GOVERNING GENERATIVE AI FOR SCALABLE MENTAL HEALTH TRAINING AND CLINICAL DECISION-MAKING

Philippe J. Giabbanelli[1], Megan A. Witherow[1], and Anca D. Dobrian[2]

[1]National Center for Collaboration in Medical Modeling & Simulation
and Office of Enterprise Research & Innovation, Old Dominion University, Suffolk, VA, USA
[2]Macon & Joan Brock Virginia Health Sciences, Old Dominion University, Norfolk, VA, USA

ABSTRACT

Driven by factors such as healthcare cost burden, poor mental health has been on the rise among aging adults for over a decade. Shortages in mental health services have heightened reliance on primary care physicians, underscoring the need for scalable, cost-effective mental health training. While generative AI (GenAI) can produce realistic images and videos for low-cost, scalable medical training, most GenAI tools are trained on non-curated internet data rather than clinically grounded portrayals. Governance mechanisms are needed to ensure GenAI systems produce clinically faithful mental health simulation content. We propose a governance-oriented, platform-agnostic framework to align AI-generated mental health content with clinically faithful portrayals and an educational study protocol to examine how governance influences interpretation, confidence, and proposed clinical actions.

1 INTRODUCTION

Poor mental health affects approximately one-quarter of adults over age 65 and more than one-third of adults aged 50–64, with both groups exhibiting gradual but persistent increases nationally since 2011 (Wright et al. 2025). Among risk factors, healthcare cost burden (particularly among low-income older adults with chronic conditions) can result in late-life symptoms of anxiety and depression such as excessive worry and reduced ability to feel pleasure (i.e., anhedonia) (Choi et al. 2024). Primary care physicians are often the front line of care for patients with mental health conditions, providing 45% of visits for patients with depression and/or anxiety (Robert Graham Center for Policy Studies in Family Medicine and Primary Care 2022). In Hampton Roads (served by Old Dominion University), drivers of anxiety and depression (e.g., socioeconomic vulnerability) co-occur with barriers to access mental health services (Faruque et al. 2025; Hodgkinson et al. 2017), increasing reliance on primary care physicians. The severe shortage of mental health professionals has led to an increased emphasis on primary care provider education in our region (Virginia Mental Health Access Program 2024). Clinical decision-making relies heavily on the ability to recognize affective cues, interpret psychological distress, and determine appropriate follow-up actions. Such demands raise questions about training clinical judgment and interpersonal skills at scale.

Training clinicians to recognize affective cues and respond appropriately to psychological distress requires repeated, structured practice with feedback over time (Schwartzkopf et al. 2025; Paulus and Meinken 2022). In medical education, this role is often filled by standardized patients (i.e., trained individuals portraying clinical scenarios), which have been shown to improve communication skills and empathic engagement. However, standardized patient programs are resource-intensive (Hamilton et al. 2024; Papanagnou et al. 2021), difficult to scale, and costly to sustain, particularly when training requires repeated interactions or longitudinal portrayals across multiple encounters. Consistency also presents a challenge, as maintaining stable affective portrayals across sessions and learners is difficult in practice. These limitations constrain the ability of training programs to provide scalable, standardized preparation for mental health care as demand continues to grow.

Recent advances in generative artificial intelligence (GenAI) enable the creation of photorealistic images, videos, and interactive simulations that are increasingly used in medical and nursing education (Ren et al. 2026; Jo et al. 2025). These tools support scalable exposure to diverse clinical scenarios, including communication-intensive and mental health–related encounters, at low marginal cost and with high reproducibility. Recent systems incorporate large language models and multimodal generation to simulate patient dialogue, affective responses, and longitudinal interactions, and have been deployed to support training in communication skills, clinical reasoning, and deliberate practice (Lee et al. 2025; Hicke et al. 2026). AI-based simulations are particularly attractive for educational contexts that require repeated practice and consistent portrayals, which are difficult to sustain using human actors alone. Conversely, these generative AI models are trained on large, non-curated collections of online data (Bianchi et al. 2023), which reflect popular representations rather than clinically grounded portrayals. As a result, AI-generated mental health content may systematically diverge from the affective patterns observed in clinical practice (Papazova et al. 2025a; Flathers et al. 2024), thus giving future physicians inaccurate expectations about symptom presentation and preventing them from learning appropriate judgments about follow-up and referral. For example, future practitioners may equate anxiety with panic attacks if GenAI outputs only popular images associated with conditions rather than covering the full (and sometimes) subtle spectrum of expressions. In sum, neither of the two options currently available is fully satisfactory: standardized patients offer tightly governed but costly and difficult-to-scale portrayals, while GenAI offers inexpensive and more consistent outputs that, without explicit governance, risk reproducing biased or clinically unfaithful representations.

