conception runs the risk of being excessively inclusive as well, incorrectly categorizing certain cases as manipulative.

For instance, incorporating "hidden influence" into a disjunctive conception runs the risk of transferring the false positive issues associated with the hidden influence criterion. One way to alleviate the concern resulting from a limited understanding of appropriateness is to view the disjunction as following a family resemblance, which would prevent individual disjuncts from being deemed sufficient for

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 4
***Article Received:** 25 January 2022 **Revised:** 12 February 2022 **Accepted:** 30 March 2022*

classification.Thirteen Even though disjunctive criteria address the issue of over-inclusivity, they still have substantial theoretical, practical, and ethical costs. 14 In theory, they hinder our ability to determine the similarities among the diverse forms of manipulation, as it's plausible that there are only distinct varieties of manipulation (refer to Coons & Weber, 2014a; Noggle, 2022). This is especially concerning because the appropriateness standard is so narrow. From a design standpoint, we would have to indicate each time what kind of manipulation we are designing against. This is a real-world issue regardless of our appropriateness standard. However, all forms will register as "manipulation" on a disjunctive conception. A measure that may work against manipulation understood as hidden influence (e.g., disclaimers) may fail to address manipulation tracked by other disjuncts, like bypassing reason. To address this, "design for non-manipulation" would always need to define precisely what kind of manipulation is within its purview, as it could be misleading given a disjunctive criterion. This shows that finding a common factor underlying all manipulation techniques has clear practical benefits, as it would facilitate clear and informative "design for non-manipulation."

Disjunctive criteria ethically complicate the development of a unified, common ethical and regulatory response to manipulative influence (see discussion in Coons & Weber, 2014a). There must be distinct ethical reactions to an influence if there are various justifications for why it counts as manipulation (a phenomenon known as supervenience). This is more intricate and differs greatly from the way ethicists and regulators currently suggest handling manipulation, which is uniformly. Therefore, a disjunctive criterion only dilutes the picture insofar as an appropriate conceptualization aids in our understanding and grasp of the phenomenon in question. The stringent appropriateness criteria make this an obvious issue.

Disjunctive conceptualizations of manipulation perform better when evaluated against a broad set of appropriateness criteria. In the field of AI ethics, there are already workable disjunctive conceptualizations of concepts other than manipulation. Text classifiers that identify hate speech, for instance, can be thought of as "tracking" a disjunctive criterion for hate speech; a criterion akin to this one might be imagined for manipulation. 15 Ultimately, though, a disjunctive conceptualization faces significant challenges. A manipulation-based text classifier would probably need to consider a wide range of difficult-to-identify contextual factors. Furthermore, it is unlikely that objectively observable characteristics of the influence, like the language employed in a text output, are inherently linked to its manipulativeness. In other words, manipulative influence doesn't "wear its manipulativeness on sleeve." The phrase "you promised to give it to me!" for instance, could be used as part of a deceptive guilt trip or as a perfectly normal, non-manipulative conversation. Without taking into account the influence's origins or motivation, such as the manipulator's intention, it doesn't seem possible that we could classify the influence with any degree of accuracy. This is due to the fact thatEliot (2023) makes the false suggestion that that generative AI can replicate objectively recognizable manipulative patterns in texts, which humans can recognize by examining the output produced by generative AI.16 Thus, a text classifier would need to examine a number of as-yet-unknown factors whose complexity needs to be taken into account in the approach's evaluation as a workable means of implementing a disjunctive conceptualization of manipulation. In conclusion, while disjunctive conceptualizations of manipulation are intriguing, they ultimately pose issues with regard to both specific and general appropriateness standards.

The trickery criterion

Understanding manipulation in terms of the influencer's goals as opposed to the influence's actual characteristics is a more fruitful strategy. According to a highly influential account, manipulation can be recognized by its deliberate attempt to deceive the recipient by making them deviate from a norm of belief, desire, or emotion (Noggle, 2020). For example, typical fraud cases are categorized as manipulation in this model because they entail an attempt to deceive the target into having an inappropriate desire or false belief. For instance, when a con artist poses as a relative over text and demands money, they aim to mislead their victim into believing something.

