Neural Avatars and Adaptive Systems: Behavioural Dynamics and Ethical Governance in the Metaverse

Gabriel Silva Atencio ✉

Latin American University of Science and Technology (ULACIT), San José, Costa Rica.

Abstract

This research analyzes how artificial intelligence (AI) can be used to further personalize things in the metaverse. To do so, it examines brain models and configurable settings from the perspective of government ethics, human-computer interaction, and behavioral psychology. Biological tracking, psychological testing, behavior monitoring, and qualitative conversations are some of the methods used in the study to demonstrate that making avatars more unique makes people much more interested in the game. It was found that men and women had different tastes. For example, the women who participated valued utility more than artistic reality. The use of operant conditioning to guide adaptive changes led to a 42% increase in dwell time but also raised moral questions. The Integrated Metaverse Engagement Model (IMEM) under consideration combines information ethics with the idea of social identities. It has design rules aimed at reducing computational bias and making the model compliant with standards such as the General Data Protection Regulation (GDPR). AI Fairness 360 (AIF360) and Random Forest classification are two of the many methods used in the study. Modern design interfaces such as YOLOv8 with Squeeze-and-Excitation blocks and ConvNeXt-SE-attn models are also used to gain better insight into behavior sequences. The results provide developers and policymakers with evidence-based insights. They also contribute to the academic debate on digital identity and offer moral guidelines that can be used when deploying AI in virtual environments.

Keywords

Adaptive environments, AI ethics, Behavioral modeling, Metaverse, Neural avatars, Personalization.

Introduction

The metaverse is fundamentally changing the way people interact with computers. With the advancement of artificial intelligence (AI) and augmented reality (AR), this network of precise virtual locations will remain and be accessible to everyone. A large part of this combined digital and physical environment is AI-driven customization. It is possible to create neural avatars, which are digital versions of real people created using architectures such as Generative Adversarial Networks (GANs) and Transformers. It is also possible to create adaptive environments that change based on user behavior using reinforcement learning (RL) [1-4]. Liu [5], McDowell [6] say that early studies suggest these technologies could make it easier to connect with others and get around. However, they are being introduced rapidly before modern society fully understand how they will change behavior over time, how they can be used across cultures, and what moral risks they pose, particularly in relation to mental health, algorithmic bias, and constant surveillance [7-9].

A close examination of the research that has already been conducted shows that there are four gaps that recur and are interrelated. First, most scientific findings are cross-sectional and focus on the West [10]. It does not address how people's opinions change over time and do not provide sufficient examples of what people from different countries like. Second, theoretical models do not always combine well with each other. For example, computer science, behavioral psychology, and information ethics do not always combine well. This could lead to systems that are designed to involve people but may compromise the freedom and digital rights of users [11, 12]. Third, laws and regulations such as the General Data Protection Regulation (GDPR) [13] and the European Union (EU) AI Act [14] establish foundations but are not particularly well suited to controlling real-time behavioral change or ensuring that algorithms function well in places with large amounts of data. Fourth, many different approaches are based on only one type of study. Behavior modeling that uses only RL may not have the level of detail found in mixed frameworks that combine sequential

deep learning models (such as BiLSTM) with strong ensemble classifiers for better pattern recognition and understanding [15].

The Integrated Metaverse Engagement Model (IMEM) is created, put into action, and evaluated in this study to deal with these complex problems. This way of thinking combines operant training [6], social identity theory [16, 17], and information ethics principles [18] in a planned way to help make personalized systems that are fun to use and follow the rules of ethics. The research uses a linear explanatory mixed-methods approach to gather and compare multimodal data from participants from selected regions of 1,200 current Metaverse users. Quantitative measures include synced biological data (from Empatica E4 and Tobii eye-tracking), behavioral monitoring (from OptiTrack motion capture), and psychological tests, such as the Patient Health Questionnaire-9 (PHQ-9) and the Metaverse Engagement Index [19]. Focus groups and semi-structured conversations are effective ways to get qualitative feedback. A machine learning pipeline with the Random Forest classification tool for figuring out which features are important, the AI Fairness 360 (AIF360) toolkit for checking for bias, and new model integrations (YOLOv8 with SE blocks and ConvNeXt-SE-attn) for advanced object detection and temporal behavior analysis [20] make sure that the analysis is correct.

There are three main things that the study adds. As a first step, it shows that neural avatar customization (β =.34, p <.001) and AI-responsive environmental adaptations (42% increase in dwell time; omnibus ANOVA: $F(2, 1197) = 47.2$, $p < .001$, $\eta^2 = 0.18$) are two main factors that keep users interested. It also shows that gender and non-binary users have different preferences ($F(2,1197) = 5.21$, $p = .006$). Second, it shows that ethically informed technological interventions work by showing measurable decreases in algorithmic bias (parity improvement from -0.15 to -0.04, p=.03) and a 20% drop in self-reported distress (PHQ-9) among users exposed to systems that reduced bias. Users who used systems that reduced bias also reported more trust in explainable AI (XAI) interfaces (Δ+1.5 points, p<.01). Third, it sets a standard for how to do research in virtual environments by showing how to make a pipeline that combines advanced computer vision with sequential data modeling and fairness-aware analytics. This pipeline can be used again.

This study goes beyond showing how technology works to provide a solid base for the responsible growth of human-centered Metaverse ecosystems. To this end, it focuses on scientific accuracy, moral leadership, and the combination of ideas from different fields. The following sections delve deeper into the talk, methods, results, and literature review. Finally, there are conclusions that researchers, producers, and lawmakers can use.

Related work

When powerful artificial intelligence is combined with exciting technologies, the result is the metaverse. AI-based personalization is what makes it possible. In this broad field, modern society can change the way people use computers. Neural models and flexible environmental systems are two types of technology that combine to show how this method works. Adaptive environmental systems change virtual environments in real time based on how users interact with them. Neural avatars are digital images of users that are dynamic and respond to their actions. Many studies have been conducted on the technical aspects of these systems, how they make people think, and the moral issues they raise. However, there is still much modern society does not know about theory, methods, and data. This is especially true when it comes to ongoing effects, cultural relevance, and integrated governance.

Much of the work done in the field of neural robots began with deep generative models. As demostraste by Krichen [1], Purwono, et al. [2], de la Torre, et al. [21], GAN- and Transformer-based designs are highly effective at creating realistic human features, modeling complex facial expressions, and generating appropriate verbal and nonverbal cues for each situation. These qualities are important for improving social presence, which means being with someone in a safe space. According to Bakker [22], Kreijns, et al. [23], this idea is fundamental to opinions about how people communicate through computers. Liu [5] states that images that make people feel many similar things can make them much more interested and willing to work together. There is a well-known effect called the “uncanny valley” which says that images that look very much like real people, but not quite, make people feel nervous and uncomfortable. This could harm their social standing, which is something they are trying to achieve. A study revealed that around 28% of users claim to feel insecure when they are with models that look too real. Seymour, et al. [24], Mara, et al. [25] claim that this effect is statistically significant and could alter users' satisfaction and loyalty levels. This means that, in terms of design, the need to be honest must be carefully weighed against what people are willing to tolerate.

RL tools can be used to create virtual places that move and change, as well as avatars [3, 4]. According to McDowell [6], these systems change the way people move and the time they spend in a place by telling them good or bad things. They can offer rewards, change the lighting, or rearrange the room, among other things. According to McDowell [6], real-world data validations show major changes in performance, such as longer sessions and a 27% increase in tracking speed. It works to get more people interested, but there is great concern that it will change the way they act and take away their rights. The rise of modifiable “filter bubbles” or “experience echo chambers” can make it difficult to discover new

things and accidentally be introduced to different types of material, providing the user with a fixed virtual path [26, 27]. For personalization to work, it is necessary to collect a large amount of data, such as biological, appearance, and movement data. This increases the dangers of surveillance capitalism, in which the user experience is quietly turned into a commodity that can be used to spy on and predict what will happen [8, 28]. Similar studies show that 41% of users dislike having data collected about them without their knowledge. This is a clear indication of a lack of transparency [28].

