Dehumanized Avatars: Unethical Behavior in the Metaverse

Carina Castagna, Sercan Demir, Markus Weinmann
University of Cologne, Cologne, Germany
castagna@wiso.uni-koeln.de, demir@wiso.uni-koeln.de, weinmann@wiso.uni-koeln.de

***Abstract*—Despite the rise of the metaverse, surprisingly, little research has explored the effects of immersion within the metaverse on individual user behavior. This study begins to fill this gap by examining the effect of dehumanized avatar identities—making avatars look less human-like—and its consequences for ethical behavior. We conducted an online experiment (N=130) to test if an avatar's level of dehumanization increases unethical behavior. Our study assessed three different levels of dehumanization (human-like vs. animal-like vs. abstract avatars). The results indicate that participants' willingness to cheat increases when engaging through an abstract and more dehumanized avatar. This study reveals the hidden risks of avatars in the metaverse and sheds light on the delicate relationship between unethical behavior and avatar dehumanization.**

Keywords—digital self, avatars, ethical behavior, dehumanization, metaverse

I. INTRODUCTION

The metaverse is expected to grow rapidly within the next years[1], so understanding users' behaviors within the metaverse is of utmost importance for companies and organizations. Users typically interact by choosing an avatar representing them in the digital sphere, also known as (digital) embodiment [1], [2]. In other words, they hide behind a digital self while interacting with others.

Hiding behind a digital self, however, may lead to unintended consequences; for example, the enhanced state of anonymity may increase unethical behavior such as cheating. Thus, this paper explores one unique form of embodiment in digital self-representations, anthropomorphism levels, and how this affects unethical behavior. We define a high anthropomorphized avatar as those who typically maintain a human-like form and low anthropomorphized avatars with dehumanized features (e.g., abstract, robotic, animal forms). Following the mechanistically dehumanized discussion [3], we raise the research question of whether ethical values found in the real world still apply to the virtual world for dehumanized avatars (i.e., low anthropomorphism). In particular, we study how ethical choices are affected by anthropomorphism levels of the digital embodiment of the self.

The present research starts to answer this important question by bridging the literature on ethical behavior and anthropomorphism to put forth the idea that when digital representations are less anthropomorphized, users may have more predisposition to cheat. Specifically, in this working paper, we begin to answer those questions by applying an initial online experiment (N = 130 participants) to show how constructions and definitions of the self (for example, as an avatar) may impact ethical behavior in digital environments. Specifically, we show that digital-self representations with lower levels of anthropomorphism—more abstract or animal-like avatars—lead users to cheat more.

In documenting these effects, this research makes significant contributions to theory and practice. To the best of our knowledge, this study is among the first to experimentally manipulate and test how avatars and their different anthropomorphism levels affect ethical behavior and decision-making. Next, we theorize about digital embodiment of the self, anthropomorphism, and cheating behavior. We then present the study that tests our prediction. The paper ends with a discussion of the theoretical implications of our findings.

II. THEORETICAL BACKGROUND AND HYPOTHESIS

We draw upon two distinct research streams: (1) Avatars and the digital embodiment of the self, and (2) Dehumanization and cheating behavior. We have chosen avatars as a study context because they are well-studied regarding their effect on users' decision-making [4], [5]. Additionally, users are increasingly interacting with avatars making decisions in the digital environment, and extending their identities to digital environments [1], [5]. This makes avatars a practically relevant and timely topic to study. Furthermore, as we are interested in the effect of dehumanization on online cheating, avatars are ideal for incorporating the concept of anthropomorphism, as avatars are malleable regarding their external appearance and individual styling.

A. Avatars and Digital Embodiment of the Self

Avatars are graphical representations of real-world users in a digital environment. They can be divided into generic and customized avatars [6]. While generic avatars are pre-defined, customized avatars are malleable and can be changed according to one's preferences. Our study builds on the extensive literature on avatar research [6].

A vast amount of literature found that users' decision-making in a digital environment is affected by an avatar's main appearance [7]. In this sense, users can extend themselves by using a digital representation (i.e., an avatar), which can also affect real-world decisions. Interestingly,

[1] https://www.statista.com/statistics/1288048/united-states-adults-reasons-for-joining-the-metaverse/

when people create an avatar, it can develop new constructions and definitions of themselves [1]. While users can alternate their avatar body, selecting and re-configuring its image, they feel free to become their ideal selves. For instance, people can have a new gender identity online or represent their aspirational self-image [8]. The idea of (digital) embodiment [1] explains changes in decision-making, as consumers are free to create new identities in the digital world, re-creating their selves and, consequently, changing their behavior.