To benefit from the scalability and consistency of GenAI systems in training future primary care physicians, we thus need *governance*: mechanisms that ensure generative AI systems produce clinically faithful portrayals of mental health conditions. In the context of mental health training for aging populations, governance entails the ability to steer AI-generated images and videos toward representations of stress, anxiety, and related affective states that align with clinical understanding rather than popular or stigmatized depictions. In principle, governance could be achieved by modifying individual GenAI systems through fine-tuning (i.e., retraining a model on specialized data) or through model-specific adapters (i.e., add-on components that alter how a given system generates outputs). However, these approaches must be repeated for each model and revisited as GenAI models are very frequently updated, making governance costly and potentially inconsistent. A lighter-weight alternative is prompt engineering, in which carefully crafted instructions guide model outputs toward more clinically appropriate depictions; however, prompt engineering remains model-dependent, its effects are difficult to standardize across platforms, and it provides no formal guarantee that outputs will meet clinical fidelity thresholds. In contrast, governance applied at the level of AI-generated outputs (i.e., as post-procesing) enables consistent oversight across heterogeneous and rapidly evolving GenAI platforms for clinically faithful portrayals of stress and anxiety among older adults. This output-focused approach aligns with the practical reality that healthcare and educational institutions may rely on multiple AI systems while requiring uniform standards for clinical fidelity.

In this paper, we propose a governance-oriented approach for aligning AI-generated mental health simulations with clinically faithful portrayals that can be applied across GenAI platforms. Our approach is rooted in AI techniques for alignment of content generation, as explained in Section 2, and supplemented by an evaluation plan with medical students at Eastern Virginia Medical School (now a part of Old Dominion University), which has a demonstrated track-record in simulation/standardized patient education. As detailed in Section 3, our protocol exposes participants to AI-generated mental health portrayals that either reflect broad, non-curated training data or have been aligned with clinically faithful affective patterns, allowing us to isolate the effects of representational quality on interpretation, confidence, and proposed clinical actions. That is, our protocol is designed to provide evidence on how governed versus ungoverned AI-generated content shapes judgments relevant to mental health care, without relying on patient-facing deployment.

2 CONCEPTUAL FRAMEWORK: GOVERNED VS UNGOVERNED AI REPRESENTATIONS

Recent advances in GenAI have enabled rapid growth in AI-based simulations for medical and nursing education, including applications to communication skills and mental health training (Ren et al. 2026; Jo et al. 2025). For example, AI-generated images have been used for health professions training to prompt reflection on stereotypes and professional roles (Stewart et al. 2025). Such systems have relied on AI-generated portrayals that are produced directly from broad, non-curated training data and incorporated into educational settings without systematic validation against clinical standards. We refer to such outputs as *ungoverned AI representations*. A substantial body of work in computer vision has documented that text-to-image generation systems systematically reflect and amplify biases present in their training data, including biases related to gender, race, age, and health-related attributes (Wan et al. 2024). In mental health contexts, these biases can manifest as exaggerated, stigmatized, or unrepresentative depictions of psychological distress. Empirical evidence shows that AI-generated images prompted with psychiatric conditions differ markedly from those associated with non-psychiatric conditions, reinforcing distorted portrayals of mental illness (Papazova et al. 2025b). As AI simulations scale in education, ungoverned representations risk standardizing mischaracterizations rather than clinically relevant understanding.