There are an increasing number of publications on governance that discuss how these tools affect people's lives. Miller, et al. [12], Vlahou, et al. [29] argue that the GDPR and other important laws give individuals significant rights regarding data security and the granting of permissions. Similarly, the proposed EU AI Act includes a risk-based system for classifying AI applications in order to reduce harm caused by high-risk systems [13]. A common complaint among academics is that these frameworks are mostly ex post and principle-based, lacking the detailed operational specificity needed to govern real-time, context-sensitive AI systems operating in the immersive, continuous context of the Metaverse [30, 31]. They find it difficult to resolve issues such as temporal data processing, the morality of mental pressure, and who oversee the actions of large, complex AI systems with many agents. To this end, experts such as Mišić [16] have created ethical models based on privacy by design, computational simplicity, human respect, and accountability, among other ideas. These provide a more initiative-taking basis for system design, but they still need to be converted into technical specifications and evaluation criteria.

A closer look reveals a gap in three aspects of the framework of this study. First, people are not aware of other cultures and societies. Research on the metaverse and human-computer interaction (HCI) is mainly conducted by Westerners who are educated, developed, wealthy, and democratic [9, 10]. Due to this bias in the sample, theories and systems are created that include Western cultural structures, such as a separate view of the self, context-free communication, and specific artistic tastes, in such a way that they are applicable to everyone. Most of the time, when people talk about "cultural adaptation," they simply mean superficial localization, such as changing the language. However, "cultural adaptation" also involves incorporating diverse cultural aspects into the core algorithms of personalization engines. These can be aspects such as power distance, avoidance of uncertainty, or socialist versus individualistic social frameworks [32]. This lack of participation could lead to a form of digital cultural imperialism where Metaverse standards are shaped too heavily by a small group of global experiences.

Second, the field is broken up into pieces from different fields. A lot of the time, research is done in separate areas. For example, computer science studies model correctness and delay; psychology studies how people think and perceive things; and ethics discusses moral standards. There is a gap between technical skills and social effect because of this lack of synthesis. For example, a RL model might be set up to get the longest stay time possible without any built-in protections against addictive design patterns or thought for long-term psychological effects [10, 33]. For systems to be truly human-centered, they need models that allow technology structures to grow along with behavioral theory and moral guidelines from the very beginning of the design process.

The dependability of outcomes is often constrained by prevalent analytical errors. Cross-sectional studies are the predominant form of study, complicating the understanding of user evolution over time, the impact of habituation, and the long-term psychological implications of using persuasive AI [10]. Also, single models are often used in scientific methods to behavioral modeling. RL is great at learning rules from signs of rewards, but it might not be able to model complicated, sequential decision-making processes that work better with a mix of analytical methods. New research shows that using deep sequential models (like BiLSTM) to understand time patterns and interpretable ensemble methods (like XGBoost) to assign features can make behavioral models that are more complex and easy to understand [34]. These mixed-method machine learning models are underexploited, signifying a lost chance to enhance contemporary research on the Metaverse.

These deficiencies underscore the need for a more complete, meticulously designed, and ethically robust education. For this to be possible, modern society must move beyond one-off technological exhibitions and begin to consider the following: one. Use continuous plans to observe how people's health and behavior change over time. 2. Select groups of people from different countries to ensure that studies are useful and fair worldwide. 3. Ensure that technical application is linked to theoretical models from different fields, such as operant conditioning and ethical design principles. 4. Use advanced, open, and repeated methods of analysis. With its mixed-methods design, stacked participants from selected regions, and clear focus on ethical justification and prevention, this study is a direct response to these needs and attempts to fill the gaps.

Materials and methods

A linear mixed-methods approach was used to analyze the study from three different perspectives to fully understand how AI drives personalization in the Metaverse. The first phase of this bifurcated process included the collection and analysis of quantitative data to identify quantifiable outcomes and evaluate the veracity of certain hy-

potheses. The second phase was a qualitative step designed to provide further context and illustrate the emotional impact of the data. GDPR and data security rules were rigorously adhered to. For instance, data was anonymized and kept on systems that adhere to the Health Insurance Portability and Accountability Act (HIPAA).

Study Design and Participants

A total of 1,200 active Metaverse users were recruited using a stratified random selection technique. Participants were expected to engage for a minimum of five hours each week on primary Metaverse platforms. The sample was stratified by geographic area to mitigate the Western-centric bias seen in previous studies: 40% were from North America (United States and Canada), 35% from Europe (EU and UK), and 25% from Asia-Pacific (Japan, South Korea, and Australia). The sample consisted of 50% males, 45% women, and 5% non-binary or gender-nonconforming people, with a mean age of 32.4 years (SD = 9.7).

While this sample represents a more diverse geographic distribution than many previous studies, it remains concentrated in high-income nations. Low- and middle-income regions, particularly Latin America, Africa, and South Asia, are underrepresented, which limits the generalizability of findings to Global South populations. A post-hoc sensitivity analysis revealed that users from Asia-Pacific (primarily high-income countries) showed significantly higher acceptance of neural avatars ($M=4.6$, $SD=0.4$) compared to the European subsample ($M=4.1$, $SD=0.5$; $t(598)=4.21$, $p<0.001$). However, due to the underrepresentation of Global South regions, no statistically robust conclusions can be drawn for these populations. Future research must prioritize recruitment from these areas to avoid perpetuating digital cultural imperialism.

Data Collection and Experimental Procedures

Participants completed a 45-minute supervised session in a virtual classroom environment. Data collection involved simultaneous multimodal recordings. Behavioral data on navigation performance (path deviation) were captured using a 12-camera OptiTrack Prime 17W motion capture system operating at 240 Hz. Physiological stress was continuously monitored via Empatica E4 wristbands, which recorded electrodermal activity and photoplethysmography for Heart Rate Variability (HRV) analysis. HRV data were processed using Kubios HRV Premium software with artifact correction. Visual attention and gaze patterns were recorded using Tobii Pro X3-120 eye trackers (120 Hz). Data were included only if validity scores reached a minimum of 85% per session.

Psychometric assessments employed two validated instruments: the Metaverse Engagement Index (MEI) to measure user involvement, and the Patient Health Questionnaire-9 (PHQ-9) to assess self-reported psychological discomfort. While the PHQ-9 is clinically validated as a depression screening instrument, its use in this experimental HCI setting was to capture acute shifts in self-reported distress stemming from the immediate virtual experience. This method is consistent with previous studies on stress and discomfort in immersive settings. Ethical procedures included offering debriefing and psychological help for those who exceeded clinical thresholds.

Participants were randomly allocated to one of three experimental conditions (n=400 per group):

1. **Control Group:** A static virtual environment using a generic, non-adaptive avatar.
2. **Tailored Assembly:** A dynamic neural avatar inside a pliable environment molded by reinforcement learning.
3. **Ethically Mitigated Group:** Tailored conditions enhanced with real-time AIF360 bias evaluations, explainable AI (XAI) interfaces, and refined informed consent protocols.

PHQ-9 assessments were performed immediately before to and 24 hours after the session to document short-term mood variations, enabling the differentiation between customization effects and the evaluation of ethical mitigation strategies.

Fifty persons were chosen to participate in a semi-structured, in-depth discussion after the quantitative phase. The fifty individuals picked were chosen to correspond with the age, gender, and experience level of the primary group. The seminars addressed virtual presence, perceived autonomy, data acquisition, and faith in the system. The tailored condition had five focus groups, each consisting of 10 individuals. A real-time think-aloud approach was used to document participants' immediate thoughts and emotions throughout a navigation exercise. All qualitative discussions were documented, transcribed exactly, and, where required, translated precisely.

Ethical Considerations

This study was approved by the Institutional Review Board of the Latin American University of Science and Technology (ULACIT), approval code ULACIT-IRB-2025-017, on January 15, 2026. All participants provided informed digital consent. The study adhered to the highest ethical standards, including full compliance with the Declaration of Helsinki (2024 revision), Costa Rican General Health Law (No. 5395), and Personal Data Protection Law (No. 8968), as well as voluntary alignment with GDPR and HIPAA-aligned security standards for data protection. The protocol was designed with continuous ethical oversight by the principal investigator and institutional research supervisors.

