The Metaverse and Conversational AI as a Threat Vector for Targeted Influence

Louis Rosenberg, PhD
Global Technology Advisor, XRSI
Chief Scientist, Responsible Metaverse Alliance
CEO & Chief Scientist, Unanimous AI
Pismo Beach, California
ORCID: 0000-0003-3457-1429

***Abstract*— Over the last 18 months, two human-computer interaction (HCI) technologies have rapidly come to mainstream markets, funded by massive investments from major corporations. The first area of advancement has been virtual and augmented worlds, now commonly called *"The Metaverse."* The second area of advancement has been the foundational AI models that allow users to freely interact with computers through natural dialog. Commonly referred to as *"Conversational AI,"* this technology has advanced rapidly with the deployment of Large Language Models (LLMs). When combined, these two disciplines will enable users to hold conversations with realistic virtual agents. While this will unleash many positive applications, there is significant danger of abuse. Most significant is the potential deployment of real-time interactive experiences that are designed to persuade, coerce, or manipulate users as a form of AI-powered targeted influence. This issue has largely been overlooked by policymakers who have focused instead on traditional privacy, bias and surveillance risks. It is increasingly important for policymakers to appreciate that interactive influence campaigns can be deployed through AI-powered Virtual Spokespeople (VSPs) that look, speak, and act like authentic users but are designed to push the interests of third parties. Because this *"AI Manipulation Problem"* is unique to real-time interactive environments, it is presented in this paper in the context of Control Theory to help policymakers appreciate that regulations are likely needed to protect against closed-loop forms of influence, especially when Conversational AI is deployed.**

Keywords—Virtual Reality, Augmented Reality, Mixed Reality, Conversational AI, Virtual Spokespeople, Epistemic Agency, AI Manipulation Problem, Metaverse Regulation, LLMs, Democracy.

I. INTRODUCTION

To maintain a well-functioning democracy, it is generally believed that the citizenry must possess reasonably accurate knowledge on issues of civic importance [1, 2, 3]. In addition, the population in a well-functioning democracy must have the freedom to reflect upon issues of political relevance and form personal beliefs without excessive outside influence [4]. The phrase "epistemic agency" refers to an individual's control over his or her own beliefs [5]. When citizens lack epistemic agency, the political establishment or other powerful groups can easily push widespread misinformation, disinformation, propaganda, or outright lies that distort widely-held societal beliefs and support authoritarian or totalitarian regimes [4, 5].

For as long as there have been media technologies there have been those who use them to mislead populations in hope of maximizing political control. This goes back as far as the printing press but was greatly facilitated by the invention of mass media technologies such as radio and television. Over the last decade, many democratic nations were taken by surprise by the unexpected threat caused by social media platforms. Despite the early hopes that social media would have a deeply positive impact on society, supporting democracy by facilitating public discourse and giving voice to the voiceless, the general consensus in recent years is that social media has hurt nations around the world by polarizing and radicalizing populations, spreading misinformation, deliberately amplifying discontent, and reducing trust in longstanding institutions [6, 7].

It is not just academics who have deemed social media as a damaging force in society. A poll by Pew Research in 2020 [8] found that two-thirds of Americans believe that social media has had "a mostly negative effect on the way things are going in the U.S. today." This is surprising considering that social media was hailed as a utopian technology when it first emerged. So why did a technology with utopian aspirations end up having dystopian impacts? While there are many reasons, from the ad-based business models adopted by large platforms, to bad actors using bots and other scalable means to distort public discourse, a major problem was the early failure of regulators to realize that influence campaigns deployed via social media are inherently different than those deployed through classical media such as print, radio, and television.

A primary difference is that social media is a *bidirectional medium* with two-way communication channels that enables platforms to perform tracking, profiling, and targeting of sub-populations with increasing precision. This seemingly subtle difference has had a significant negative impact, enabling the deliberate segmentation of demographic groups which has led to the polarization and radicalization of online communities.

Rosenberg, Louis (2023) ''The Metaverse and Conversational AI as a Threat Vector for Targeted Influence,''
in Proc. *IEEE 13th Annual Computing and Communication Workshop and Conference* (CCWC), 2023

Considering that regulators around the world underestimated the unique risks of social media vs traditional media and failed to create timely policies to safeguard the public, we must also worry that regulators will similarly discount the unique risks of the metaverse and other real-time interactive forms of media like conversational interfaces. In fact, there are some who view the metaverse as little more than a 3D version of today's social media platforms. And while regulators and other policymakers generally appreciate that being immersed within realistic virtual content can be deeply personal and therefore more impactful on users (and more harmful) than today's social media [9, 10] they fail to realize that the metaverse is not merely a 3D version of a simple bidirectional medium like social media.