B. Dehumanization and Cheating Behavior

Individuals' offline behavior can be affected by even the smallest changes between their real bodies and virtual selves. However, studies show that different levels of avatar anthropomorphism—making avatars look more human-like by attributing human-like qualities [9]—can diminish this impact, thus increasing the similarity between the virtual and offline behavior [10]. For instance, avatars wearing a user's real face increase their identification level with this avatar [11], which can lead to consequences in the real world. For example, people, who observed their own real faces on an avatar, had a significantly higher willingness to quit smoking, in comparison with the other-avatar condition [10]. While the relation between anthropomorphism and avatars is optimistic, dehumanized characteristics on avatars are increasing within the metaverse context. It is becoming more and more popular to create avatars with lower levels of anthropomorphism; we call this phenomenon dehumanization, which is the attribution of animal or object traits to oneself (i.e., avatars with geometric forms) [12].

Most previous literature focuses on making avatars more human-like, that is, increasing the levels of anthropomorphism; however, less research focuses on the opposite effect, that is, making avatars look less human-like, for example, the consequences of using animal-like or abstract avatars [12]. This effect is also referred to as "dehumanization," that is, removing human traits. Removing human traits might also have unintended consequences; for example, people acting via a dehumanized avatar might also act less morally or ethically. Given that users are typically free to choose any type of avatar in most virtual settings—human-like, animal-like or even something else—we investigate how dehumanizing avatars affects real behavior. Specifically, we will focus on ethical behavior for two reasons: (1) ethical behavior is usually associated with humans, and using dehumanized avatars may also lead to less ethical behavior; (2) understanding how avatar choice influences ethical behavior may help to design better avatars, which might lead to less cheating in virtual worlds.

Online cheating has been shown to be a major problem in digital environments [13]. Cheating comprises conscious unethical behavior, such as lying or manipulating performance standards to promote self-interests [14]. Cheating behavior can increase when users are behind their digitally reembodied selves. Previous research explains this effect as users feel more anonymous and lack accountability when being in a digital environment [15].

Thus, in the same way, that an anthropomorphized avatar endorses healthy offline habits [10], we posit that dehumanized avatars can possibly affect people's ethical behavior. Specifically, we predict that people can hide behind an avatar; and that the more dehumanized an avatar look, the more users tend to act less human-like. That is, moral standards go down, which leads users with dehumanized avatars cheating more. Therefore, we formally hypothesize:

Hypothesis 1 (H1): *Users exposed to more (less) dehumanized avatars cheat more (less).*

Our hypothesis highlights a significant contribution to avatars and ethical behavior discussion. Although previous studies applied avatars to verify ethical behavior, these investigations typically demonstrate the manipulation aspects of anthropomorphism and not dehumanization [16], [17]. Further, no manipulation of dehumanization was associated with unethical behavior, as most of the studies are also related to man vs. machine, that is, the comparison between AI and human interaction [18]. Another example is [19], who proposed to study realistic vs. unrealistic avatars. The authors manipulated both conditions using human features rather than geometric or animal features.

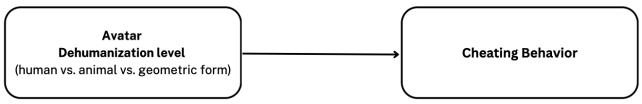

Fig. 1: Conceptual framework

III. METHOD

A. Material and Procedure

We conducted a between-group online experiment to test our hypothesis. We randomly assigned participants to one of three avatar conditions, only differing in the level of dehumanization: (a) the human-like condition, in which the avatar looked like a human and the level of dehumanization was relatively low, (b) the animal condition, in which the avatar was a dog, and the level of dehumanization was relatively moderate and (c) the abstract condition, in which the avatar was a white geometric form and the level of dehumanization was relatively high (see also Figure 2). All material was presented via Qualtrics.