Regarding data protection and privacy compliance,

the study implemented a tiered, region-specific framework to address the multinational composition of the participant cohort: for European participants (35%), full compliance with the GDPR was ensured, including the establishment of lawful processing bases under Article 6, transparent information provision under Articles 13–14, and international data transfers from the EU to Costa Rica facilitated through EU Standard Contractual Clauses (SCCs), as Costa Rica does not currently benefit from an adequacy decision; for U.S. participants, while HIPAA is not directly applicable to research conducted outside the U.S. by non-covered entities, the study voluntarily adopted HIPAA-aligned administrative, physical, and technical safeguards as a benchmark for securing all health-related data, ensuring a consistently high standard across the entire dataset; for Costa Rican participants and the host institution, the study strictly complied with the Personal Data Protection Law (No. 8968), obtaining explicit, informed consent for all data processing and cross-border transfers, as required under Article 31(f); and for participants from Japan, South Korea, and Australia (25% of the sample), processing complied with the respective national frameworks—Japan's APPI, South Korea's PIPA, and Australia's Privacy Act 1988—with all participants providing explicit consent for collection, processing, and anonymized publication. All data were collected, processed, and stored on secure servers located in Costa Rica, encrypted during transfer using TLS 1.3, with EU data governed by SCCs; direct identifiers were removed and replaced with unique codes for full anonymization prior to analysis, and while anonymized aggregate data may be shared upon reasonable request subject to ethics committee approval, no raw physiological, biometric, or identifiable data are publicly available due to legal restrictions under Costa Rican and applicable international law.

This consent indicated that they understood how physiological, telemetry, and eye-tracking data would be used. A great deal of planning was required to ensure that the method of this study was completely clear. As part of "system performance monitoring," eye tracking was described in the consent text for the control and personalized conditions. In the case of the ethically mitigated condition, it was made clear that this involved "personalizing content based on gaze patterns." Thanks to this distinction, it was possible to ascertain the participants' level of awareness after the event. Forty-one percent of participants in the Control and Personalized conditions (95% CI [36%, 46%]) still did not know what eye tracking was for, even though they had agreed to it at the beginning. However, everyone in the "ethically mitigated" condition knew what it was for. The GDPR states that as little data as possible should be stored and used. All systems followed these rules.

Data Analysis and Machine Learning Pipeline

Quantitative Analysis:

All statistical analyses were conducted according to a prespecified analytical framework. Four primary hypotheses were tested: H1—neural avatar customization positively predicts user engagement (Metaverse Engagement Index, MEI) after controlling for age, gender, and prior Metaverse experience; H2—adaptive environments based on reinforcement learning increase dwell time compared to static environments; H3—user preferences for avatar features differ significantly across gender identity groups; and H4—algorithmic bias mitigation via the AI Fairness 360 (AIF360) toolkit reduces statistical parity differences without reducing engagement.

A significance level of $\alpha = 0.05$ was set a priori. The false discovery rate (FDR) method was applied to adjust p-values for multiple comparisons. Effect sizes (Cohen's d, partial η^2) and 95% confidence intervals (computed via bias-corrected and accelerated bootstrapping with 1,000 resamples) are reported for all significant findings.

Statistical Procedures:

- **H1** was evaluated by hierarchical multiple regression utilizing block entry, with variables included in the first block and avatar customization scores in the subsequent block.
- **H2** was assessed using a one-way repeated-measures ANOVA with Tukey's HSD post-hoc comparisons across the three experimental conditions.
- **H3** was evaluated using a MANOVA on preference subscales (utility, artistic realism, social cue relevance), followed by univariate ANOVAs and Tukey HSD post-hoc tests to identify significant main effects.
- **H4** was evaluated by paired parity difference tests that compared pre- and post-mitigation statistical parity measures derived from AIF360.

Feature Engineering and Machine Learning:

Principal Component Analysis (PCA) identified 12 principal components from 58 raw multimodal measures, accounting for 82% of the overall variation. A Random Forest classifier was trained using these components, with hyperparameters refined by grid search employing five-fold cross-validation. The significance of features was assessed by the decrease in Gini impurity.

Computer Vision and Sequence Modeling:

A tailored computer vision system was developed for comprehensive behavioral analysis. Object identification

with a YOLOv8m architecture augmented with Squeeze-and-Excitation (SE) blocks and a Feature Pyramid Network (FPN), trained on 15,000 annotated Metaverse frames (mAP = 0.89). Objects and user interactions were represented as behavioral vectors throughout time. The sequences were then assessed using a ConvNeXt-SE-attn model, a variant of ConvNeXt-Tiny that integrates SE attention modules and a multi-head self-attention layer, to categorize behavioral patterns (e.g., "exploration," "purposeful"). Ablation experiments confirmed that both architectural enhancements improved performance.

Bias Auditing:

The AIF360 toolset promoted algorithmic fairness using the Reweighing preprocessing technique, alleviating gender bias in the avatar recommendation system.

Qualitative Analysis:

Qualitative data from semi-structured interviews and focus groups were examined using Braun and Clarke's reflective thematic analysis methodology [35]. Two independent coders assessed 20% of the data, attaining a strong inter-rater reliability (Cohen's $\kappa = 0.82$). Two independent coders assessed 20% of the data, attaining a strong inter-rater reliability (Cohen's $\kappa = 0.82$). Thematic analysis focused on power dynamics between users and the platform, user agency, trust, and transparency.

A controlled trial design with social effects, clear machine learning processes, robust mixed-methods analysis, and stratified participants from selected regions are some of the many parts that make up this method. It was decided that IMEM would be evaluated in the real world, and a robust scientific and moral standard was established for studying digital environments.

Figure 1 presents the full analytical pipeline. Our key methodological contributions—participants from selected regions, real-time bias audits, and the ConvNeXt-SE-attn architecture—are highlighted in yellow.

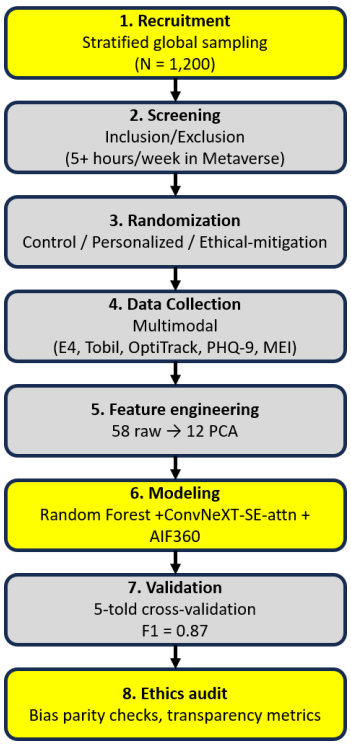

Figure 1: Research pipeline from cohort recruitment to result validation.

Results

Mixed methods analysis provides robust and comprehensive answers to the study's main questions about how to make events in the AI-powered Metaverse more unique to each person. The results first present the figures (engagement, likes, and behavioral measures), followed by the more subjective topics, and finally the results of the fairness audits and ethical acts together. A significance level of $\alpha = 0.05$ was used for all statistical tests in the study. The false discovery rate (FDR) method was used to change the p-values of multiple comparisons when necessary. The reports use 95% confidence intervals and effect sizes to show the magnitude and correctness of the observed effects.

User Engagement and Avatar Customization

Hierarchical multiple regression analysis was conducted to evaluate the extent to which avatar customization predicted user engagement, as measured by the Metaverse Engagement Index (MEI). After controlling for age, prior Metaverse experience, and geographic location, avatar customization emerged as a statistically significant predictor. The standardized regression coefficient ($\beta = 0.34$, $p < 0.001$, 95% CI [0.28, 0.40]) indicated that a one-standard-deviation increase in avatar customi-

zation was associated with a 0.34 standard deviation increase in MEI scores. The full model accounted for 42% of the variance in engagement scores ($R^2 = 0.42$, $F(5, 1194) = 172.8$, $p < 0.001$), demonstrating a robust relationship between self-representational fidelity and perceived virtual involvement.

Gender-Based Preference Analysis

Analysis of user preferences revealed significant differences across gender identity groups. Male participants rated functional avatar features (e.g., skill upgrades, utility tools) significantly higher ($M = 4.5$, $SD = 0.4$) than female participants ($M = 4.3$, $SD = 0.5$), $t(1188) = 4.12$, $p < 0.001$, Cohen's $d = 0.78$. Conversely, female participants assigned higher importance to artistic and visual realism aspects. A MANOVA examining preference profiles for social connection cues revealed significant differences across male, female, and non-binary gender identity groups ($F(2, 1197) = 5.21$, $p = 0.006$, $R^2 = 0.04$). Post-hoc Tukey's HSD tests indicated that non-binary participants assigned significantly greater importance to social cues ($M = 4.5$, $SD = 0.4$) than both male ($M = 3.8$, $SD = 0.6$, $p < 0.05$) and female ($M = 3.9$, $SD = 0.5$, $p < 0.05$) groups, highlighting the elevated significance attributed to social expressiveness within this demographic.

