Human-Centered Metaverse and Digital Environments, Vol. 99, 2023, 122–130
https://doi.org/10.54941/ahfe1003938

The Metaverse as an AI-mediated Tool of Targeted Persuasion

Louis Rosenberg[1,2]

[1]Unanimous AI, Pismo Beach, CA 93448, USA
[2]Responsible Metaverse Alliance (RMA), Sydney, Australia

ABSTRACT

Over the next five to ten years, the metaverse will transform our digital lives from *flat media* viewed in the 3rd person to *immersive media* experienced in the 1st person. While the core technologies of virtual reality (VR) and augmented reality (AR) are not inherently dangerous to consumers, many advocacy groups and policymakers have raised concerns about the expansive surveillance capabilities that large platforms could deploy across populations. What is often overlooked, however, is how surveillance-related risks could be significantly amplified if metaverse platforms were also allowed to simultaneously deploy targeted promotional content as real-time immersive experiences. When considered in the context of Control Theory (CT), the pairing of *real-time surveillance* with *real-time influence* raises serious new concerns, for it could enable interactive platforms to become extremely efficient instruments for targeted deception, manipulation, and persuasion. These risks are further increased by the growing likelihood that AI-controlled avatars (i.e., Virtual Spokespeople) that look, speak, and behave like authentic human users will be used as a primary tool of targeted influence within immersive environments. For these reasons, policymakers must consider limiting the ability of metaverse platforms to deploy real-time immersive influence campaigns, especially when conversational AI technologies are utilized. As described herein, conversational AI paired with real-time behavioural and emotional monitoring could enable profoundly dangerous tools for deploying targeted influence campaigns at scale.

Keywords: Virtual reality, Augmented reality, Mixed reality, Conversational AI, Virtual spokespeople, Virtual product placements, Metaverse regulation, Large language models

INTRODUCTION

Across the fields of political science, economics, and philosophy, it is generwally believed that a well-functioning democracy must have a citizenry that possesses basic knowledge on issues of civic importance (BozDag et al. 2015; Caplan. 2008; Hardin, 2008). It is also widely believed that the population in a well-functioning democracy must have the freedom to reflect upon issues of political relevance and form beliefs without excessive outside influence (Coeckelbergh, 2022). The phrase “epistemic agency” refers to an individual’s control over his or her own beliefs (Gunn and Lynch, 2021). When citizens lack epistemic agency, democracy is threatened, as the political establishment can easily push widespread misinformation, disinformation, and propaganda, that distort societal beliefs and support authoritarian or totalitarian regimes (Coeckelbergh, 2022).

Mass media technologies are often abused to influence populations and weaken epistemic agency in hope of maximizing political control. This goes back to the printing press but was accelerated by modern media technologies such as radio and television. Over the last decade, the world was caught mostly unprepared for the unique threat to democracy caused by social media technologies. Despite early hopes that social media would support democracy by giving voice to the voiceless, the consensus in recent years is that social media has hurt democracies around the world by polarizing populations, spreading misinformation, and reducing trust in longstanding institutions (Aspen Institute, 2021; Rosenberg, 2022).

It is not just academics who find social media to be a damaging force in society. A recent poll by Pew Research (Auxier, 2020) 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 ironic 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 influence-based business models adopted by social media companies to bad actors using bots and other means to distort public discourse, I believe a major problem was the failure of regulators to realize that influence campaigns deployed via social media are inherently different than those deployed via classical media such as print, radio, and television. The difference is that social media is a bidirectional medium that allows for tracking, profiling, and targeting of sub-populations. This seemingly subtle difference has had a major impact, contributing to the polarization and radicalization of online communities.

With that background, it is deeply concerning that many regulators currently underestimate the potential dangers of the metaverse, categorizing the problems as equivalent to those currently encountered on social media platforms. In fact, there are some who view the metaverse as little more than a 3D version of today's social media platforms. And while they acknowledge that immersive content can be significantly more impactful (and therefore more harmful) than today's social media (Breves, 2021; Han et al. 2022) they fail to realize that the metaverse is not merely a bidirectional medium like social media, it is a real-time bidirectional medium which means it can impart closed-loop influence on target users (Rosenberg, 2022) They also fail to realize that with recent advances in Conversational AI such as Large Language Models like ChatGPT, metaverse platforms are increasingly likely to deploy influence campaigns using realistic Virtual Spokespeople (VSPs) that could adapt and optimize their promotional tactics in real-time based on behavioural and emotional monitoring of target users (Rosenberg, 2023).