From a governance perspective, representational quality matters because repeated exposure to visual and affective cues shapes clinical judgment, expectations, and thresholds for concern. Prior work shows that AI-based simulations can influence learners' confidence, empathy, and clinical reasoning, with outcomes depending strongly on realism, consistency, and alignment with educational objectives (Hicke et al. 2026; Lee et al. 2025). In mental health care, facial expressions and affective signals constitute part of the clinical information used to assess severity, risk, and need for follow-up, particularly among older adults where symptoms of anxiety and depression may present subtly. Systematic distortions in these cues may therefore influence empathic orientation, interpretation of distress, and proposed clinical actions, even when AI is used only for training rather than patient-facing care. These concerns are compounded by evidence that AI-supported assessment of clinical performance varies substantially with representational choices and scoring frameworks, emphasizing the need for standardization (Geathers et al. 2025). In this setting, governance serves as a quality control mechanism preceding downstream deployment and decision support.

We define *governed AI representations* as AI-generated portrayals that are systematically evaluated and aligned with clinically faithful standards prior to use in healthcare training. Building on prior bias mitigation research in text-to-image generation, which has proposed model retraining, pipeline modifications, and post-processing corrections as viable strategies (Prerak 2024; Fu et al. 2025), we focus on governance at the level of outputs rather than individual models. Specifically, we propose to anchor governance in a curated, expert-labeled reference dataset of facial expressions associated with anxiety and depression, drawn from clinical video collections (e.g., Symptom Media (Mlakar et al. 2024; Iammarino et al. 2020)). Under our proposed framework, inputs could be photorealistic facial expressions generated using any text-to-image or text-to-video AI tool (e.g., Flux, Google Imagen/Veo, Kling, Wan). Then, standard distribution-level measures such as Wasserstein distance (Dimitrakopoulos et al. 2020), precision, recall, and F-scores (Kynkäänniemi et al. 2019) quantitatively compare these AI-generated expressions to expert-validated clinical examples to assess how closely the AI reflects clinically meaningful patterns. Based on these assessments, an automated correction layer will adjust AI-generated outputs to match clinically faithful affective portrayals, without requiring access to or modification of the underlying AI models. The proposed framework will use Generative Adversarial Networks (Goodfellow et al. 2020) for domain adaptation, which have shown success in aligning affective (Choi et al. 2018) and clinical content (Kirchhoff et al. 2025). In summary, we evaluate AI-generated images and videos against clinically meaningful affective patterns, then we apply an automated correction to produce governed representations.

Enforcing governance by post-processing of AI outputs is an established mitigation strategy for bias in text-to-image generation, particularly for demographic imbalances or stereotyped associations (Wan et al. 2024; Prerak 2024; Fu et al. 2025). Prior works have focused on improving representational fairness with generic benchmarks and fairness-oriented metrics. Our approach to GenAI governance differs

in several ways. First, we anchor governance to clinically validated reference standards, focusing on affective patterns associated with anxiety and depression rather than demographic balance alone. Second, we target affective meaning and clinical plausibility, which are central to mental health judgment. Third, we design governance as a platform-agnostic, output-level process that can be applied consistently across heterogeneous and frequently updated GenAI systems, addressing practical deployment constraints in healthcare and education. Finally, our mechanism of AI governance is directly relevant to workforce preparation and healthcare quality rather than merely as an isolated image correction technique.

3 EMPIRICAL DESIGN AND IMPLICATIONS FOR AI GOVERNANCE IN HEALTHCARE

Our protocol for pilot evaluation is to conduct a low-risk educational study with medical students across all four years of training ($n \approx 600$) at Eastern Virginia Medical School. Based on comparable studies at our institution (Berry et al. 2025), a conservative estimate of $\approx$10% participation rate ($n \approx 60$ students) provides adequate power to detect medium effect sizes (Cohen's $d \leq 0.40$) at $\alpha = 0.05$. Participants will be randomly assigned to a *governed condition*, receiving AI-generated stimuli corrected by our framework, or an *ungoverned condition*, receiving uncorrected AI-generated stimuli (i.e., based on non-curated online training data that may lack clinical fidelity). This between-subjects design isolates the educational effect of clinical alignment from mere exposure to affective content (Imbens and Rubin 2015). Both groups will complete identical tasks using facial expression stimuli generated from the same prompts and demographic specifications; the only difference is whether those stimuli have been processed through our correction layer (governed condition) or are presented as originally generated by the AI model (ungoverned condition).