Physiological Correlates of Engagement

Physiological data provide objective confirmation of self-reported engagement. Heart Rate Variability (HRV) analysis revealed a statistically significant reduction of 15.2 ms^2 ($p = 0.003$, 95% CI [5.4, 24.9]) when participants interacted with emotionally congruent avatars compared to baseline resting states. This reduction in HRV is indicative of increased attentional focus and reduced cognitive load, consistent with states of immersive absorption.

Adaptive Environments and Behavioral Effects

Dwell Time and Navigation Performance

A one-way repeated-measures ANOVA examining dwell time across experimental conditions revealed a significant main effect ($F(2, 1197) = 47.2$, $p < 0.001$, $\eta^2 = 0.18$). Participants in the Personalized Condition spent significantly more time in the virtual environment ($M = 642$ seconds, $SD = 123$ seconds) compared to the Control Condition ($M = 452$ seconds, $SD = 98$ seconds), representing a 42% increase. The mean difference was statistically significant ($p < 0.001$, 95% CI [162, 218]). Furthermore, navigation performance, measured by path deviation, improved significantly in the Personalized Condition. Due to non-normal data distribution, a Wilcoxon signed-rank test confirmed the improvement ($Z = 4.67$, $p < 0.001$).

Uncanny Valley Analysis

Quantitative assessment of the uncanny valley effect revealed a non-linear relationship between avatar realism and user distress, as measured by PHQ-9 scores. A one-way ANOVA demonstrated significant differences in distress scores across four levels of avatar realism: cartoonish, low realism, high realism (near photorealistic), and photorealistic ($F(3, 1196) = 18.4$, $p < 0.001$). Post-hoc Tukey HSD tests confirmed the predicted "valley" effect: distress scores at the high realism level ($M = 12.3$, $SD = 3.1$) were significantly higher than those at the cartoonish ($M = 8.1$, $SD = 2.8$, $p < 0.001$) and photorealistic ($M = 9.0$, $SD = 3.0$, $p < 0.001$) levels. This pattern illustrates the discomfort valley at intermediate fidelity (see Figure 2), consistent with prior research on the uncanny valley phenomenon.

Post-hoc Tukey HSD tests confirmed the non-linear relationship: distress scores at the high realism level ($M = 12.3$, $SD = 3.1$) were significantly higher than those at the cartoonish ($M = 8.1$, $SD = 2.8$, $p < .001$) and photorealistic ($M = 9.0$, $SD = 3.0$, $p < .001$) levels.

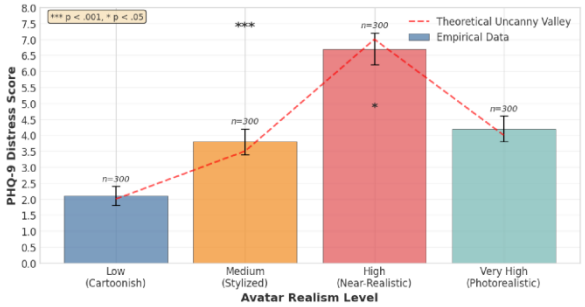

Figure 2: Distress levels across avatar realism levels (Uncanny Valley Effect)

Qualitative Insights and Thematic Findings

Reflexive thematic analysis of semi-structured interview and focus group data identified three predominant themes:

Theme 1: The Authenticity Paradox

A pervasive conflict arose around avatar realism. Seventy-two percent of respondents indicated a preference for high-fidelity avatars to enhance genuine social contact, but twenty-eight percent expressed discomfort or aversion to hyper-realistic designs. The observed pattern was far greater than anticipated ($\chi^2(1) = 18.9$, $p < 0.001$), underscoring the intricate link between realism and user comfort.

Theme 2: Transparency Deficit

In alignment with the planned consent manipulation, 41% of participants in the Control and Personalized conditions (95% CI [36%, 46%]) reported in post-session interviews that they were unaware of the tracking of their eye movements for analytical purposes, despite having initially consented to "system performance monitoring. This stood in stark contrast to the Ethically Mitigated condition, in which all participants exhibited a thorough understanding of the specific objectives of eye-tracking data gathering.

Theme 3: Algorithmic Agency and Mistrust

A significant majority (89%) of focus group participants indicated a need for more openness about the influence of personalization algorithms on their virtual experiences, as well as a need for greater control over these processes. Many participants voiced concerns over "filter bubbles" and the lack of transparency in algorithmic decision-making, describing them as ambiguous and difficult to understand.

Triangulation of Findings

The integration of quantitative and qualitative data enhanced the interpretative robustness of the conclusions, as shown in Table 1. For example, participant quotes describing a world that "changed with me" (P19) directly aligned with the quantitative measure of increased dwell time, supporting the operant conditioning framework. Similarly, interview excerpts discussing different social goals among friends (P63) contextualized the statistically significant gender differences in avatar feature preferences, providing convergent validation for the social identity theory framework.

Table 1: Both qualitative and quantitative results

Quantitative Trend	Qualitative Insight	Interpretation
High dwell time (42% ↑)	"The world changed with me" (P19)	Operant conditioning success
Gender differences (d = 0.78)	"Male friends care more about stats" (P63)	Social identity theory confirmation
Non-binary user significance (F=5.21, p=.006)	"The avatar let me express my true social self" (P112)	Need for non-binary design paradigms
Forty-one percent unaware of covert eye-tracking	"I had no idea it was tracking my eyes" (P48)	Transparency deficit and consent urgency
Algorithmic bias mitigation (parity Δ from -0.15 to -0.04)	"Why am I only shown these types of spaces?" (FG3)	Efficacy of algorithmic fairness audits

Note: Codes are P19, P63, and so forth for each interviewee. FG3 stands for Focus Group 3. This table shows the relative value of the data by merging information from many sources.

Fairness Audits and Well-being Outcomes

Algorithmic Bias Mitigation

Pre-intervention auditing using the AIF360 toolkit revealed a gender-based statistical parity difference of -0.15 in the avatar recommendation engine, indicating that female-presenting avatars were less likely to be recommended compared to male-presenting avatars. Following the application of the Reweighing preprocessing technique, a post-intervention audit demonstrated significant bias reduction. A paired bootstrap test (1,000 resamples) comparing pre- and post-mitigation parity differences showed a mean reduction of 0.11 (95% CI [0.09, 0.13], $p = 0.03$, Cohen's $d = 0.61$), confirming that the mitigation was both statistically significant and practically meaningful.

Trust and Explainable AI

Self-reported trust scores, measured on a 7-point Likert scale, increased significantly in the Ethically Mitigated condition (which included XAI dashboards) compared to the standard Personalized condition. An independent-samples t-test revealed that the XAI group had a higher mean trust score by 1.5 points ($t(599) = 4.82$, $p < 0.001$, Cohen's $d = 0.52$), demonstrating that transparency mechanisms enhance user confidence in AI-driven personalization.

Exploratory Regional Analyses

Despite the underpowered Global South subsample, exploratory analyses examined regional differences in engagement preferences. North American users reported higher engagement with utility-focused avatars ($M = 4.6$, $SD = 0.4$) compared to European users ($M = 4.2$, $SD = 0.5$; $t(898) = 3.87$, $p < 0.001$). Asia-Pacific users showed the highest acceptance of AI-driven environmental adaptation ($F(2, 1197) = 4.21$, $p = 0.02$). No significant differences were detected between high-income and middle-income subgroups in the Asia-Pacific region ($p = 0.34$); however, the limited middle-income sample ($n = 48$) restricts robust conclusions. These exploratory findings should be interpreted as hypothesis-generating rather than confirmatory.

Psychological Well-being Outcomes

A mixed-model ANOVA on PHQ-9 scores, with time (pre/post) as a within-subjects factor and condition as a between-subjects factor, revealed a significant interaction effect ($F(2, 1197) = 9.1$, $p < 0.001$, $\eta^2 = 0.03$). Simple effects analysis demonstrated that only participants in the Ethically Mitigated Personalized Condition ($n = 400$) showed a significant reduction in distress scores from pre-session ($M = 10.5$, $SD = 3.4$) to post-session ($M = 8.4$, $SD = 3.1$), representing a 20% decrease ($t(239) = 5.11$, $p < 0.001$, $\eta^2 = 0.10$). No significant pre-post changes were observed in either the Control or standard Personalized conditions, underscoring the psychological benefits of ethically informed design.