This increases the concern that policymakers, who underestimated the increased risk that social media poses compared to traditional media, will also underestimate the increased risk that immersive media poses as compared to social media. As will be described below, the combination of real-time mass surveillance and real-time AI-mediated influence enabled by metaverse platforms could turn immersive worlds into the most dangerous tools of influence and persuasion ever created (Robertson, 2022). To raise awareness about the risks, researchers have written about dangers of metaverse technologies and the urgent need to protect basic human rights (Rosenberg, 2021; Rosenberg, 2022), but the unique risks of real-time interactive influence have not been conveyed in a form that has sufficiently motived regulators and policymakers. To address this, this paper uses basic ideas from Control Theory and frame emerging risks in a more rigorous way.

CONTROL THEORY AND METAVERSE RISKS

As policymakers plan the guardrails that can protect the public in immersive environments, it's important to consider the impact that a bidirectional real-time medium can have on epistemic agency. To help convey the risks, we can use the discipline of Control Theory, which formally represents how a "controller" can influence the behaviors of any interactive system. A classic example is a simple thermostat that regulates the temperature in a house. You set a temperature goal and if your house gets too cold, the heater turns on. If your house gets too hot, the heater turns off. When working as designed, the thermostat keeps your home very close to the temperature goal you defined. That's feedback control.

Referring to Fig. 1 above, the *System* being controlled is a house, the *Sensor* is a thermometer, and the *Controller* is a thermostat. An input signal called the *Reference* is the temperature the user sets as the goal. The goal is compared to the actual temperature in the house (i.e., *Measured Output*). The difference between the goal and measured temperature is fed into the thermostat which determines what the heater should do – turn on or turn off. This creates a real-time *feedback loop* that continually detects behaviors (e.g., temperature) and imparts influence (e.g., modulates the heat), to guide the system towards a desired goal. I give this background to help policymakers appreciate the feedback loops.

When considering "influence campaigns" (Glorin, 2022; Sedova et al., 2021; Waltzman 2022) in immersive worlds, *the system* being controlled is the user. That's because a user who puts on a headset is sinking themselves into a fabricated world controlled by a third party – an environment that has the potential to *act on the user* more than they *act upon it*. In the diagram above, the *System Input* to the user are the immersive sights, sounds, and touch sensations that are fed into the user's eyes, ears, hands, and body via interface devices. There is also an arrow labeled *System Output*. In the simple

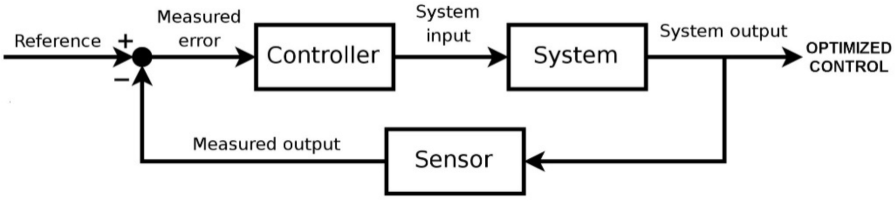

Figure 1: Basic Control System Diagram.

thermostat example, the output is the real-time temperature of the house. In the metaverse example, the output is real-time actions and reactions of the human user.

This brings us to the *Sensor* box in the diagram above. In the metaverse, an extensive array of sensors will track the user in real-time – the motions of their hands, head, and body, including the direction they're looking, the dilation of their pupils, the changes in their posture and gait – even their biometrics (i.e., vital signs) are likely to be tracked including respiration rate, heart rate and blood pressure. Already, commercially available headsets like the Meta Quest Pro can track facial expressions and eye motions. Some contend this is not a privacy concern because facial expressions and eye motions are easily visible to any person nearby. This is a misconception. That's because AI based tracking systems can sense eye motions and facial expressions that are not perceptible to human observers. This includes subconscious expressions that are too fast or subtle for human observers to detect. Known as "micro-expressions," these faint changes can convey emotions that users did intend to express and are unaware of revealing.