The clever conceptualization appears to be useful in addressing the numerous purposefully deceptive applications of generative AI. The devious idea, in particular, is effective when generative AI is employed as a tool to enable manipulative influence. Goldstein et al. (2023) provide a critical evaluation of AI-driven influence operations, highlighting the potential of generative AI to increase fraud's scale and profitability. For instance, employing generative AI to produce convincing phishing content, like texts or emails, can exacerbate phishing and other attempts to trick people into requesting information or resources. In these situations, the intention to deceive the victim is easily discernible. But it's crucial to recognize a distinct kind of manipulation made possible by generative AI, where the cunning criterion seems to be appropriate Specifically, the deceptive conceptualization leads to false negatives in a minimum of two pertinent, albeit less common, use cases.18 First, while it cannot be claimed that someone intentionally tries to deceive

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 4
***Article Received:** 25 January 2022 **Revised:** 12 February 2022 **Accepted:** 30 March 2022*

anyone, they may inadvertently use generative AI to create manipulative influences. Brignull (2023), for instance, explains how users can conduct automated A/B testing and have the "winning" design automatically implemented.

A user of this feature might only be motivated to create a website design that effectively increases sales or engagement. However, because the "winning" design might contain conventional dark patterns, the user might nevertheless be considered to be acting manipulatively due to their disregard for the true extent of their influence or lack of concern for it. This kind of inadvertent manipulation is not easily explained by the trickery explanation.19 Second, the deception account's emphasis on intents causes issues since generative AI has the potential to be manipulative in and of itself rather than just being a tool for manipulation. Although the issue over whether AI systems have intention has resurfaced in light of generative AI advancements, AI systems are usually believed to lack intention. It's possible to talk about generative AI as manipulating (Nyholm, 2022) and use a criterion to determine whether the creators or deployers of a system intended to manipulate.

However, if the system is perceived as the source of manipulation, either intentionally or through opaque (quasi-)intention, the deceptive account will provide a false negative: even if these cases appear to be manipulation, they will not be classified as manipulation.20 According to Cappuccio et al. (2022), new AI-driven manipulation techniques might be "emergent" and can't be reduced to something like a human user's goals. The significance of taking into account emergent, unintentional types of manipulation that originate in the automated behaviour of AI-driven systems is also emphasised by Pham et al. (2022). We cannot recognise unintentional manipulation that results from the automatic activity of the system if an explanation of manipulation places too much emphasis on the goal to deceive or mislead. potentially if using AI as a tool presents the most immediate risk of manipulation, there is also a threat from emergent, unintentional manipulation, which may potentially be far more dangerous than humans using generative AI for manipulative reasons. Therefore, a critical examination of the trickery criterion is necessary.21 In conclusion, the trickery notion is most challenged in situations where generative AI raises the possibility of exacerbating preexisting worries about manipulation by increasing the extent of manipulative influence.

The standard of indifference Identifying manipulation as an apathy towards an ideal condition rather than as a malevolent intent to cause harm or induce error is a notion that seeks to address these issues (Klenk, 2020, 2021c). The goal of manipulation, as defined by the indifference criterion, is to exert an effective influence without providing the other person with an explanation or an explanation that makes sense to them (Klenk, 2021c, 2023).22 For instance, the concern of the fraudster will probably have an effective influence if they use a generative AI application to create a text message that appears to be from a distressed youngster asking for money from a worried parent (i.e. successful fraud).

They don't care how they accomplish their intended objective, though, at the same moment. The indifference account highlights the fraudster's incentive to employ any strategy that works to achieve their objective, as opposed to the trickery conceptualisation, which views the fraudster as attempting to deceive the victim.23 Similar to this, using generative AI to produce a political campaign advertisement that evokes the idea of "foreign" people and those images are selected because they are believed to maximise a desired effect of the campaign (such as inciting racial hatred and xenophobia), then that use of the system qualifies as manipulative (cf. Mills, 1995). In a similar vein, automated system behaviour manipulation can be explained by the indifference theory. When a recommender system is configured, for instance, to show content that successfully grabs users' attention and shows it for that reason instead of giving them advice on who to vote for, what to buy, or what to think, then the recommender system is being abused. Furthermore, one may argue that the system operates in a manipulative manner (Klenk, 2020, 2022b). This has implications for potential applications of generative AI in the future. Even though ChatGPT and other generative AI applications can't currently be used to fine-tune their output for purposes other than text-sequence prediction, attempts to do so in the future with the intention of effectively influencing outcomes are a possibility (and have been previously discussed, for example, by Matz et al., 2023). Future generative AI applications may not be intentionally manipulative if they are optimised for effective user influence (e.g., to boost sales through a customer service application) (see the discussion below).24