We told participants that they were part of a metaverse study. After participants were randomly assigned to one of the three avatar conditions, we let them play the mind game [20], [21]. In this game, participants were asked to think of a number between one and six (the range of a die). Afterward, on a new page, a random numbers generator rolled a six-sided die. Participants were then asked to report "yes" if the number they thought of matched the

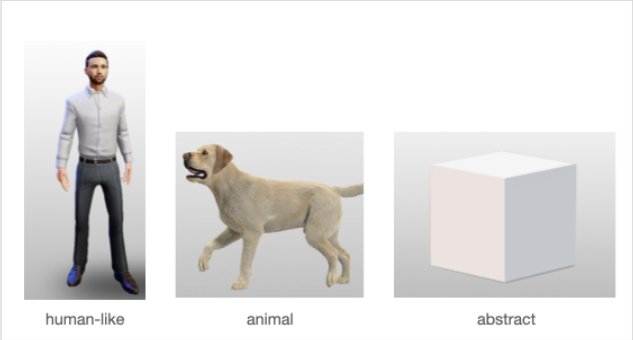

Fig. 2: Treatment conditions

die outcome, otherwise "no." To incentivize cheating, participants received £1 when they reported "yes," and £0 if they reported "no." At the end of the experiment, we gathered demographic data and debriefed the participants.

B. Participants

We used Prolific (https://prolific.co) to recruit participants who are older than 18 years and reside in the US. In total, we recruited 150 participants. Twenty participants did not pass an attention check, resulting in a final sample size of $N = 130$ (age = 40.3 years, 44% female). We paid all participants £1 (about US$1.21) for a 7-minute task, a payment equivalent to an hourly wage of £9.11 (US$11). In addition, they received a bonus depending on their decisions made in the experiment (described above).

C. Measures and Model Specification

The purpose of our analysis is to compare reported matches—i.e., the number people thought of and the results of the random numbers generator—across conditions. According to our theorizing, we expect more matches in more dehumanized conditions, that is, individuals cheat more when they are exposed to animal-like or abstract avatars. Our dependent variable was binary, that is if an individual indicated a match (=1) or not (=0). Hence, we specified a logistic regression model to estimate the effect of being in a particular condition—human-like vs. animal vs. abstract—on a participant's likelihood to report a match:

$$Pr(\text{match}_i = 1) = logit(p_i) = \alpha + \beta \times \text{condition}_\text{i}, \quad (1)$$

where α is the intercept representing reported matches in the human-like condition, which also served as baseline; β represents the differences of matches reported in the other conditions, either the animal or abstract condition. According to our hypothesis, we expected β to be positive and significant for the animal and abstract conditions. That is, participants report more matches in those conditions compared to the human-like condition.

IV. RESULTS

A. Descriptive Statistics

Figure 3 presents the descriptive statistics by condition. In the human condition, 47.91% of the participants reported a match. This indicates cheating, given an expected value of 16.67% (the probability that a participant's report matches the random numbers generator is 1 out of 6). In the other condition, the share was even higher: 63.26% in the animal condition, and 66.67% in the abstract condition reported a match.

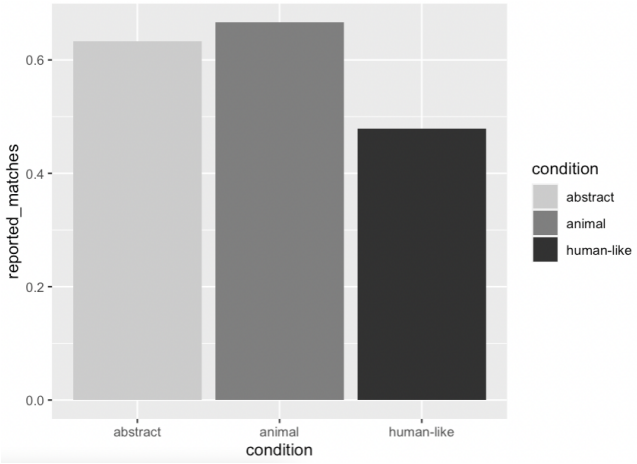

Fig. 3: Conditions and respective cheating share

B. Estimation Results

Table I presents our estimation results. Coefficients are presented on an odds scale. The results show that the odds of reporting a match in the abstract condition are 2.21 times higher than in the human condition (the reference group). Likewise, the odds of reporting a match in the animal condition are 2.13 times higher. Both coefficients are positive and significant, thus supporting our hypothesis.

TABLE I: Logistic Regression Results (Odds-Ratio)

	Dependent variable:
	Binary: reported match (yes = 1)
Abstract condition	2.21*
	(0.88–5.76)
Animal condition	2.13*
	(0.93–5.02)
Intercept	0.41
	(0.09–1.78)
Controls	yes
Observations	130
Log Likelihood	−84.316
Akaike Inf. Crit.	178.631
Notes:	*p<0.1; **p<0.05; ***p<0.01
	We used age and gender as controls.
	Confidence intervals in parentheses.