In addition to tracking user behaviors, AI technologies already exist that can infer emotions from user posture, facial expressions, vocal inflections, and eye motions (Heller & Bar-Zeev 2021). Other AI technologies already exist to detect emotions from the blood-flow patterns on your face and the vital signs detected from sensors in your earbuds (Benitez-Quiroz et al., 2018). This means that when a user immerses themselves in a virtual or augmented environment, sensors will be able to track almost everything that user does and says while also assessing what that user feels (emotionally) during each of such actions and reactions.

Based on the paragraphs above, we can update the system diagram as shown in Figure 2 in which we replace the *System* with the *Human User* and replace the *Sensor* with the broad abilities of *Metaverse* platforms to directly track or indirectly infer a user's behaviors and emotions.

It is important to note that extensive behavioral and emotional data tracked by metaverse platforms are likely to be stored over time unless regulators prevent this. Data collected in this way could be extensive, documenting the behaviors and emotions of individual users over periods of days, weeks, months or years. Even worse, this data could be processed with machine learning (ML) technology to build behavioral models and emotional models that predict how individuals will react within a wide range of circumstances.

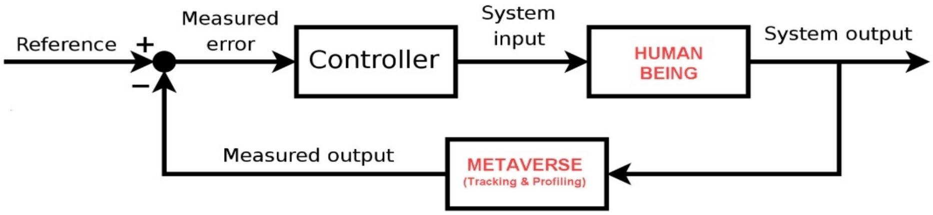

Figure 2: Control System Diagram for Metaverse Environments.

In this context, we must consider the primary risk from metaverse platforms as their ability to (i) track users during their daily life, (ii) profile user behaviors and emotions during thousands of interactions each day, (iii) process this data using ML to create behavioral and emotional models, and (iv) use these behavioral and emotional models to predict the actions and reactions of users in response to injected virtual content. And this won't just happen in fully virtual worlds but also in augmented worlds. This means that the tracking, storing, profiling and modeling of behaviors and emotions could occur throughout our daily life (Rosenberg 2021).

Of course, the threat from metaverse platforms is not just their ability to track and profile users, but how this data could be used to impart targeted influence. This brings us to the *Controller* element of the system in Fig. 2 above. As shown, the controller receives a *Measured Error* as input, which is the difference between a *Reference Input* (i.e., desired behaviors) and the *Measured Output* (i.e., sensed behaviors). When assessing influence campaigns, the *Reference Input* represents the "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 services or it could be a *state actor* that wishes to drive users towards pieces of propaganda, misinformation, ideology, or outright lies. In either scenario, we can update the Control System diagram by replacing the word Reference with the *Agenda* of the third-party influencer as shown in Fig. 3.

Finally, we must address the *Controller* element. Its function is to reduce the error between the *Reference Input* and the *Measured Output* (i.e., the difference between what you want the system to do and what the sensors report the system is currently doing). While the controller can be as simple as a thermostat, it also can be quite complex. For example, self-driving vehicles use AI-controllers to navigate busy traffic, achieving difficult objectives in rapidly changing environments. In the metaverse, the controller's goal could be to impart targeted influence on a human user. The controller could do this through real-time feedback loops that continually adjusts its persuasive tactics based on the sensed behaviors and emotions of the user, gradually guiding that user towards the desired agenda.