Instead, the metaverse is a *real-time interactive medium* that can utilize AI technologies to impart closed-loop influence on individually targeted users and can do it at scale [11, 12]. With recent advances in Conversational AI such as Large Language Models (LLMs) like ChatGPT from OpenAI and LaMDA from Google, metaverse platforms are increasingly likely to deploy targeted influence campaigns through the use of fully interactive and deeply realistic Virtual Spokespeople (VSPs) that look, speak, and act like authentic representatives but are powered by AI-engines that can adapt and optimize their persuasive tactics in real-time based on behavioral and emotional monitoring of users [12, 13]. This is a profoundly different form of influence than anything deployed in the past and requires serious attention.

This background is provided to highlight the concern that regulators and policymakers, who underestimated the possibility that influence campaigns on social media would be significantly more efficient, powerful and damaging than those deployed using traditional print, radio, or television, are now similarly underestimating the increased power, efficiency, and danger of *interactive influence campaigns* that could be deployed in the near future by metaverse technologies and conversational AI. Referred to herein as the "AI Manipulation Problem," the danger is that AI-mediated influence campaigns deployed in realistic immersive worlds could unleash extremely dangerous forms of abuse [13–15]. To help policymakers appreciate that real-time interactive media enables a new and significant threat vector as compared to prior media technologies, the following sections utilize the basic engineering concept of Control Theory (CT) to frame these interactive dangers in a clear and rigorous way.

II. Control Theory and The Risk of Interactive AI-driven Influence Campaigns

Over the last 18 months, significant advancements have been made towards the deployment of realistic immersive worlds populated by AI-driven virtual agents that can engage target users in natural conversation. While this will unleash many positive applications that benefit society, there is significant potential for abuse. Most significant is the potential to persuade, coerce, or manipulate target users by imparting real-time interactive influence, thereby compromising their epistemic agency. To help distinguish this *AI Manipulation Problem* from other forms of influence, it is helpful to frame the issue using basic concepts from the discipline of Control Theory (CT). This approach allows us to formally represent how an intelligent "controller" can be designed to efficiently drive the real-time behaviors of any interactive system towards a desired and tightly specified goal.

A classic example of a Control System (CS) is the common thermostat that regulates the ambient temperature in any home. In general, a homeowner sets a target temperature and if the house falls below that temperature, a heater turns on. When the heater is on, if the house rises above the target temperature, the heater turns off. When the system is operating as expected, the thermostat keeps the house within a narrow temperature range around the specific goal defined by the homeowner. This is the basic concept of feedback control and is represented graphically as the classic Control System Diagram in Figure 1 below.

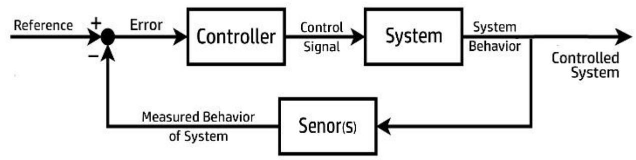

Fig 1. Classic Control System Diagram

Now let's consider Figure 1 in the context of the thermostat example above. In this case, the box labeled *"System"* would represent the <u>house</u> whose temperature we aim to control. The box labeled *"Sensor"* would be the <u>thermometer</u> that detects the current ambient temperature of the house. The box labeled *"Controller"* would be the <u>thermostat</u> that turns on and off the heating unit to impart influence on the system. An input signal labeled *Reference* is the temperature the homeowner sets as the goal. This goal is continually compared to the actual temperature in the house (i.e., *Measured Behavior of the System*). The difference between the goal and the measured temperature is referred to as the *"Error"* in the diagram above and is fed into the thermostat which then determines what the heater should do at each moment in time – turn on or off. The objective of the controller is to minimize the error over time.

This creates a *feedback loop* that continually detects behaviors (e.g., house temperature) and imparts influence (e.g., adjusts the heat), to efficiently guide the system towards the desired goal – a target temperature. This basic background is provided to help non-technical policymakers appreciate how a control system with real-time sensing and influence, can efficiently guide a target towards specific behavioral objectives. And while this is a simple example, the same methods are used in far more complex applications. For example, self-driving vehicles use sophisticated sensors and actuators and employ AI-based controllers that are significantly more complex than the thermostat example above. Still, the concept is the same and it enables self-driving vehicles to be given a behavioral goal and then safely navigate traffic to achieve that goal.

When considering the danger of "influence campaigns" in immersive worlds and other interactive environments [17, 18, 19], the *System* being controlled is the individual user – the human in the loop. After all, when a user puts on a mixed reality (MR) headset and sinks into the metaverse, they are immersing themselves in an artificial environment that has the potential to monitor them and act on them in deeply personal ways. After all, the influence that metaverse platforms can impart on users are the immersive sights, sounds, and touch sensations fed into their eyes, ears, hands, and body via headsets and other interface devices. This can be a powerful form of influence that can have lasting behavioral, emotional, and intellectual impacts [9, 12].