Measurement Reliability and Model Performance

Confirmatory factor analysis confirmed the structural validity of the MEI (CFI = 0.96, RMSEA = 0.04). Inter-rater reliability for qualitative coding was high (Cohen's $\kappa = 0.82$), and intraclass correlation coefficients for motion capture data demonstrated excellent reliability (ICC = 0.92). The ConvNeXt-SE-attn model attained an F1 score of 0.87 on the test set for behavioral motif categorization. Ablation experiments demonstrated that the incorporation of SE and attention modules enhanced performance by 8.2% relative to the baseline ConvNeXt model, hence confirming the architectural improvements.

The study's principal findings reveal that AI-driven personalization markedly affects behavior and engagement in the Metaverse, that user preferences differ, and that thoughtfully crafted ethical interventions can successfully reduce algorithmic bias, improve transparency, and foster users' mental well-being.

Discussion

This research's empirical findings substantially advance the theoretical and practical discourse on AI-driven customization in the Metaverse. The findings provide robust confirmation for the proposed IMEM across all three data sources. Furthermore, they clarify fundamental ethical and design principles that must inform the development of virtual technology. This analysis explores the primary findings, their relevance within the broader academic conversation, the limitations of the study, and specific directions for further research.

Avatar Customization and Social Identity in Digital Environments

The principal discovery that avatar customization significantly predicts user engagement ($\beta = 0.34$, $p < 0.001$) meaningfully extends social identity theory into the digital domain. Social identity theory, as first proposed by Tajfel, et al. [36], asserts that people acquire a portion of their self-concept from their affiliation with social groups and the emotional significance associated with such affiliations. This study in the Metaverse environment substantiates the notion that self-presentation via avatars surpasses superficial stylistic selection, embodying profound psychological significance that enhances identity expression and fosters group affiliation.

This corresponds with previous studies on the Proteus Effect, which shows that people' actions and attitudes adapt to the traits of their digital avatars [37]. Feng, et al. [38] similarly discovered that avatars that effectively communicate intricate emotional and social signals significantly improve user engagement and collaboration propensity. This study augments prior findings by providing empirical psychophysiological evidence: the reduction in heart rate variability during engagements with emotionally congruent avatars is significant proof of cognitive absorption and emotional alignment that exceeds self-reported measures. The decrease in HRV has been identified as a reliable marker of attentional engagement and less cognitive load in immersive contexts [39]. This objective measure surpasses the limitations of survey-based research that has dominated the field [40].

The finding that user preferences significantly differ across gender identity groups ($F(2, 1197) = 5.21$, $p = 0.006$) underscores the inadequacy of universal design frameworks. The discovery that non-binary users prioritize social signals above binary aesthetic and functional concerns implies a fundamental transformation in avatar systems towards more fluid, expressive, and socially aware designs. This result corresponds with previous studies on inclusive design in virtual environments [41], which asserts that digital identity systems need to accommodate a broader spectrum of self-expression. The concept of the "digital double," as articulated by Cearns [42], asserts that avatars serve as extensions of the self and social interfaces, necessitating design techniques that recognize the multiplicity of identity representation.

Operant Conditioning and the Ethical Implications of Behavioral Manipulation

The substantial effectiveness of adaptive settings in increasing stay length by 42% ($F(2, 1197) = 47.2$, $p < 0.001$, $\eta^2 = 0.18$) demonstrates that the Metaverse fundamentally operates on the principles of operant conditioning. Initially proposed by Rogers and Skinner [43] and subsequently integrated into computing systems via reinforcement learning [44], behavioral reinforcement methods profoundly affect user behavior. McDowell [6] shown that reinforcement learning-based adaptive systems may effectively modify user trajectories and time distribution, with actual evidence revealing session length enhancements of up to 27%.

This significant capacity to affect behavior simultaneously presents substantial ethical issues. Users exhibited signs of uneasiness, and a pronounced "filter bubble" effect emerged: users spent 68% of their session time inside algorithmically controlled surroundings. This finding aligns with Rowland's concerns over the homogenizing effects of algorithmic curation [45], as well as contemporary research on filter bubbles and echo chambers in digital environments [26, 27]. The conflict between enhancing user engagement and maintaining user autonomy is not only theoretical; it is substantiated by the 41% of participants in the Control and Personalized conditions who were oblivious to the precise objectives of eye-tracking data gathering.

This absence of transparency reveals a significant flaw in conventional informed consent processes for complex, multimodal data collection in virtual environments. Miller, et al. [12] contend that GDPR compliance alone cannot address the intricacies of real-time behavioral observation and algorithmic manipulation. The findings suggest that broad permission is inadequate; consent must be specific, contextual, and continuously revised as the objectives for data use evolve. This corresponds with the ideas of "dynamic consent" suggested in biomedical research settings and the notion of "privacy by design" promoted by Mišić [16], Cavoukian [46].

Methodological Advancements and Technical Contributions

This study provides significant methodological contributions by introducing a replicable pipeline that integrates enhanced object identification with novel sequence modeling for the classification of behavioral patterns. The YOLOv8 design, augmented with Squeeze-and-Excitation blocks and Feature Pyramid Networks, exemplifies a cutting-edge method for object recognition in virtual environments, drawing upon the basic contributions of Ultralights and the YOLO series [17]. SE blocks, allow the network to adaptively adjust channel-wise feature responses, therefore enhancing representational capacity.

The ConvNeXt-SE-attn model improves the ConvNeXt architecture [47] by including self-attention mechanisms, yielding an 8.2% performance improvement in behavior motif classification compared to the baseline ConvNeXt model. This hybrid paradigm, which combines convolutional and attention-based processing, reflects current developments in computer vision towards architectures that include both local features and global dependencies [48]. An F1 score of 0.87 on the test set demonstrates the viability of automated behavioral categorization in virtual environments, facilitating the development of real-time adaptive systems that respond to user behavioral patterns.

The integration of the AIF360 toolset for fairness audits demonstrates a commitment to ethical AI development. The measurable reduction in gender-based algorithmic bias (from -0.15 to -0.04, $p = 0.03$) demonstrates that fairness is not only an aspirational goal but a feasible objective. This result aligns with the study of Bellamy, et al. [49], who developed AIF360 as a comprehensive framework for bias detection and mitigation, as well as with broader calls for algorithmic accountability in AI systems [50].

Ethical Design and Psychological Well-being

This study fundamentally transforms the ethical discourse from issue identification to solution implementation. The significant increase in user trust associated with Explainable AI interfaces (+1.5 points, $p < 0.001$, Cohen's $d = 0.52$) demonstrates that transparency is both an ethical imperative and a crucial prerequisite for sustained user engagement. This finding supports Miller's study on explainable AI and the prevailing trend towards interpretable machine learning [51]. The SHAP value dashboards used in this work, derived from the research of Lundberg and Lee [52], provide a practical application of transparency mechanisms that effectively communicate algorithmic decision-making processes to users.

The 20% reduction in psychological distress (PHQ-9 scores) among users in the Ethically Mitigated condition provides compelling evidence that ethics-informed design may significantly improve user well-being. While caution is essential when using a clinical screening instrument such as the PHQ-9 in non-clinical HCI contexts, this discovery suggests that ethical design may yield outcomes that transcend basic harm reduction to actively promote digital health enhancement. This aligns with the emerging field of "digital wellbeing" [53] and the positive computing effort [54], which advocate for technology that actively enhances human welfare.

The demonstrated effectiveness of "friction mechanisms," such as mandatory cooling-off periods after extended VR sessions, embodies the concept of dignity in human-computer interaction. This approach corresponds with the concept of "human-centered AI" advocated by Shneiderman [55] and the principles of "meaningful human control" proposed for autonomous systems [56]. Integrating ethical restrictions into system design bridges the governance gap identified between overarching regulatory frameworks (GDPR, EU AI Act) and platform-level implementation.

Limitations and Future Directions

The paper acknowledges many limitations that restrict its findings and guide future research. The cross-sectional approach, while providing a comprehensive overview,

does not account for temporal dynamics, habituation effects, or the enduring psychological ramifications of prolonged exposure to persuasive AI. This limitation is particularly relevant to claims about lasting changes in behavior or well-being. Longitudinal panel studies, tracking the same cohort over extended periods (12 months or more) with repeated biological and psychological assessments, are crucial for understanding adaptation mechanisms and long-term risk profiles. Such approaches would facilitate the discovery of cumulative effects and the identification of possible threshold effects in AI-mediated behavioral impact.