At baseline and post-intervention, participants will complete validated measures including the 20-item Jefferson Scale of Empathy–Student Version (Hojat et al. 2001) (JSE-S) and Likert-scale items developed for this study to assess students' self-rated confidence in identifying facial expressions suggestive of psychological distress, deciding when clinical follow-up is warranted, and initiating an empathic response. Participants will complete *two assessment tasks to evaluate complementary aspects of clinical competence*. The *video task approximates real-world patient encounters*. Students will view 15-second facial expression clips depicting clinically concerning or unremarkable affect across diverse demographic profiles and respond as they would in a clinical encounter. The 15-second time window reflects both the practical output limits of current text-to-video GenAI models and the brief observational windows typical of initial clinical assessments (Dhall et al. 2011). These clips will be created by text-to-video GenAI models and, for the aligned condition, automatically corrected by our innovative approach. For each clip, students will rate the degree to which the displayed affect suggests a need for mental health evaluation, select the most suitable next clinical action, and select a verbal response to the patient, followed by a brief free-text justification. In the image vignette task, students will view sequences of still images depicting the same individual across multiple life contexts (e.g., home, work, clinical settings) and complete structured items to interpret affective cues and propose clinical actions. To control for order effects, half of participants within each condition will complete the video task first, and half will complete the image vignette task first. *These vignettes assess clinical reasoning based on affective information across situations, thus complementing the video tasks which simulate direct clinical encounters*.

Primary analyses will compare pre/post change scores between governed and ungoverned conditions using appropriate between-group tests, assessing differences in empathic orientation, self-efficacy, and clinical preparedness. Secondary analyses will examine between-group differences in task-based performance, including affective cue recognition and appropriateness of proposed clinical actions. We acknowledge that self-report and structured task responses provide *preliminary* evidence of educational impact. As noted in recent reviews of AI-based virtual patient simulations (Jo et al. 2025), current studies in this area mostly rely on self-report measures and lack validated performance metrics, limiting the ability to draw firm conclusions about clinical competence. Future work building on this framework will incorporate objective performance measures, such as Objective Structured Clinical Examination scores or standardized clinical assessments, to more rigorously evaluate the effects of governed AI representations on clinical skill development.

4 CONCLUSION

Although AI in healthcare and medical training has potential risks in amplifying biased or stigmatized portrayals of mental health conditions, we assert that robust governance can effectively mitigate these risks. We propose a governance framework that aligns AI-generated affective content with an expert-labeled reference data set of clinically faithful depictions of anxiety and depression, independent of the specific underlying GenAI model. Furthermore, we present a protocol for a low-risk educational study with medical students to assess how governed versus ungoverned AI-generated content shapes judgments relevant to clinical care. The proposed framework and study will support scalable, cost-effective mental health education, targeting an unmet need in primary care physician training.

ACKNOWLEDGMENTS

PJG and MAW are supported by funding provided by the Virginia Innovation Partnership Authority of the Commonwealth of Virginia and Old Dominion University's National Center for Collaboration in Medical Modeling and Simulation.

REFERENCES

Berry, A., A. Campbell, D. Li, C. Bay, and U. Ikonne. 2025. "An investigation of second-year medical students' use of outside resources at two institutions". *Medical Science Educator* 35(2):847–862.

Bianchi, F., P. Kalluri, E. Durmus *et al*. 2023. "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale". In *Proc. ACM Conf. on fairness, accountability, and transparency*, 1493–1504.

Choi, N. G., C. N. Marti, B. Y. Choi, and M. M. Kunik. 2024. "Healthcare cost burden and self-reported frequency of depressive/anxious feelings in older adults". *Journal of gerontological social work* 67(3):349–368.

Choi, Y., M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. 2018, June. "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation". In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Dhall, A., R. Goecke, S. Lucey, and T. Gedeon. 2011. "Acted facial expressions in the wild database". *Australian National University, Canberra, Australia, Technical Report TR-CS-11* 2(1).

Dimitrakopoulos, P., G. Sfikas, and C. Nikou. 2020. "Wind: Wasserstein Inception Distance For Evaluating Generative Adversarial Network Performance". In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 3182–3186.