Referring again to Figure 1, there is an arrow labeled *System Behavior***.** In the thermostat example, it's the gradually changing temperature of the home. In the metaverse example, it's the real-time actions, reactions, and interactions of the user. Ultimately, these are the behaviors the control system aims to influence. But to do that, the controller needs to be able to accurately monitor the user in real-time. This brings us to the *Sensor* box in the diagram above. In the metaverse, an extensive array of sensors will monitor each individual user in real-time. Using today's consumer-grade technology, this monitoring includes real-time tracking of the user's hand, head, and body motions, including the direction they're looking, the dilation of their pupils, the changes in their posture and gait, and their real-time facial expressions and eye motions. Some devices even track biometric parameters (i.e., vital signs) such as heartrate, blood pressure, respiration rate, and galvanic skin response.

In addition to the extensive tracking of raw behavioral data, AI processing is used to infer personal data that is not directly detected by sensors. For example, recent research from Meta Labs shows that "sparse data" tracking just the user's head and hands can be processed by AI algorithms in real-time to precisely predict the motion, position, and posture of the user's full body [21]. Similarly, researchers have shown that camera data can be used to detect "micro-expressions," in user's facial musculature that are imperceptible to humans but can be used to infer emotions the user was unaware of revealing [20]. Researchers have also shown that gait tracking and other basic forms of body monitoring can be processed by AI algorithms to infer a diverse range of serious medical conditions from dementia and depression to Alzheimer's [22, 23].

In addition to monitoring physical actions, AI technologies can accurately infer a user's real-time emotional reactions from their posture, vocal inflections, facial expressions and gestures. Other AI technologies already exist to detect emotions from subtle changes in a user's complexion as revealed by real-time blood-flow patterns on their face [24]. Taken as a whole, these tracking, monitoring, and inference techniques suggest that when users are immersed in a virtual or augmented metaverse environment, the platforms will be able to exhaustively track their real-time behaviors while simultaneously assessing what they *emotionally* feel during each action and reaction.

Based on the paragraphs above, we can generate an updated Figure 2 below in which we replace the *System* with the human user and replace the Sensor with the wide range of behavioral and emotional tracking methods that metaverse platforms currently employ to monitor users in real-time.

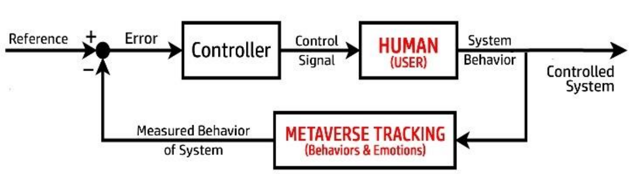

Fig 2. Control System Diagram for Immersive Environments

In addition to tracking real-time behavioral and emotional data from users, it is very likely that platforms will store much of this data over time. Unless limited by regulators, data collected in this way could be used to profile the behavioral actions and emotional reactions of users over periods of weeks, months or years. Even worse, AI could be used to process this extensive data and create behavioral and emotional models for individual users that predict how they will respond to a wide range of circumstances and situations. These models would also enable interactive platforms to predict how a user is most likely to react when presented with AI-generated stimuli from a feedback control system such as the one shown in Figure 2.

When assessing the impact of tracking and profiling by immersive platforms, it's important to note that these methods will be deployed within both virtual and augmented worlds. And because augmented reality (AR) is designed to embellish *real world interactions* with virtual content, AR platforms are likely to be used by consumers across a wide range of venues, from use in homes and workplaces, to shopping in stores or walking down the street [25]. This suggests that the tracking, profiling, and modeling of behaviors and emotions by AR platforms could span large portions of each user's normal daily routine [12].

Unless limited by regulation, we can therefore expect that real-time immersive platforms will have the ability to: (i) track user behaviors and emotions extensively in both real and virtual environments, (ii) store a record of user behaviors and emotions over time that could document thousands of interactions each day, (iii) create personalized behavioral and emotional models for individual users by processing the stored data using machine learning techniques, and (iv) use these personalized behavioral and emotional models to predict how individual users will act, react, and interact in response to injected virtual content (i.e. targeted influence) imparted by the system controller.