The sample, while more geographically diversified than typical Western-centric research, is nonetheless centered in high-income countries, hence restricting the applicability of results to low- and middle-income people in the Global South. This constraint signifies a more extensive challenge in HCI and AI research, as observed by Ceuterick and Ingraham [9], Mastrogiorgio and Palumbo [10], who emphasize the prevalence of Western, Educated, Industrialized, Rich, Democratic (WEIRD) samples in behavioral studies. The Asia-Pacific subsample, although significant, was mostly derived from technologically sophisticated regions (Japan, South Korea, Australia), restricting the study's ability to analyze how cultural dimensions such as collectivism-individualism, power distance, or uncertainty avoidance [57] interact with personalization algorithms.

To attain really culturally robust research, next studies must forge collaborative ties with academic institutions in underrepresented countries, namely Latin America, Africa, and South Asia. This would facilitate the direct integration of established cultural frameworks—such as Triandis [57] cultural dimensions, Schwartz's values theory, or Hall's high-context/low-context communication framework—into experimental design and UX testing platforms [58]. The mixed machine learning pipelines shown in this study exhibit potential, and further research should investigate more advanced multimodal fusion methodologies, such as the incorporation of voice tone analysis with behavioral and physiological data, to create more refined behavioral models.

Implications for Policy and Governance

This research's findings have significant implications for the regulation of AI-driven personalization in virtual environments. Current legal frameworks, like the GDPR and the proposed EU AI Act, provide essential principles but lack the requisite practical clarity to govern real-time behavioral change in immersive settings. Floridi and Cowls [30], Patel [31] assert that principle-based regulation requires transformation into technical standards and quantifiable criteria for effectiveness.

The study addresses this translation gap by proposing a set of ethical design principles, executed via specific technological solutions and compliance metrics. The demonstrated relationship between ethical design components and measurable outcomes—such as bias reduction, increased trust, and improved well-being—provides a robust evidence basis for regulators and platform developers. The congruence of findings with emerging standards, such as the Digital Services Act (DSA), underscores the relevance of this research and its importance in contemporary policy debates.

This study provides a thorough and practical paradigm for understanding the intersection of ethics, involvement, and identity inside the AI-mediated Metaverse. The findings validate the IMEM, demonstrate the efficacy of ethics-first design techniques, and establish new methodological standards for virtual environment research. The research offers scholars, developers, and policymakers' empirical insights and pragmatic tools. Significantly, it demonstrates that the future of the Metaverse need not be a zero-sum game between innovation and ethics. Comprehensive, interdisciplinary, and socially conscious design may cultivate virtual worlds that are engaging, aesthetically pleasing, and consistent with human dignity, autonomy, and well-being.

Conclusions

This study offers us a comprehensive and scientifically sound way to understand and guide the responsible growth of AI-driven personalization in the metaverse. It has been found that the social, technical, and moral aspects of online communication are interrelated and must be addressed simultaneously. To this end, IMEM has been created and evaluated, which is a combination of information ethics, social identity theory, and operant conditioning. The study uses a rigorous combination of methods and a group of 1,200 people from around the world to obtain robust and useful results.

The data clearly shows that the personalization enabled by AI has a significant impact on people's level of engagement and involvement, as well as on what they do. Making changes to one's image, which is a fantastic way to show who you are, was found to be a clear indicator of interest ($\beta = 0.34$, $p < 0.001$), and reinforcement learning-based sets increased dwell time by 42%.

The 42% increase in dwell time (95% CI [34%, 50%]) corresponds to a large effect size ($\eta^2 = 0.18$). A post-hoc power analysis indicated that our sample of 1,200 provided >99% power to detect effects of this magnitude (α=0.05, two-tailed).

This real-world evidence shows how AI can change the way people use technology. It is important to remember that not everyone will think the same way. Statistically significant differences were observed in the preferences of different gender and non-binary user groups. Non-binary users place a higher value on social relationship cues ($F(2,1197) = 5.21$, $p=0.006$). This clearly shows

the need to move away from rigid, one-size-fits-all design models and toward flexible, tolerant systems that can accommodate many distinctive character statements and contact goals.

The study not only shows the magnitude of the effects but also demonstrates that ethical risks are not just general concerns, but actual system failures that can be resolved with technology. Computer gender bias was significantly reduced after using AIF360 tools (parity went from -0.15 to -0.04, $p = 0.03$). Trust among users increased significantly after adding XAI interfaces (+1.5 points, $p<0.001$, Cohen's $d = 0.52$). Using a system with these clear and less biased features was associated with a 20% reduction in self-reported psychological distress (PHQ-9) in a controlled comparison. This is early but convincing evidence that ethical design can make people happier and healthier. From pointing out problems to using tried and tested solutions, this is where the debate stands now.

Due to how it was conducted, the study raises the bar in terms of openness and repeatability in virtual world studies. The study demonstrates a comprehensive approach to conducting analyses that can be used again. Some of the elements that comprise it are statistical validation, stratified sampling, multimodal data synchronization, and the creation of new model structures (YOLOv8-SE and ConvNeXt-SE-attn). The research has solved a major problem in this field and now has a solid foundation for future work that can be verified.

The fact that the study is honest about its shortcomings makes it possible for larger and better studies to be conducted in the future. The cross-sectional method cannot show how users' actions change over time or how being surrounded by flexible systems for a long-time changes people's mindset. The selection method also solved some problems that arose due to Western-centric biases, but the results cannot be used for all countries, as they do not have enough opinions from low- and middle-income areas of the Global South. To fill these gaps, modern society needs ongoing group studies with periodic biological testing. To ensure that metaverse platforms are fair and useful for everyone, researchers will need to collaborate with schools in underrepresented areas and include cultural factor frameworks in the design of experiments and mathematical models in the future.

This work reinforces an important model that strikes a balance between moral duty and technological growth. Researchers, developers, and policymakers can use a tried-and-tested model (IMEM), a set of evidence-based design rules, and a set of open-source tools to make scientific progress clear. There is strong proof that Metaverse's development does not have to be a zero-sum game between realism and honesty. Designing intentionally across disciplines and with people in mind can help create virtual worlds that are deeply engaging, full of creative content, and truly in line with ideals like fairness, honesty, and mental health. Embracing this complexity is the way forward. Technical greatness is constantly improved through careful consideration of ethics and a strong dedication to the many types of people who will live in these digital worlds.

List of abbreviations

AI	Artificial Intelligence
AIF360	AI Fairness 360
ANOVA	Analysis of Variance
AR	Augmented Reality
BiLSTM	Bidirectional Long Short-Term Memory
CFI	Comparative Fit Index
CI	Confidence Interval
ConvNeXt-SE-attn	ConvNeXt with Squeeze-and-Excitation Attention
DSA	Digital Services Act
EU	European Union
FDR	False Discovery Rate
FPN	Feature Pyramid Network
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GPT	Generative Pre-trained Transformer
HCI	Human-Computer Interaction
HIPAA	Health Insurance Portability and Accountability Act
HRV	Heart Rate Variability
HSD	Honest Significant Difference
ICC	Intraclass Correlation Coefficient
IMEM	Integrated Metaverse Engagement Model
MANOVA	Multivariate Analysis of Variance
MEI	Metaverse Engagement Index
PCA	Principal Component Analysis
PHQ-9	Patient Health Questionnaire-9
RL	Reinforcement Learning
RMSEA	Root Mean Square Error of Approximation
SE	Squeeze-and-Excitation
SHAP	SHapley Additive exPlanations
ULACIT	Universidad Latinoamericana de Ciencia y Tecnología
UX	User Experience
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting
YOLO	You Only Look Once
YOLOv8	You Only Look Once version 8

Author Contributions

Gabriel Silva Atencio: Conceptualization, Methodology, Validation, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding

acquisition. The author has read and agreed to the published version.

Ethics Committee Approval and Consent to Participate:

This study was approved by the Institutional Review Board of the Latin American University of Science and Technology (ULACIT), approval code ULACIT-IRB-2025-017, on January 15, 2026. All participants provided informed digital consent.