Faruque, F., G. H. Shah, and R. M. Bohler. 2025. "The Association Between Social Determinants of Health (SDoH) and Mental Health Status in the US". *European Journal of Investigation in Health, Psychology and Education* 15(5):87.

Flathers, M., G. Smith, E. Wagner, C. E. Fisher, and J. Torous. 2024. "AI depictions of psychiatric diagnoses: a preliminary study of generative image outputs in Midjourney V. 6 and DALL-E 3". *BMJ Ment Health* 27(1).

Fu, Z., R. Brown, S. Shao, K. Rawal, E. Delaney, and C. Russell. 2025. "FairImagen: Post-Processing for Bias Mitigation in Text-to-Image Models". In *Advances in Neural Information Processing Systems*. NeurIPS 2025.

Geathers, J., Y. Hicke, C. Chan, N. Rajashekar, S. Young, J. Sewell, *et al*. 2025. "Benchmarking Generative AI for Scoring Medical Student Interviews in Objective Structured Clinical Examinations (OSCEs)". arXiv:2501.13957v2.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, *et al*. 2020, October. "Generative adversarial networks". *Commun. ACM* 63(11):139–144.

Hamilton, A., A. Molzahn, and K. McLemore. 2024. "The evolution from standardized to virtual patients in medical education". *Cureus* 16(10).

Hicke, Y., J. Geathers, K. Vu, J. Sewell, C. Cardie, J. Talwalkar, *et al*. 2026. "MedSimAI: Simulation and Formative Feedback Generation to Enhance Deliberate Practice in Medical Education". arXiv:2503.05793v2.

Hodgkinson, S., L. Godoy, L. S. Beers, and A. Lewin. 2017. "Improving mental health access for low-income children and families in the primary care setting". *Pediatrics* 139(1):e20151175.

Hojat, M., S. Mangione, T. J. Nasca, M. J. Cohen, J. S. Gonnella, J. B. Erdmann, *et al*. 2001. "The Jefferson Scale of Physician Empathy: development and preliminary psychometric data". *Educational and psychological measurement* 61(2):349–365.

Iammarino, C. S., J. Johnson, M. Zolezzi, and Z. S. Hasnani. 2020. "Simulation in mental health interprofessional education". *J Nurs Educ Pract* 10(46):10–5430.

Imbens, G. W., and D. B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Jo, Y.-W., M. Lee, and H.-J. Yang. 2025. "Large Language Model-Based Virtual Patient Simulations in Medical and Nursing Education: A Review". *Applied Sciences* 15:11917.

Kirchhoff, A., A. Hustinx, B. Javanmardi, T.-C. Hsieh, F. Brand, F. Hellmann, *et al*. 2025, Mar. "GestaltGAN: synthetic photorealistic portraits of individuals with rare genetic disorders". *European Journal of Human Genetics* 33(3):377–382.

Kynkäänniemi, T., T. Karras, S. Laine, J. Lehtinen, and T. Aila. 2019. "Improved precision and recall metric for assessing generative models". In *Proc. 33rd Int. Conf. on Neural Information Processing Systems*.

Lee, K., S. Lee, E. H. Kim *et al*. 2025. "Adaptive-VP: A Framework for LLM-Based Virtual Patients that Adapts to Trainees' Dialogue to Facilitate Nurse Communication Training". In *Findings of the Association for Computational Linguistics: ACL 2025*, 2319–2352.

Mlakar, I., U. Arioz, U. Smrke, N. Plohl, V. Šafran, and M. Rojc. 2024. "An End-to-End framework for extracting observable cues of depression from diary recordings". *Expert Systems with Applications* 257:125025.

Papanagnou, D., M. R. Klein, X. C. Zhang, K. A. Cameron, A. Doty, D. M. McCarthy, *et al*. 2021. "Developing standardized patient-based cases for communication training: lessons learned from training residents to communicate diagnostic uncertainty". *Advances in Simulation* 6(1):26.

Papazova, I., A. Hasan, and N. Khorikian-Ghazari. 2025a. "Biased AI generated images of mental illness: does AI adopt our stigma?". *European Archives of Psychiatry and Clinical Neuroscience*:1–3.