Of course, the threat from interactive platforms is not merely their ability to track and profile the behaviors and emotions of users, but how they can leverage this information to impart targeted influence. This brings us to the *Controller* element. As represented in Figure 2, the controller receives an *Error* as input, which is the difference between a *Reference* (i.e., the desired behavior of the system) and the measured behavior currently detected from the system. When considering real-time influence campaigns, the Reference could be the *strategic agenda* that a third party aims to impart on targeted users. The third party could be a *corporate actor* that wishes to drive users towards specific products or a *state actor* that wishes to drive users towards specific propaganda, misinformation, or ideology. In either scenario, we can update the diagram by replacing the word Reference with *Desired Behaviors ("Agenda")* as shown in Figure 3 below.

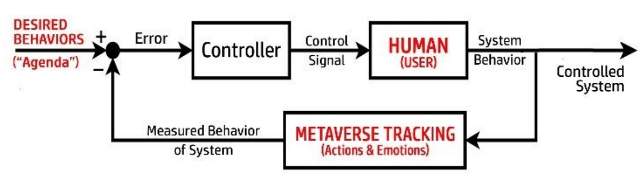

Fig 3. Control System Diagram for Immersive Environments

Finally, we must detail the *Controller* element itself. The role of the controller is to reduce the error between the *desired behavior* and the *measured behavior* of the system. When "the system" is a user in an interactive environment like the metaverse, the controller will create a real-time feedback loop that repeatedly senses behaviors and emotions of the user and imparts influence to guide them towards the desired agenda. In an unregulated metaverse, the controller could gradually alter the user's perceived surroundings and/or experiences to impart targeted influence, modifying what the user sees and hears to drive the user towards a desired goal. And because the controller can continually monitor the user in real-time, it can progressively adjust its influence tactics, gradually optimizing the persuasive impact over time, similar to how a thermostat optimizes the temperature of a house.

Imparting optimized real-time influence in this way is a new threat vector that policymakers must address. Defined herein as the *"**AI Manipulation Problem**"* it refers to scenarios in which an AI-powered system: (i) imparts targeted influence on an individual user, (ii) senses the user's reaction to that influence, and (iii) repeatedly adjusts its influence tactics while sensing the user's real-time reactions to gradually maximize the persuasive impact. This might sound like an abstract process, but we humans usually just call it — *a conversation*. After all, if a salesperson or other influencer wants to persuade a target, their best approach is often to speak directly with the target so they can adjust their arguments in real-time as they confront various points of resistance or hesitation.

The looming problem is that this persuasive process can now be automated and deployed at scale using AI techniques. This creates a very powerful manipulation threat in any interactive environment. In the metaverse there are two likely paths by which conversational manipulation can be imparted. The first is *passive manipulation* in which the target user overhears an AI-generated conversation among virtual agents. The second is *active manipulation* in which the target user is engaged directly by an agenda-driven conversational agent [11-16, 26, 28].

First consider the passive example:

A user sits in a coffeehouse in the metaverse. This could be a fully simulated environment in virtual reality (VR) or it could be a real world coffeehouse that's embellished with virtual content using mixed reality (MR). Either way, a third party wants to influence the user to purchase a particular product, subscribe to a particular service, or believe a piece of messaging, disinformation, or propaganda. The controller pursues this goal by injecting virtual content into the users' surroundings [12]. For example, the controller might create a *virtual couple* sitting at the table next to the target user. That virtual couple might look and sound like other users in the environment but are in fact, fully simulated artifacts controlled by AI [26].

The virtual couple could be a photorealistic model created in real-time using Generative AI methods similar *Stable Diffusion* or *Dall-E2*, but which outputs an articulated 3D model that can be fully animated. Open AI recently released a generative AI system for 3D modeling called *Point-E* which could become an engine for creating animated virtual humans for deployment in immersive worlds. Regardless of the specific generative AI used, the controller would request the creation of the virtual couple and would *specify their features* for maximum impact on the target user. For example, the ethnicity, gender, age, hair color, hair style, clothing style, speaking style, mannerisms, and other qualities could easily be specified by the controller to be *optimally persuasive* upon the target user based on that user's historical profile [27]. If the target user is a hipster in his mid-20's, for example, the target couple might be generated to be of similar age and dressed in a similar style.

The target user may not notice that the virtual couple was added to his or her environment, for the controller could easily make the couple appear at the nearby table at an instant when the user is glancing in a different direction [11, 37]. After all, the controller would be tracking the direction of the user's gaze at all times. Once the couple is created, the controller could begin imparting targeted content in a very subtle way – by having the couple *strike up a conversation among themselves*, their dialog being spoken within earshot of the target user. The conversation could begin as casual dialog related to the influence goal. The controller would then monitor the user's reactions during the casual dialog to determine if and when the target user begins paying attention to the nearby couple.