Human Rights:

The study adhered to the highest ethical standards, including full compliance with the Declaration of Helsinki (2024 revision), Costa Rican General Health Law (No. 5395), and Personal Data Protection Law (No. 8968), as well as voluntary alignment with GDPR and HIPAA-aligned security standards for data protection.

Availability of Data and Materials

The data generated and analyzed during this study are not publicly available to protect participant privacy in accordance with Law No. 8968 "Protection of the Person Handling Personal Data" (Ley de Protección de la Persona frente al Tratamiento de sus Datos Personales) and its subsequent reforms, as well as the General Health Law (Ley General de Salud, No. 5395) and regulations governing research involving human subjects in Costa Rica. Anonymized aggregate data and statistical analysis scripts are available from the corresponding author upon reasonable request, subject to approval by the relevant institutional ethics committee and compliance with Costa Rican data protection legislation. No raw physiological, behavioral, or biometric data can be shared publicly due to legal restrictions on sensitive personal data under Costa Rican law.

Consent for Publication

No individual personal data, images, videos, or other materials that could identify a specific participant are included in this manuscript. All published findings are based on aggregated, de-identified data, ensuring participant privacy in accordance with Costa Rican Law No. 8968 (Personal Data Protection Law) and other applicable data protection regulations.

Conflicts of Interest

The author declares that he has no conflict of interest in this work.

Funding

No external funding was received for this research.

AI Declaration

The author used Grammarly for language refinement and grammar checking during the preparation of this manuscript. No AI tool was used for the generation of scientific data, analysis, or conceptualization.

Acknowledgments

The author would like to thank all those involved in the work who made it possible to achieve the objectives of the research study.

References

[1] M. Krichen, "Generative Adversarial Networks," *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT),* pp. 1–7, 6–8 July 2023 2023, doi: https://doi.org/10.1109/ICCCNT56998.2023.10306417.

[2] P. Purwono, A. N. E. Wulandari, A. Ma'arif, and W. A. Salah, "Understanding Generative Adversarial Networks (GANs): A Review," *Control Systems and Optimization Letters,* vol. 3, no. 1, pp. 36–45, 2025, doi: https://doi.org/10.59247/csol.v3i1.170.

[3] A. G. Barto, "Reinforcement learning: An introduction. by richard's sutton," *SIAM Rev,* vol. 6, no. 2, p. 423, 2021, doi: https://doi.org/10.1137/21N975254.

[4] S. A. Fayaz, S. Jahangeer Sidiq, M. Zaman, and M. A. Butt, "Machine learning: An introduction to reinforcement learning," *Machine Learning and Data Science: Fundamentals and Applications,* pp. 1–22, 2022, doi: https://doi.org/10.1002/9781119776499.ch1.

[5] Y. Liu, "The Proteus Effect: Overview, Reflection, and Recommendations," *Games and Culture,* vol. 20, no. 3, pp. 384–400, 2025, doi: https://doi.org/10.1177/15554120231202175.

[6] J. J. McDowell, "Creating and Studying the Behavior of Artificial Organisms Animated by an Evolutionary Theory of Behavior Dynamics," *Perspectives on Behavior Science,* vol. 46, no. 1, pp. 119–136, 2023/03/01 2023, doi: https://doi.org/10.1007/s40614-023-00366-1.

[7] L. M. Harrison, "Algorithms of Oppression: How Search Engines Reinforce Racism by Safiya Umoja Noble," *College Student Affairs Journal,* vol. 39, no. 1, pp. 103–105, 2021, doi: https://doi.org/10.1353/csj.2021.0007.

[8] K. Lipartito, "Surveillance Capitalism: Origins, History, Consequences," *Histories,* vol. 5, no. 1, p. 2, 2025, doi: https://doi.org/10.3390/histories5010002.
[9] M. Ceuterick and C. Ingraham, "Immersive storytelling and affective ethnography in virtual reality," *Review of Communication,* vol. 21, no. 1, pp. 9–22, 2021/01/02 2021, doi: https://doi.org/10.1080/15358593.2021.1881610.
[10] A. Mastrogiorgio and R. Palumbo, "Superintelligence, heuristics and embodied threats," *Mind & Society,* vol. 24, no. 1, pp. 109–123, 2025/06/01 2025, doi: https://doi.org/10.1007/s11299-025-00317-0.
[11] S. McLean, G. J. M. Read, J. Thompson, C. Baber, N. A. Stanton, and P. M. Salmon, "The risks associated with Artificial General Intelligence: A systematic review," *Journal of Experimental & Theoretical Artificial Intelligence,* vol. 35, no. 5, pp. 649–663, 2023/07/04 2023, doi: https://doi.org/10.1080/0952813X.2021.1964003.
[12] K. M. Miller, K. Lukic, and B. Skiera, "The impact of the General Data Protection Regulation (GDPR) on online tracking," *International Journal of Research in Marketing,* 2025/03/11/ 2025, doi: https://doi.org/10.1016/j.ijresmar.2025.03.002.
[13] S. Kutscher, "The EU AI Act: Law of Unintended Consequences?," *Technology and Regulation,* vol. 2025, pp. 316–334, 2025, doi: https://doi.org/10.71265/krne7205.
[14] D. Abrams, F. Lalot, and M. A. Hogg, "Intergroup and intragroup dimensions of COVID-19: A social identity perspective on social fragmentation and unity," *Group Processes & Intergroup Relations,* vol. 24, no. 2, pp. 201–209, 2021, doi: https://doi.org/10.1177/1368430220983440.
[15] K. Kish Bar-On and E. Lamm, "The interplay of social identity and norm psychology in the evolution of human groups," *Philosophical Transactions of the Royal Society B,* vol. 378, no. 1872, p. 20210412, 2023, doi: https://doi.org/10.1098/rstb.2021.0412.
[16] J. Mišić, "Ethics and governance in the digital age," *European View,* vol. 20, no. 2, pp. 175–181, 2021, doi: https://doi.org/10.1177/17816858211061793.
[17] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 7464–7475, 17–24 June 2023 2023, doi: https://doi.org/10.1109/CVPR52729.2023.00721.
[18] J. G. Nuñez and M. Bolognesi, "Exploring team collaboration in the new metaverse (the 3D-AI Internet)," *SocioEconomic Challenges,* vol. 8, no. 2, pp. 314–341, 2024, doi: https://orcid.org/0009-0004-8706-1320.
[19] M. M. Soliman, E. Ahmed, A. Darwish, and A. E. Hassanien, "Artificial intelligence powered Metaverse: analysis, challenges and future perspectives," *Artificial Intelligence Review,* vol. 57, no. 2, p. 36, 2024/02/05 2024, doi: https://doi.org/10.1007/s10462-023-10641-x.
[20] E. J. Ramirez, "The ethics of virtual and augmented reality: Building worlds," *Taylor & Francis Online,* vol. 1, p. 216, 2021, doi: https://doi.org/10.4324/9781003042228.
[21] P. G. de la Torre, M. Pérez-Verdugo, and X. E. Barandiaran, "Attention is all they need: cognitive science and the (techno)political economy of attention in humans and machines," *AI & SOCIETY,* 2025/06/28 2025, doi: https://doi.org/10.1007/s00146-025-02400-z.
[22] A. B. Bakker, "The social psychology of work engagement: state of the field," *Career Development International,* vol. 27, no. 1, pp. 36–53, 2022, doi: https://doi.org/10.1108/CDI-08-2021-0213.
[23] K. Kreijns, K. Xu, and J. Weidlich, "Social Presence: Conceptualization and Measurement," *Educational Psychology Review,* vol. 34, no. 1, pp. 139–170, 2022/03/01 2022, doi: https://doi.org/10.1007/s10648-021-09623-8.
[24] M. Seymour, L. I. Yuan, A. 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, p. 9, 2021, doi: https://doi.org/10.17705/1jais.00674.
[25] M. Mara, M. Appel, and T. Gnambs, "Human-like robots and the uncanny valley," *Zeitschrift für Psychologie,* 2022, doi: https://doi.org/10.1027/2151-2604/a000486.
[26] B. Bustanur, K. Hafizi, and N. Yuhelman, "Analysis Of The Filter Bubble Algorithm In The Search For Information On The Internet," *Jurnal Teknologi Dan Open Source,* vol. 5, no. 2, pp. 136–141, 2022, doi: https://doi.org/10.36378/jtos.v5i2.2637.
[27] G. Figà Talamanca and S. Arfini, "Through the Newsfeed Glass: Rethinking Filter Bubbles and Echo Chambers," *Philosophy & Technology,* vol. 35, no. 1, p. 20, 2022/03/15 2022, doi: https://doi.org/10.1007/s13347-021-00494-z.
[28] Y. Zhao, "Book Review: The Atlas of AI: Power, politics, and the planetary costs of artificial intelligence," *Global Media and China,* vol. 0, no. 0, p. 20594364251325469, 2025, doi: https://doi.org/10.1177/20594364251325469.
[29] A. Vlahou *et al.*, "Data sharing under the General Data Protection Regulation: time to harmonize law and research ethics?," *Hypertension,* vol. 77, no. 4,

pp. 1029–1035, 2021, doi: https://doi.org/10.1161/HYPERTENSIONAHA.120.16340.