Thus, manipulation is recognised by the indifference approach according to two standards. Firstly, it examines impact solely with a certain objective in mind. That being said, the perspective does not consider impact that is entirely coincidental to be manipulation, which is consistent with the most, if not all, of the literature on manipulation (see Noggle, 2018).25 Next, the indifferent view enquires as to why a specific influence method was selected in order to accomplish the pertinent objective. The selection of an influence method that is not justified by the desire to provide the target of the influence with explanations is a bad characteristic of

manipulative influence. In this respect, the manipulator is "careless" (Klenk, 2021c) or unconcerned with disclosing to their victims the rationale behind their selection of persuasion tactics. It's important to remember that the indifference view might be unintentionally understood by considering the purpose of a selected influence mechanism. A recommender system's option to "watch next video," for instance, has a specific purpose, such as encouraging the user to engage in a target behaviour.

According to the indifference view, this is manipulation because the purpose of the influence method is not to "disclose reasons." Notably, emergent, unintentional manipulation brought about by the perception that generative AI systems behave like "stochastic parrots" can be captured by the indifference criteria (Bender et al., 2021). One of the main benefits of the indifference approach over the deceptive conceptualisation of manipulation is this. In Frankfurt's definition of bullshitting, which is a form of speech act indifferent to truth, generative AI systems might be viewed as "bullshitters" (Frankfurt, 2005). According to Klenk (2022a), manipulation is a super-category of bullshit that is more broadly associated with apathy towards the truth and investigation than it is with malevolent intent.26 This sums up the "behaviour" of generative AI systems quite nicely. They behave similarly to a "trickster," consuming vast amounts of data and regurgitating what appears to be information. It is advisable to closely examine the methods, motivations, and effects of the information's production if we require the "tape" of their data (Floridi, 2023).

Conclusion

Generative AI has great potential but also great risk. It might make large-scale, automated influence possible. This can be beneficial, for example, in the creation of digital health assistants or in meaningful and moral communication. However, there is a chance that it could be manipulated. This paper presented a research strategy centred on creating generative AI systems that are non-manipulative in order to fulfil its promise and stay safe. It illustrated that starting with a suitable conceptualisation of the phenomenon is necessary if we are to design for non-manipulation, which is something that everyone interested in responsible and reliable AI should be worried about.

References

[1] Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., et al. (2021). *A general language assistant as a laboratory for alignment*. Retrieved from http:// arxiv. org/ pdf/ 2112. 00861. pdf

[2] Barnhill, A. (2014). What is manipulation? In C. Coons & M. Weber (Eds.), *Manipulation: Theory and practice* (pp. 51–72). Oxford University Press.

[3] Barnhill, A. (2022). How philosophy might contribute to the practical ethics of online manipulation. In M. Klenk & F. Jongepier (Eds.), *The philosophy of online manipulation* (pp. 49–71).

[4] Routledge.

[5] Baron, M. (2003). Manipulativeness. *Proceedings and Addresses of the American Philosophical Association, 77*, 37. https:// doi. org/10. 2307/ 32197 40

[6] Baron, M. (2014). The mens rea and moral status of manipulation. In C. Coons & M. Weber (Eds.), *Manipulation: Theory and practice* (pp. 98–109). Oxford University Press.

[7] Beauchamp, T. L. (1984). Manipulative advertising. *Business and Professional Ethics Journal, 3*, 1–22.

[8] Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics*. Oxford University Press.

[9] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. ACM Digital Library*FAccT '21: 2021 ACM Conference on Fairness, Accountability,*

[10] *and Transparency, Virtual Event Canada, 03 03 2021 10 03 2021* (pp. 610–623). Association for Computing Machinery. https:// doi.org/ 10. 1145/ 34421 88. 34459 22

[11] Buijsman, S., Klenk, M., & van den Hoven, J. (forthcoming). Ethics of AI. In N. Smuha (Ed.), *Cambridge handbook on the law, ethics and policy of artificial intelligence*. Cambridge University Press.

[12] Cappuccio, M. L., Sandis, C., & Wyatt, A. (2022). Online manipulation and agential risk. In M. Klenk & F. Jongepier (Eds.), *The philosophy of online manipulation* (pp. 72–90). Routledge.

[13] European Commission (2021). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (ARTIFICIAL INTELLIGENCE*

[14] *ACT) and amending certain Union legislative acts*. European Commission.