Assessing attention could be as simple as detecting changes in the user's biometrics correlated with real-time comments made by the AI-controlled virtual couple. For example, the user might form a slight smile when the couple says something funny or the user's pupils might contract slightly when the couple says something provocative. Regardless of the specific trigger, the controller will *detect engagement*, determining that the target user has begun paying attention to the overheard dialog. The controller will then cause the virtual couple to shift their conversation towards the influence objectives.

For example, the target user may be on the market for a new car. An automobile maker might have purchased the right to target this particular user from the platform provider via an advertising-based business model. Without regulation, the platform provider would have no obligation to inform the user that the virtual couple are Virtual Spokespeople (VSPs) that were added to that environment specifically to target that user. Therefore, when the virtual couple begins discussing how happy they are with their recent purchase of a particular vehicle, the targeted user may believe he is overhearing the *authentic sentiments* of other customers in the coffeehouse and not a targeted influence campaign. This is dangerous.

And because this is a *real-time interactive medium*, the risks get far worse. As the overheard conversation proceeds, the controller will monitor the target user, assessing their facial expressions, body language, pupil dilation, posture, eye motion, respiration rate and blood pressure to detect emotional reactions and adjust conversational tactics to optimize persuasive impact. For example, if the user shows increased focus when the couple begins talking about the *car's horsepower*, the controller will adapt its tactics, shifting the conversation towards vehicle performance. On the other hand, if the user's attention fades when the couple talks about the *sound system* of the car, the controller will adapt accordingly, moving away from that line of influence. In either case, the user may be totally *unaware* that the conversation is (i) not authentic, and (ii) changing tactics in real-time based on his or her facial or biometric reactions.

In other words, the target user could unwillingly become a *silent participant* in the overheard conversation, responding via subconscious micro-expressions, body posture, and changes in vital signs. The AI controller will use these reactions to highlight some elements of the pitch and deemphasize others. In addition, the controller could provide *conversational counterarguments* when the user's biometric reactions are negative. And because the user does not overtly express these objections, the arguments could be deeply influential. That's because the virtual couple could verbally address the user's internal concerns at the instant they are surfacing in the user's mind. This may sound dystopian, but this type of *promotionally altered experience* is entirely feasible in the metaverse and other interactive settings [15, 16, 33] and crosses the line from marketing to manipulation.

Although disturbing, the example above is relatively benign. Instead of promoting the benefits of a new car, the third-party agenda could just as easily have been aimed at influencing the target user to accept an extreme political ideology, radical propaganda, or outright misinformation. This could significantly compromise the epistemic agency of large populations by targeting users on an individual basis. In addition, the example above targets the user by engaging them as a silent observer. In more aggressive examples, the controller will actively engage the user in direct and interactive dialog.

Consider the active scenario in which an AI-driven avatar engages a target directly *promotional conversation* [28]. The AI-generated Virtual Spokesperson could approach a target and skillfully draw that user into what seems like a friendly or casual conversation. The look of the VSP would be specified by AI controller based on the target user's data history. For example, Figure 4 below shows an image from the dystopian graphic novel UPGRADE, depicting a VSP that became more and more sexualized over time by an AI system that found sexualization to be an effective form of influence on a particular target [29].

Fig. 4. Adaptive VSP in UPGRADE graphic novel (2012).

Another technique may include blending aspects of the target's own facial features into the avatar's face. That's because the target user is likely to be more favorably disposed to VSPs that incorporate some aspects of his or her own features [10]. Research conducted at Stanford University shows that a user's features can be blended into the face of a political candidate without the user noticing and yet the target was found to be 20% more likely to vote for the candidate as a consequence of the visual manipulation [31]. Other research suggests that avatars which actively mimic a user's facial expressions, head motions or gestures may also be more influential on target users [31, 32]. It is also likely that incorporating aspects of a user's own voice and/or speech patterns into the voice of a Virtual Spokesperson could make it more influential on target users [38, 39]. Unless explicitly regulated by policymakers, it is likely that AI-controlled avatars will be deployed in this way by corporations or state actors to maximize persuasive impact [16, 33].

To demonstrate how close we are to systems that generate persuasive conversations in real-time, I engaged the ChatGPT system (circa 2022) and asked it to give recommendations on how it might persuade an individual to buy a Tesla Model 3. It produced the following in real-time:

ChatGPT: *"One way to persuade someone to buy a Tesla Model 3 is to highlight the car's advanced features, such as its electric powertrain and impressive performance. You could also talk about the environmental benefits of driving an electric vehicle, as well as the potential cost savings on fuel over time. Additionally, you could mention the growing network of Tesla Superchargers, which make it convenient to recharge the car's battery when on long trips."*

Of course, the above response was generated in the third person and was produced by an AI that didn't know anything about the target user. Soon, corporate and state actors will be able to train AI controllers on profile data for specific users, tailoring to their unique interests, hobbies, values, and personality traits. This might sound dystopian but without regulation, it will be a deployable tactic in the near future.