Consider this example: a user sits in a coffeehouse in the metaverse. A third party aims to influence that user about a product, service, or piece of messaging, propaganda, or disinformation. The controller pursues this goal by

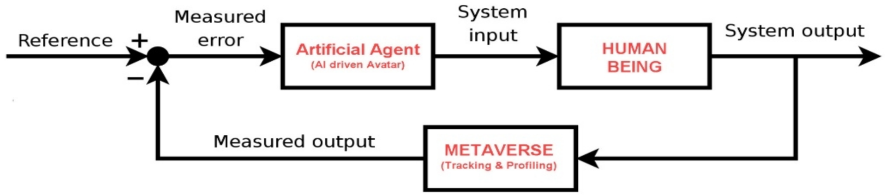

Figure 3: Control System Diagram with Third Party Agenda.

injecting virtual content into the users' surroundings (Rosenberg, 2021). This content could take the form of Virtual Product Placements (VPPs) or Virtual Spokespeople (VSPs). It could be so realistic it is indistinguishable from non-promotional elements in the virtual or augmented world. For example, the controller may create a *virtual couple* sitting at a nearby table. The look of the couple is custom crafted for maximum impact. This means the age, ethnicity, gender, clothing, hair style, speaking style, and other qualities will be selected by AI algorithms for optimal persuasion of the target user. The controller will then cause the virtual couple to have a conversation among themselves, *deliberately within earshot* of the target. The overheard dialog will support the promotional agenda set by the third party. And because the virtual couple could appear highly realistic, the target user may have no idea that he or she is not overhearing the genuine discussion among other patrons.

Thus, when the virtual couple starts discussing *how pleased they are* with a new car they recently purchased, the target may believe he is overhearing *authentic sentiments*, not a targeted advertising campaign (Rosenberg, 2022). Even worse, as the virtual couple continues their conversation, the controller could monitor the target user in real-time, assessing his or her facial expressions, pupil dilation, and blood pressure to detect real-time emotional reactions. If the user's pupils dilate when the couple talks about the car's impressive horsepower, the controller could cause the AI-generated dialog to focus on vehicle performance. Or if the user's blood pressure rises when the couple talks about the car's self-driving features, the controller could adjust to avoid that topic further. In this way, the target user is an unwitting participant in a manipulative *feedback control system* that optimizes targeted impact. (Rosenberg 2005, Rosenberg 2021). Although disturbing, this example merely promotes a new car. The same tactics could just as easily promote extreme ideology, radical propaganda, or disinformation. Deployed at scale, such methods could compromise the epistemic agency of large populations.

The scenario above targets users as silent observers. More aggressive tactics will target users with AI-mediated avatars that engage users in agenda-driven conversation. These conversational agents (i.e., Virtual Spokespeople) will draw users into seemingly friendly dialog. The visual appearance of the VSP will be optimized for maximum impact. This could involve traditional media tactics such as sexualizing the VSP in ways that are directly appealing to target users. More devious methods may include blending aspects of a target user's own facial features into the VSP's face. This is because the target user is likely to be more favorably disposed to conversational agents that incorporate some aspects of his or her own features. Research conducted at Stanford University has shown that a viewer's features can be blended into the faces of political candidates without the viewer noticing. And yet, the target is 20% more likely to vote for the candidate as a consequence of the manipulation (Bailenson, 2008). Additional research suggests that avatars that mimic a target 's facial expressions, head motions or body gestures can be more influential on users (Bailenson & Yee, 2005; Chartrand et al., 1999). Unless explicitly banned, these highly manipulative techniques are likely to be deployed by corporate and state actors at scale (Rosenberg, 2023).

Again, the AI-controlled avatars will look, act, and speak like authentic human users. To demonstrate how close we are to systems that can produce realistic and persuasive dialog in real-time, the ChatGPT generative AI system (circa 2022) was engaged and and asked for advice on how to persuade people 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 response above was generated by an AI system that had no data about the target user. Soon, platforms will be able to train LLMs on profile data for specific individuals, tailoring responses to their unique interests, values, and traits. Such methods could be used to sell products, or to conversationally convey disinformation, extreme ideology, or hate. In addition, the controller could have access to real-time emotional data, including the facial expressions, vocal inflections, body posture, pupil dilation, facial blood patterns, and vital signs of the target. The AI-driven controller could use these deeply personal cues to optimize conversational tactics in real-time. This sounds dystopian but without regulation, it will be a deployable technology in the very near future.

In addition, AI controllers will likely include modules trained on sales tactics and other forms of human persuasion. AI technologies can already outplay the best human competitors at chess, go, and other complex games. In 2022, DeepMind used a system called DeepNash to show for the first time that an AI could learn to bluff human players and sacrifice gaming pieces for the sake of long-term strategy (Perolat et al., 2022). From that perspective, a typical consumer will be highly vulnerable to persuasion when engaged in strategic conversation with AI-agents.