V. GENERAL DISCUSSION

Summary. The present research sheds light on the role of digital self-representations (i.e., avatars) on ethical

behavior. Specifically, we examined how different levels of dehumanized avatars—human-like, animal-like, abstract—can increase cheating behavior.

Theoretical contributions. The present research makes primary contributions to the disposal literature. First, we provide clear and causal evidence that users are more willing to cheat within a digital context when acting through a dehumanized avatar. Specifically, by manipulating three different levels of dehumanization (human-like vs. animal-like vs. abstract avatars), we found that users acting through a dehumanized avatar cheat more in 20% of the cases. Second, our results demonstrate that participants in a human-like (i.e., anthropomorphic avatar) condition act less dishonestly (approx. 16% less) than when assigned to an abstract or animal avatar. We contribute to the digital self-extension theory [1] by showing how the embodiment of a self on a dehumanized level can affect ethical behavior. The effect of anonymity on unethical behavior [22] can be even more substantial on the metaverse as users create and identify their selves with a new digital self, diminishing their ethical values.

We also contribute to avatar research [10], [11] by showing that anthropomorphism and dehumanized characteristics can influence ethical behavior. Dehumanized avatars (i.e., low anthropomorphism level) appear to boost unethical behavior. Thus, our results add to the novel research vain of dehumanization and digital selves [3], [7].

Limitations and future research. This research has some limitations that offer avenues for future research. In this study, we relied mainly on an online experiment, manipulating generic avatars, which participants could not actually customize. Despite manipulating a fictitious metaverse scenario, our online study was not tested on a real digital platform. Another limitation is the use of only one gender in avatar manipulation. Thus, future work may wish to change how avatars' pictures are displayed on the manipulation, emphasizing the avatar as the participant's identity.

Thus, an additional study could examine whether different genders impact our main effect. Further, the salience of the avatar image could also be increased by keeping it visible through each decision stage, thereby increasing the manipulation impact. This can be executed in two ways: First, by presenting avatar images during the mind game stage. Second, the avatar presence may also be highlighted by increasing the experiment's sense of reality. For instance, future research could examine the impact of digital self-representations on dishonest behavior by applying Virtual Reality scenarios. As users are able to customize avatars according to their own preferences, a personalized avatar could act as an extended self and thus could be strengthening the effect. Future research could further examine whether ethical behavior is affected by underlying processes such as socioeconomic vulnerability and personal control.

Managerial implications. Finally, our findings bring important implications for the design of digital environments for several stakeholder groups. As our results showcase the potential unethical behaviors of digital-self representations that are distant from the human form, we invite behavioral researchers and practitioners to consider how interventions might be used to improve ethical behavior in the metaverse. As the immersion in the metaverse may daze boundaries between the real world and the virtual world, we highlight the importance of ethical procedures during the primary phase of virtual worlds, the avatar creation and self-identification. In this sense, our results featured the need for ethical guidance associated with avatars' dehumanization level, increasing cheating probabilities. For instance, educative programs in order to increase users' identification with avatars, despite their dehumanized form, shall decrease this phenomenon. Further, companies may want to nudge metaverse users to consider the unethical cues that dehumanized avatars can portend for their virtual network. Cheating behavior could also be improved with the development of ownership nudges, making users more accountable for their actions even on dehumanized avatars. Our findings also suggest that consumers might want to consider shifting their avatars into a more anthropomorphized form to collaborate in an ethical digital environment. A universal ethical user certification could also be developed in order to create individual accountability and credibility online. Finally, we urge that ethical considerations and feasible solutions need to be discussed and built into the first step of the metaverse immersion, which has also shown to be one of the most important: the development and identification with an avatar.

VI. CONCLUSION

In this paper, we hypothesized and tested the effect of dehumanized avatars on online cheating behavior. In particular, results show that users are generally more willing to cheat when acting through an avatar in a digital context. A more dehumanized avatar (animal or geometric form) leads to more cheating behavior.

REFERENCES

[1] Belk, Russell W., "Extended self in a digital world," *Journal of Consumer Research*, vol. 40, no. 3, pp. 477–500, 2013.

[2] ——, "Possessions and the extended self," *Journal of Consumer Research*, vol. 15, no. 2, pp. 139–168, 1988.

[3] N. Castelo and D. R. Lehmann, "Be careful what you wish for: Unintended consequences of increasing reliance on technology," *Journal of Marketing Behavior*, vol. 4, no. 1, pp. 31–42, 2019.