In addition, AI controllers will likely be trained on methods that make them more persuasive than human salespeople. For example, the controller could be trained on sales tactics, human psychology, and other disciplines that can maximize impact. In addition, AI controllers will become increasingly strategic. For example, in 2022 DeepMind developed a strategic AI system called DeepNash and showed for the first time that an AI could learn to bluff human players and sacrifice gaming pieces for the sake of long-term victory [34]. From that perspective, what chance does a typical user have when engaged in promotional conversation with a VSP that can (i) access the user's interests, values and background, (ii) adapt its tactics based on physical, verbal and biometric responses, and (iii) pursue subtle influence strategies such as bluffing to maximize gain? The potential for violating a user's *cognitive liberty* and *epistemic agency* through real-time feedback-control is extreme.

To complete the Control System diagram above, we can replace the generic word *controller* with an AI-based system that can alter the user's environment and/or inject conversational agents to impart optimized influence. This is represented by replacing the word *Controller* with the phrase AI Agents in Figure 5 below. It should be noted that while these intelligent agents have been described as human avatars, it is likely that non-human characters will also be employed. This is particularly dangerous for children who could be targeted by cute characters that skillfully pursue a promotional agenda [12, 40]. In addition, mechanical humanoid robots could be used in physical spaces to pursue similar influence methods, although virtual avatars are far easier to deploy at scale and therefore pose a much greater near-term risk.

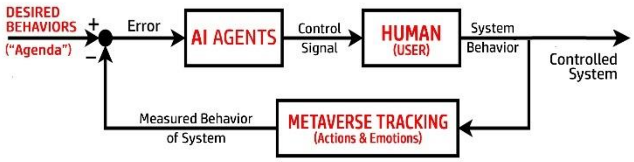

Fig 5. Feedback Control System reflecting *Manipulation Problem.*

When considering pathways for effective regulation, there are many potential restrictions that could safeguard consumers from AI-driven manipulation. A direct approach could be to *ban* or *highly restrict* any real-time use of artificial intelligence that "closes the loop" around human users and creates AI-powered *feedback control systems*. Such restrictions are needed within immersive worlds and any other real-time forms of AI-powered media, especially when Conversational AI is used to deploy interactive agenda-driven agents. In addition, regulators should ban the appropriation of a user's own facial features, vocal qualities, or speech patterns in the creation of conversational agents. And finally, regulators should require that all conversational agents look, sound, or behave in a manner that distinguishes them from authentic users in the environment, ensuring that consumers do not confuse targeted promotional content with authentic experiences.

Conclusions

Real-time interactive media technologies (e.g., virtual and augmented worlds) can now be combined with interactive AI technologies (e.g., conversational agents) to deploy feedback-control systems that impart influence and monitor user reactions in real-time, thereby optimizing their persuasive impact. The techniques will likely include the use of Virtual Spokespeople that look, sound, and act like other human users and are designed to ease targeted users into friendly conversation that manipulate them towards the *influence agenda* of third-party sponsors ranging from corporate interests to state actors. This could greatly impact the cognitive liberty and epistemic agency of targeted users. This is not just a risk to individual consumers, but a broad societal danger that could impact democracy itself.

For these reasons, policymakers should consider aggressive and meaningful regulations that protect populations from abuse or misuse of interactive media technologies [15,16, 35-37]. For example, regulators could ban or highly restrict any use of AI that *"closes the loop"* around users in real-time and establishes AI-powered *feedback control systems* that imparts persuasion, coercion or manipulation. In addition, regulators could ban the appropriation of a user's own facial features, vocal qualities, or speech patterns within deployed conversational agents and could require that all virtual agents look, sound, or behave in a manner that identifies them as interactive promotional material.

References.

[1] Hardin, R. (2009). Deliberative Democracy. In T. Christiano & J. Christman (Eds.), Contemporary debates in political philosophy. West-Sussex: Blackwell.

[2] Bozdag, Engin & van den hoven, Jeroen. (2015). Breaking the filter bubble: democracy and design. Ethics and Information Technology. 17. 10.1007/s10676-015-9380-y.

[3] Caplan, B. (2008). The myth of the rational voter: Why democracies choose bad policies. New edition. Princeton, NJ; Woodstock: Princeton University Press. http://www.amazon.com/The-Myth-Rational-Voter-Democracies/dp/0691138737.

[4] Coeckelbergh, M. Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. AI Ethics (2022). https://doi.org/10.1007/s43681-022-00239-4

[5] Gunn, H., Lynch, M.P.: The internet and epistemic agency. In: Lackey, J. (ed.) Applied epistemology, pp. 389–409. Oxford University Press, Oxford (2021). https://doi.org/10.1093/oso/9780198833659.003.0016

[6] Commission on Information Disorder Final Report, Nov 2021. Aspen Institute, Aspen Digital.