Papazova, I., A. Hasan, and N. Khorikian-Ghazari. 2025b. "Biased AI generated images of mental illness: does AI adopt our stigma?". *European Archives of Psychiatry and Clinical Neuroscience* 275:2563–2565.

Paulus, C. M., and S. Meinken. 2022. "The effectiveness of empathy training in health care: a meta-analysis of training content and methods". *Int. Journal of Medical Education* 13:1.

Prerak, S. 2024. "Addressing Bias in Text-to-Image Generation: A Review of Mitigation Methods". In *2024 Third Int. Conf. on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, 1–6. IEEE.

Ren, Y., Y. Wang, S. Dong *et al*. 2026. "Application of Artificial Intelligence in Medical Education: A Systematic and Narrative Review of Pedagogical Potential and Ethical Implications". *Advances in Medical Education and Practice*:1–13.

Robert Graham Center for Policy Studies in Family Medicine and Primary Care 2022. "The State of Integrated Primary Care and Behavioral Health". U.S.-focused report on integration of behavioral health into primary care.

Schwartzkopf, C. T., R. T. Alves, P. C. Lopes *et al*. 2025. "The role of training and education for enhancing empathy among healthcare students: a systematic review of randomised controlled trials". *BMC Medical Education* 25(1):469.

Stewart, M. P., S. H. Park, M. K. Sanders, A. Holland, J. Luk, S. Sloan *et al*. 2025. "Implementation of an AI activity to teach interprofessional roles and responsibilities". *Journal of the American College of Clinical Pharmacy* 8(8):816–822.

Virginia Mental Health Access Program 2024. "Virginia Mental Health Access Program (VMAP) Impact Report". Comprehensive statewide report on VMAP services, provider engagement, and impact on mental health access in Virginia.

Wan, Y., A. Subramonian, A. Ovalle, Z. Lin, A. Suvarna, C. Chance, *et al*. 2024. "Survey of Bias In Text-to-Image Generation: Definition, Evaluation, and Mitigation". *arXiv preprint arXiv:2404.01030*.

Wright, E., E. C. Dore, K. C. Koenen, C. Mangurian, D. R. Williams, and R. Hamad. 2025. "Prevalence of and inequities in poor mental health across 3 US surveys, 2011 to 2022". *JAMA Network Open* 8(1):e2454718.

AUTHOR BIOGRAPHIES

PHILIPPE J. GIABBANELLI is a Research Professor at Old Dominion University (ODU). He received a B.S. in Computer Science from the University of Nice (France) and holds M.S. and Ph.D. degrees from Simon Fraser University (Canada). He has held positions at several universities (e.g., the University of Cambridge) prior to being tenured at Miami University and then promoted to Full Professor at the Virginia Modeling, Analysis, and Simulation Center (VMASC). He authored over 180 articles, mostly in modeling & simulation and AI applied to human behavior. As principal investigator, his research on AI in health has been supported by organizations including the CDC, Microsoft AI for Health, and the NIH. His e-mail address is pgiabban@odu.edu.

MEGAN A. WITHEROW is a Research Assistant Professor at ODU. She received her Ph.D. in Electrical and Computer Engineering and her B.S. in Computer Engineering from ODU. She is certified in the Facial Action Coding System (FACS). Her research focuses on AI for affective and human-centered computing, with applications in mental health and healthcare education. She has published on facial expression analysis for clinical applications and has received support from the National Science Foundation. She contributes technical expertise on how AI-generated visual content represents human affect and how such representations can be evaluated and aligned with clinically meaningful patterns. Her e-mail address is mwithero@odu.edu.

ANCA D. DOBRIAN is a Professor and Senior Associate Vice President for Faculty Affairs and Professional Development at Macon & Joan Brock Virginia Health Sciences at ODU. An inducted Master Educator, she is experienced in curriculum design, course development, and competency-based assessment across all learner levels. Her educational scholarship includes publications in medical education and national leadership through multiple AAMC committees. Concurrently, she has secured over $7M in extramural funding as PI or Co-I from the NIH, DoD, and American Heart Association. Her e-mail address is dobriaad@odu.edu.