[15] European Commission. (forthcoming). *Meaningful and ethical communications*.

[16] European Commission. Coons, C., & Weber, M. (2014a). Introduction. In C. Coons & M.

[17] Weber (Eds.), *Manipulation: Theory and practice*. Oxford University Press.

[18] Coons, C., & Weber, M. (Eds.). (2014b). *Manipulation: Theory and practice*. Oxford University Press.

[19] European Parliamentary Research Services. (2020). *European framework on ethical aspects of artificial*

intelligence, robotics and related technologies: European added value assessment. European Parliamentary Research Services.

[20] European Commission, Directorate-General for Justice and Consumers, Lupiáñez-Villanueva, F., Boluda, A., Bogliacino, F., Liva, G., Lechardoy, L., & Rodríguez de las Heras Ballell, T. (2022).

[21] *Behavioural study on unfair commercial practices in the digital environment: Dark patterns and manipulative personalisation*. Final report.

[22] Flynn, J. (2022). Theory and bioethics. In E. N. Zalta & U. Nodelman (Eds.), *Stanford encyclopedia of philosophy: Winter 2022*. Stanford University.

[23] Frankfurt, H. G. (2005). *On bullshit*. Princeton University Press.

[24] Friedman, B., & Hendry, D. (2019). *Value sensitive design: Shaping technology with moral imagination/Batya Friedman and David G. Hendry*. The MIT Press.

[25] Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437. Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669. https:// doi. org/ 10. 1037/ 0033- 295x.103.4. 650

[26] Gorin, M. (2014a). Do manipulators always threaten rationality? *American Philosophical Quarterly, 51*(1), 51–61.

[27] Gorin, M. (2014b). Towards a theory of interpersonal manipulation. In C. Coons & M. Weber (Eds.), *Manipulation: Theory and practice* (pp. 73–97). Oxford University Press.

[28] Hacking, I. (1999). *The social construction of what?* (8th ed.). Harvard University Press.

[29] Himmelreich, J., & Köhler, S. (2022). Responsible AI through conceptual engineering. *Philosophy and Technology, 35*, 1–30. https:// doi. org/ 10. 1007/ s13347- 022- 00542-2

[30] IEEE. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems*. Retrieved from https:// stand ards. ieee. org/ conte nt/ dam/ ieee- stand ards/ stand ards/ web/ docum ents/ other/ ead1e. pdf

[31] Kahneman, D. (2012). *Thinking, fast and slow* Penguin psychology (1st ed.). Penguin. Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V., & Irving,

[32] G. (2021). *Alignment of language agents*. Retrieved from https:// arxiv. org/ pdf/ 2103. 14659

[33] Klenk, M., & Hancock, J. (2019). Autonomy and online manipulation. *Internet Policy Review*.

[34] Klenk, M. (2020). Digital well-being and manipulation online. In C. Burr & L. Floridi (Eds.), *Ethics of digital well-being: A multidisciplinary perspective* (pp. 81–100). Springer.

[35] Klenk, M. (2021a). How do technological artefacts embody moral values? *Philosophy and Technology, 34*, 525–544. https:// doi. org/ 10.1007/ s13347- 020- 00401-y

[36] Klenk, M. (2021b). Interpersonal manipulation. *SSRN Electronic Journal*. https:// doi. org/ 10. 2139/ ssrn. 38591 78

[37] Klenk, M. (2021c). Manipulation (Online): Sometimes hidden, always careless. *Review of Social Economy*. https:// doi. org/ 10. 1080/00346 764. 2021. 18943 50

[38] Klenk, M. (2022a). Manipulation as indifference to inquiry. *SSRN Electronic Journal*. https:// doi. org/ 10. 2139/ ssrn. 38591 78

[39] Klenk, M. (2022b). Manipulation, injustice, and technology. In M. Klenk & F. Jongepier (Eds.), *The philosophy of online manipulation* (pp. 108–131). Routledge.

[40] Klenk, M., & Jongepier, F. (2022a). Introduction and overview of chapters. In M. Klenk & F. Jongepier (Eds.), *The philosophy of online manipulation* (pp. 1–12). Routledge.

[41] Klenk, M., & Jongepier, F. (2022b). Manipulation online: Charting the field. In M. Klenk & F. Jongepier (Eds.), *The philosophy of online manipulation* (pp. 15–48). Routledge.