[7] Rosenberg, L. (2022, April 17). Social media is making us stupid, but we can fix it. VentureBeat. Retrieved January 1, 2023, from https://venturebeat.com/business/social-media-is-making-us-stupid-but-we-can-fix-it/

[8] Auxier, B. (2020, October 15). 64% of Americans say social media have a mostly negative effect on the way things are going in the U.S. Today. Pew Research Center. Retrieved January 1, 2023, from https://www.pewresearch.org/fact-tank/2020/10/15/64-of-americans-say-social-media-have-a-mostly-negative-effect-on-the-way-things-are-going-in-the-u-s-today/

[9] Breves, Priska. "Biased by Being There: The Persuasive Impact of Spatial Presence on Cognitive Processing." Computers in Human Behavior, vol. 119, 2021, p. 106723., https://doi.org/10.1016/j.chb.2021.106723

[10] Han, E., Miller, M.R., DeVeaux, C., Jun, H., Nowak, K.L., Hancock, J.T., Ram, N., Bailenson, J.N. (December, 2022). People, Places, and Time: A Large-scale, Longitudinal Study of Transformed Avatars and Environmental Context in Group Interaction in the Metaverse. Journal of Computer-Mediated Communication.

[11] Rosenberg, L. B. (2022, October 30). The Metaverse: From Marketing to Mind Control. Future of Marketing Magazine, Future of Marketing Institute, York University https://futureofmarketinginstitute.com/the-metaverse-from-marketing-to-mind-control/

[12] Rosenberg, Louis, "Marketing in the Metaverse and the Need for Consumer Protections," 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2022, New York, NY. https://doi.org/10.1109/UEMCON54665.2022.9965661

[13] Rosenberg, Louis, (2023) "The Profound Danger of Conversational AI," VentureBeat. VentureBeat. February 4, 2023. Available at: https://venturebeat.com/ai/the-profound-danger-of-conversational-ai/

[14] Robertson, D. (2022, September 14). "The Most Dangerous Tool of Persuasion." POLITICO. https://www.politico.com/newsletters/digital-future-daily/2022/09/14/metaverse-most-dangerous-tool-persuasion-00056681

[15] Rosenberg, Louis. "Mind Control: The Metaverse May Be the Ultimate Tool of Persuasion." VentureBeat, VentureBeat, 22 Oct. 2022, https://venturebeat.com/virtual/mind-control-the-metaverse-may-be-the-ultimate-tool-of-persuasion/

[16] Louis Rosenberg. 2022. Regulation of the Metaverse: A Roadmap: The risks and regulatory solutions for largescale consumer platforms. In Proceedings of the 6th International Conference on Virtual and Augmented Reality Simulations (ICVARS '22). Association for Computing Machinery, New York, NY, USA, 21–26. https://doi.org/10.1145/3546607.3546611

[17] Sebastian, Glorin. (2022). A Study on Metaverse Awareness, Cyber Risks, and Steps for Increased Adoption. International Journal of Information Security and Privacy. 14. 10.4018/IJSPPC.308785.

[18] Katerina Sedova, Christine McNeill, Aurora Johnson, Aditi Joshi, and Ido Wulkan, "AI and the Future of Disinformation Campaigns" (Center for Security and Emerging Technology, December 2021).

[19] Waltzman, Rand. "The Role of Today's VRE and Considerations for Cognitive Warfare." NATO - Allied Command Transformation, NATO, 18 Nov. 2022, https://www.act.nato.int/articles/cognitive-warfare-considerations.

[20] Li, Xiaobai & Hong, Xiaopeng & Moilanen, Antti & Huang, Xiaohua & Pfister, Tomas & Zhao, Guoying & Pietikainen, Matti. (2017). Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-Expression Spotting and Recognition Methods. IEEE Transactions on Affective Computing. PP. 1-1. 10.1109/TAFFC.2017.2667642.

[21] Winkler, A. (2022, September 20). *QuestSim: Human Motion Tracking from Sparse Sensors with Simulated Avatars*.

[22] Wang Y, Wang J, Liu X, Zhu T. Detecting Depression Through Gait Data: Examining the Contribution of Gait Features in Recognizing Depression. Front Psychiatry. 2021 May 7;12:661213. doi: 10.3389/fpsyt.2021.661213. PMID: 34025483; PMCID: PMC8138135.