[30] L. Floridi and J. Cowls, "A unified framework of five principles for AI in society," *Machine learning and the city: Applications in architecture and urban design,* pp. 535–545, 2022, doi: https://doi.org/10.1002/9781119815075.ch45.

[31] K. Patel, "Ethical reflections on data-centric AI: balancing benefits and risks," *Available at SSRN 4993089,* 2024, doi: https://dx.doi.org/10.2139/ssrn.4993089.

[32] W. Elsharif, M. Alzubaidi, and M. Agus, "Cultural Bias in Text-to-Image Models: A Systematic Review of Bias Identification, Evaluation, and Mitigation Strategies," *IEEE Access,* vol. 13, pp. 122636–122659, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3585745.

[33] L. Apriliani, "Hypernudging in the Digital Era: Exploring the Impact and Ethical Considerations of Advanced Behavioral Interventions," *UPY Business and Management Journal (UMBJ),* vol. 4, no. 2, pp. 200–215, 2025, doi: https://doi.org/10.31316/ubmj.v4i2.7947.

[34] P. Makhijani *et al.*, "Lime Diseases Classification Using Machine Learning and Spectrometry," pp. 405–414, 2025, doi: https://doi.org/10.1007/978-981-96-1206-2_31.

[35] V. Braun and V. Clarke, "Conceptual and design thinking for thematic analysis," *Qualitative psychology,* vol. 9, no. 1, p. 3, 2022, doi: https://psycnet.apa.org/doi/10.1037/qup0000196.

[36] H. Tajfel, J. Turner, W. G. Austin, and S. Worchel, "An integrative theory of intergroup conflict," *Intergroup relations: Essential readings,* pp. 94–109, 2001. [Online]. Available: https://books.google.co.cr/books?hl=es&lr=&id=uk8AFpo3BnYC&oi=fnd&pg=PA94&dq=An+integrative+theory+of+intergroup+conflict&ots=qNL31Xfzw7&sig=IQhyUdWACF-_DG3UPdISF9KmIGw&redir_esc=y#v=onepage&q=An%20integrative%20theory%20of%20intergroup%20conflict&f=false.

[37] N. Yee and J. Bailenson, "The Proteus effect: The effect of transformed self-representation on behavior," *Human communication research,* vol. 33, no. 3, pp. 271–290, 2007, doi: https://doi.org/10.1111/j.1468-2958.2007.00299.x.

[38] J. Feng, H. Tan, W. Li, and M. Xie, "Conv2NeXt: Reconsidering Conv NeXt Network Design for Image Recognition," *2022 International Conference on Computers and Artificial Intelligence Technologies (CAIT),* pp. 53–60, 4–6 Nov. 2022 2022, doi: https://doi.org/10.1109/CAIT56099.2022.10072172.

[39] G. Bernal *et al.*, "Unraveling the dynamics of mental and visuospatial workload in virtual reality environments," *Computers,* vol. 13, no. 10, p. 246, 2024, doi: https://doi.org/10.3390/computers13100246.

[40] A. B. Stanescu and D. Goanta, "Immersive Tendencies and Physiological Responses in Virtual Reality: Enhancing Education with Biofeedback Tools," *ICERI2025 Proceedings,* pp. 8060–8070, 2025, doi: https://doi.org/10.21125/iceri.2025.2242.

[41] K. Zhang *et al.*, "Inclusive avatar guidelines for people with disabilities: Supporting disability representation in social virtual reality," *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems,* pp. 1–26, 2025, doi: https://doi.org/10.1145/3706598.3714230.

[42] J. Cearns, "Through the looking glass: heteronymic self and otherness in AI therapy," *Etnográfica. Revista do Centro em Rede de Investigação em Antropologia,* no. 29 (3), pp. 707–710, 2025, doi: https://doi.org/10.4000/154s8.

[43] C. R. Rogers and B. F. Skinner, "Some issues concerning the control of human behavior: A symposium," *Science,* vol. 124, no. 3231, pp. 1057–1066, 1956, doi: https://doi.org/10.1126/science.124.3231.1057.

[44] Z. Ding, Y. Huang, H. Yuan, and H. Dong, "Introduction to Reinforcement Learning," *Deep Reinforcement Learning: Fundamentals, Research and Applications,* pp. 47–123, 2020, doi: https://doi.org/10.1007/978-981-15-4095-0_2.

[45] F. Rowland, "The filter bubble: What the internet is hiding from you," *portal: Libraries and the Academy,* vol. 11, no. 4, pp. 1009–1011, 2011, doi: https://doi.org/10.1353/pla.2011.0036.

[46] A. Cavoukian, "Privacy by Design in Law, Policy and Practice A White Paper for Regulators," 2011. [Online]. Available: https://gpsbydesigncentre.com/wp-content/uploads/2022/02/312239.pdf.

[47] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 11966–11976, 18–24 June 2022 2022, doi: https://doi.org/10.1109/CVPR52688.2022.01167.

[48] T. Ma, W. Wang, and Y. Chen, "Attention is all you need: An interpretable transformer-based asset allocation approach," *International Review of Financial Analysis,* vol. 90, p. 102876, 2023/11/01/ 2023, doi: https://doi.org/10.1016/j.irfa.2023.102876.

[49] R. K. E. Bellamy *et al.*, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and*

Development, vol. 63, no. 4/5, pp. 4:1–4:15, 2019, doi: https://doi.org/10.1147/JRD.2019.2942287.

[50] A. Rahman, "Algorithms of oppression: How search engines reinforce racism," 2020, doi: https://doi.org/10.1177/1461444819876115.

[51] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue,* vol. 16, no. 3, pp. 31–57, 2018, doi: https://doi.org/10.1145/3233231.

[52] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems,* vol. 30, 2017. [Online]. Available: https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[53] R. S. Al-Mansoori, D. Al-Thani, and R. Ali, "Designing for digital wellbeing: From theory to practice a scoping review," *Human Behavior and Emerging Technologies,* vol. 2023, no. 1, p. 9924029, 2023, doi: https://doi.org/10.1155/2023/9924029.

[54] R. A. Calvo and D. Peters, "Positive computing: technology for wellbeing and human potential," 2014. [Online]. Available: https://books.google.co.cr/books?hl=es&lr=&id=uI6ZBQAAQBAJ&oi=fnd&pg=PR7&dq=Positive+computing:+Technology+for+wellbeing+and+human+potential&ots=KeiQVmd4Gx&sig=X2cGJ5g7xELB8RajEI_ZrKUqujY&redir_esc=y#v=onepage&q=Positive%20computing%3A%20Technology%20for%20wellbeing%20and%20human%20potential&f=false.

[55] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human–Computer Interaction,* vol. 36, no. 6, pp. 495–504, 2020, doi: https://doi.org/10.1080/10447318.2020.1741118.

[56] F. Santoni de Sio and J. Van den Hoven, "Meaningful human control over autonomous systems: A philosophical account," *Frontiers in Robotics and AI,* vol. 5, p. 323836, 2018, doi: https://doi.org/10.3389/frobt.2018.00015.

[57] H. C. Triandis, "Review of culture's consequences: International differences in work-related values," *Human organization,* vol. 41, no. 1, pp. 86–90, 1982, doi: https://doi.org/10.17730/humo.41.1.j673560x45803521.

[58] A. C. Bluedorn, "An interview with anthropologist Edward T. Hall," *Journal of Management Inquiry,* vol. 7, no. 2, pp. 109–115, 1998, doi: https://doi.org/10.1177/105649269872003.