[4] M. Seymour, K. Riemer, and J. Kay, "Actors, avatars and agents: Potentials and implications of natural face technology for the creation of realistic visual presence," *Journal of the Association for Information Systems*, pp. 953–981, 2018.

[5] A. Davis, J. Murphy, University of Nebraska at Omaha, D. Owens, D. Khazanchi, and I. Zigurs, "Avatars, people, and virtual worlds: Foundations for research in metaverses," *Journal of the Association for Information Systems*, vol. 10, no. 2, pp. 90–117, 2009.

[6] H. M. Aljaroodi, M. T. P. Adam, R. Chiong, and T. Teubner, "Avatars and embodied agents in experimental information systems research: A systematic review and conceptual framework," *Australasian Journal of Information Systems*, vol. 23, pp. 1–37, 2019.

[7] N. Yee, J. N. Bailenson, and N. Ducheneaut, "The proteus effect: Implications of transformed digital self-representation on online and offline behavior," *Communication Research*, vol. 36, no. 2, pp. 285–312, 2009.

[8] J. Martin, "Use-Value, Exchange-Value, and the Role of Virtual Goods in Second Life."

[9] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: A three-factor theory of anthropomorphism." *Psychological Review*, vol. 114, no. 4, pp. 864–886, 2007.

[10] H. Song, J. Kim, R. J. Kwon, and Y. Jung, "Anti-smoking educational game using avatars as visualized possible selves," *Computers in Human Behavior*, vol. 29, no. 5, pp. 2029–2036, 2013.

[11] Y. Seo, M. Kim, Y. Jung, and D. Lee, "Avatar face recognition and self-presence," *Computers in Human Behavior*, vol. 69, pp. 120–127, 2017.

[12] K. Karanika and M. K. Hogg, "Self–object relationships in consumers' spontaneous metaphors of anthropomorphism, zoomorphism, and dehumanization," *Journal of Business Research*, vol. 109, pp. 15–25, Mar. 2020.

[13] M. Weinmann, J. S. Valacich, C. Schneider, J. L. Jenkins, and M. Hibbeln, "The path of the righteous: Using trace data to understand fraud decisions in real time," *MIS Quarterly*, vol. 46, no. 4, pp. 2317–2336, 2022.

[14] M. S. Mitchell, M. L. Ambrose, R. Folger, M. D. Baer, and N. F. Palmer, "Cheating Under Pressure: A Self-Protection Model of Workplace Cheating Behavior," *Journal of Applied Psychology*, vol. 103, no. 1, pp. 54–73, 2018.

[15] J. Suler, "The online disinhibition effect," *CyberPsychology & Behavior*, vol. 7, no. 3, pp. 321–326, 2004.

[16] J. F. Nunamaker, D. C. Derrick, A. C. Elkins, J. K. Burgoon, and M. W. Patton, "Embodied Conversational Agent-Based Kiosk for Automated Interviewing," *Journal of Management Information Systems*, vol. 28, no. 1, pp. 17–48, Jul. 2011. [Online]. Available: https://www.tandfonline.com/doi/full/10.2753/MIS0742-1222280102

[17] R. M. Schuetzler, G. M. Grimes, and J. S. Giboney, "The effect of conversational agent skill on user behavior during deception," *Computers in Human Behavior*, vol. 97, pp. 250–259, Aug. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0747563219301311

[18] M. D. Pickard, C. A. Roster, and Y. Chen, "Revealing sensitive information in personal interviews: Is self-disclosure easier with humans or avatars and under what conditions?" *Computers in Human Behavior*, vol. 65, pp. 23–30, Dec. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0747563216305684

[19] M. Seymour, L. Yuan, A. R. Dennis, and K. Riemer, "Have We Crossed the Uncanny Valley? Understanding Affinity, Trustworthiness, and Preference for Realistic Digital Humans in Immersive Environments," *Journal of the Association for Information Systems*, vol. 22, no. 3, pp. 591–617, 2021. [Online]. Available: https://aisel.aisnet.org/jais/vol22/iss3/9/

[20] A. Kajackaite and U. Gneezy, "Incentives and cheating," *Games and Economic Behavior*, vol. 102, pp. 433–444, 2017.

[21] T. Jiang, "Cheating in mind games: The subtlety of rules matters," *Journal of Economic Behavior & Organization*, vol. 93, pp. 328–336, 2013.

[22] K. C. Yam and S. J. Reynolds, "The effects of victim anonymity on unethical behavior," *Journal of Business Ethics*, vol. 136, no. 1, pp. 13–22, 2016.