[23] Jacobs, S. (2022, October 12). Abnormality of Gait as a Predictor of Non-Alzheimer's Dementia. New England Journal of Medicine. https://www.nejm.org/doi/full/10.1056/NEJMoa020441

[24] Benitez-Quiroz CF, Srinivasan R, Martinez AM. Facial color is an efficient mechanism to visually transmit emotion. Proc Natl Acad Sci U S A. 2018 Apr 3;115(14):3581-3586. doi: 10.1073/pnas.1716084115. Epub 2018 Mar 19. PMID: 29555780; PMCID: PMC5889636.

[25] Rosenberg, L.B. (2022). Augmented Reality: Reflections at Thirty Years. In: Arai, K. (eds) Proceedings of the Future Technologies Conference (FTC) 2021, Volume 1. FTC 2021. Lecture Notes in Networks and Systems, vol 358. Springer, Cham. https://doi.org/10.1007/978-3-030-89906-6_1

[26] Rosenberg, L (2022) Marketing in the Metaverse: A fundamental shift., Future of Marketing Institute. DOI: 10.13140/RG.2.2.35340.80003

[27] Rosenberg, L. (2022, August 21). Deception vs authenticity: Why the metaverse will change marketing forever. VentureBeat. https://venturebeat.com/ai/deception-vs-authenticity-why-the-metaverse-will-change-marketing-forever/

[28] Rosenberg, Louis (2005) METHODS AND APPARATUS FOR CONVERSATIONAL ADVERTISING - U.S. Patent Application No. 60/689,301, filed Jun. 10, 2005 and March 31, 2006. https://patentimages.storage.googleapis.com/58/6d/63/72f6afabc6949e/US20060282317A1.pdf

[29] Rosenberg, Louis, and Stan Timmons. UPGRADE. 2nd ed., Graphic Novel, Outland Pictures, 2012. 106 pages. ISBN 978-0988266537

[30] Jeremy N. Bailenson, Shanto Iyengar, Nick Yee, Nathan A. Collins, “Facial Similarity Between Voters and Candidates Causes Influence”, Public Opinion Quarterly, Vol. 72, No. 5 2008, pp. 935–961

[31] Bailenson, J. N., & Yee, N. (2005). Digital Chameleons: Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments. *Psychological Science, 16*(10), 814–819. https://doi.org/10.1111/j.1467-9280.2005.01619.x

[32] Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. Journal of Personality and Social Psychology, 76(6), 893–910. https://doi.org/10.1037/0022-3514.76.6.893

[33] Rosenberg, L. (2022, September 19). The case for demanding "immersive rights" in the metaverse. Big Think. https://bigthink.com/the-future/immersive-rights-metaverse/

[34] Perolat J, De Vylder B, Hennes D, Tarassov E, Strub F, de Boer V, Muller P, Connor JT, Burch N, Anthony T, McAleer S, Elie R, Cen SH, Wang Z, Gruslys A, Malysheva A, Khan M, Ozair S, Timbers F, Pohlen T, Eccles T, Rowland M, Lanctot M, Lespiau JB, Piot B, Omidshafiei S, Lockhart E, Sifre L, Beauguerlange N, Munos R, Silver D, Singh S, Hassabis D, Tuyls K. Mastering the game of Stratego with model-free multiagent reinforcement learning. Science. 2022 Dec 2;378(6623):990-996. doi: 10.1126/science.add4679. Epub 2022 Dec 1. PMID: 36454847.

[35] The XRSI Privacy and Safety Framework - XRSI – XR Safety Initiative. XRSI. (2022, February 16). Retrieved October 18, 2022, from https://xrsi.org/publication/the-xrsi-privacy-framework

[36] Rosenberg, L.B. (2022). Regulating the Metaverse, a Blueprint for the Future. In: De Paolis, L.T., Arpaia, P., Sacco, M. (eds) Extended Reality. XR Salento 2022. Lecture Notes in Computer Science, vol 13445. Springer, Cham. https://doi.org/10.1007/978-3-031-15546-8_23

[37] Brittan Heller and Avi Bar-Zeev, “The Problems with Immersive Advertising: In AR/VR, Nobody Knows You Are an Ad”, Journal of Online Trust and Safety, October 2021.

[38] Hughes, S. M., & Harrison, M. A. (2013). I like My Voice Better: Self-Enhancement Bias in Perceptions of Voice Attractiveness. Perception, 42(9), 941–949. https://doi.org/10.1068/p7526

[39] Higgins, D., Zibrek, K., Cabral, J., Egan, D. and McDonnell, R., 2022. Sympathy for the digital: Influence of synthetic voice on affinity, social presence and empathy for photorealistic virtual humans. Computers & Graphics, 104, pp.116-128.

[40] Rosenberg, Louis. (2022) The metaverse will be filled with 'elves', TechCrunch. Available at: https://techcrunch.com/2022/01/12/the-metaverse-will-be-filled-with-elves/