To complete the system diagrams (see Fig. 4), we can replace the word *controller* with *AI Agents* that alter a user’s surroundings and/or engages users in conversation. And while AI agents are envisioned as human avatars, non-human characters will also be used. This is particular dangerous for

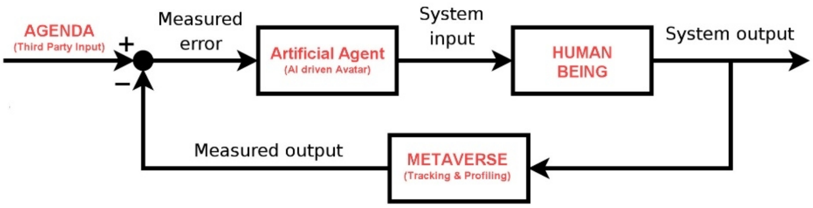

Figure 4: Dystopian Control System deployed in an AI-moderated Metaverse.

children, who could be targeted by cute characters that deploy an AI-driven promotional agenda.

CONCLUSION

Virtual and augmented worlds (i.e., the metaverse) will likely make extensive use of *generative AI* to adapt immersive content and *conversational AI* to engage users in interactive dialog. This combination of technologies can easily be used to create *feedback-control systems* that maximize real-time persuasion. The techniques will likely include the use of Virtual Spokespeople that look, sound, and act like human users and are designed to meet the goals of third-party sponsors or state actors. This could significantly impact the cognitive liberty of target users, compromising their epistemic agency. This is not just a risk to individuals, but also a societal risk that could impact the foundations of democracy. For these reasons, policymakers must consider aggressive and meaningful actions aimed at protecting populations from abuse of AI-driven experiences in immersive environments.

REFERENCES

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

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

Bailenson, J. 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

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

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.

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

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

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

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

Commission on Information Disorder Final Report, Nov 2021. Aspen Institute.

Glorin, Sebastian. (2022). A Study on Metaverse Awareness, Cyber Risks, and Steps for Increased Adoption. International Journal of Information Security and Privacy.

Gunn, H., Lynch, M. P.: The internet and epistemic agency. In: Lackey, J. (ed.) Applied epistemology, pp. 389–409. Oxford University Press, Oxford (2021).

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.

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

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

Perolat J, et al. 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.

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

Rosenberg, L (2022) Marketing in the Metaverse: A fundamental shift., Future of Marketing Institute. August 15, 2022. DOI: 10.13140/RG.2.2.35340.80003

Rosenberg, L. (2005) Methods and Apparaturs for Conversational Advertising - U. S. Patent Application No. 60/689, 301, filed Jun. 10, 2005.

Rosenberg, L. (2021) *Metaverse: Augmented reality pioneer warns it could be far worse than social media,* Big Think. Nov 6, 2021. Available at: https://bigthink.com/the-future/metaverse-augmented-reality-danger/

Rosenberg, L. (2022) Deception vs authenticity: Why the metaverse will change marketing forever. VentureBeat. August 21, 2021. VenureBeat.com.

Rosenberg, L. (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

Rosenberg, L. (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

Rosenberg, L. (2022). 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/

Rosenberg, L., (2022) "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.

Rosenberg, Louis (2023) "The Metaverse: The Ultimate Tool of Persuasion." Metaverse Applications for New Business Models and Disruptive Innovation, edited by Muhammad Anshari, et al., IGI Global, 2023, pp. 1–11. https://doi.org/10.4018/978-1-6684-6097-9.ch001

Sedova, Katerina. Christine McNeill, Aurora Johnson, Aditi Joshi, and Ido Wulkan, "AI and the Future of Disinformation Campaigns" (Center for Security and Emerging Technology, December 2021). https://doi.org/10.51593/2021CA011

Waltzman, Rand. "The Role of Today's VRE and Considerations for Cognitive Warfare." NATO - Allied Command Transformation, NATO, 18 Nov. 2022, https://doi.org/10.1109/UEMCON54665.2022.9965661 https://www.act.nato.int/articles/cognitive-warfare-considerations