[42] Knobe, J., & Nichols, S. (2017). Experimental philosophy. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy: Winter 2017*. Stanford University.

[43] Mills, C. (1995). Politics and manipulation. *Social Theory and Practice, 21*(1), 97–112.

[44] Noggle, R. (1996). Manipulative actions: A conceptual and moral analysis. *American Philosophical Quarterly, 33*(1), 43–55.

[45] Noggle, R. (2018). The ethics of manipulation. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy: Summer 2018* (2018th ed.). Stanford University.

[46] Noggle, R. (2020). Pressure, Trickery, and a unified account of manipulation. *American Philosophical Quarterly, 57*, 241–252. https:// doi. org/ 10. 2307/ 48574 436

[47] Noggle, R. (2022). The ethics of manipulation. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy: Summer 2022* (2022nd ed.). Stanford University.

[48] Nyholm, S. (2022). Technological manipulation and threats to meaning in life. In M. Klenk & F. Jongepier (Eds.), *The philosophy of online manipulation*. Routledge.

[49] Osman, M. (2020). Overstepping the boundaries of free choice: Folk beliefs on free will and determinism in real world contexts. *Consciousness and Cognition, 77*, 102860. https:// doi. org/ 10. 1016/j. concog. 2019. 102860

[50] Osman, M., & Bechlivanidis, C. (2021). Public perceptions of manipulations on behavior outside of awareness. *Psychology of Consciousness: Theory, Research, and Practice*. https:// doi. org/ 10.

[51] 1037/ cns00 00308

[52] Osman, M., & Bechlivanidis, C. (2022). Impact of personalizing experiences of manipulation outside of awareness on autonomy. *Psychology of Consciousness: Theory, Research, and Practice*. https:// doi. org/ 10. 1037/ cns00 00343

[53] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder,

[54] P., Christiano, P., … Lowe, R. (2022). *Training language models to follow instructions with human feedback*. Retrieved from http:// arxiv. org/ pdf/ 2203. 02155. Pdf

[55] Pepp, J., Sterken, R., McKeever, M., & Michaelson, E. (2022). Manipulative machines. In M. Klenk & F. Jongepier (Eds.), *The philosophy of online manipulation* (pp. 91–107). Routledge.

[56] Pham, A., Rubel, A., & Castro, C. (2022). Social media, emergent manipulation, and political legitimacy. In M. Klenk & F. Jongepier (Eds.), *The philosophy of online manipulation.* Routledge.

[57] Sunstein, C. R. (2016). *The ethics of influence: Government in the age of behavioral science*. Cambridge University Press.

[58] Susser, D., Roessler, B., & Nissenbaum, H. (2019a). Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review, 4*(1), 1–45.

[59] Susser, D., Roessler, B., & Nissenbaum, H. (2019b). Technology, autonomy, and manipulation. *Internet Policy Review, 8*, 1–22. https:// doi. org/ 10. 14763/ 2019.2. 1410

[60] Tremblay, M. S., Colley, R. C., Saunders, T. J., Healy, G. N., & Owen, N. (2010). Physiological and health implications of a sedentary lifestyle. *Applied Physiology, Nutrition, and Metabolism, 35*,

[61] 725–740. https:// doi. org/ 10. 1139/ H10- 079

[62] van de Poel, I. (2013). Translating values into design requirements. *Philosophy and engineering: Reflections on practice, principles and process* (pp. 253–266). Springer.

[63] van de Poel, I. (2015). Conflicting values in design for values. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of ethics, values, and technological design: Sources, theory, values*

[64] *and application domains* (pp. 89–116). Springer.

[65] van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines, 30*, 385–409. https:// doi. org/ 10. 1007/ s11023- 020- 09537-4

[66] van den Hoven, J., Vermaas, P. E., & van de Poel, I. (2015). Design for values: An introduction. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of ethics, values, and technological design: Sources, theory, values and application domains* (pp. 1–7). Springer.

[67] Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., et al. (2022). Taxonomy of risks posed by language models. *FAccT '22: 2022 ACM Conference on Fairness, Accountability,and Transparency, Seoul Republic of Korea, 21 06 2022 24 06 2022* (pp. 214–229). ACM. https:// doi. org/ 10. 1145/ 35311 46. 35330 88

[68] Wilkinson, T. M. (2013). Nudging and manipulation. *Political Studies, 61*, 341–355. https:// doi. org/ 10. 1111/j. 1467- 9248. 2012. 00974.x