From Multimedia Content to Multimodal Agents

JIANNAN LI, Singapore Management University, Singapore

Generative artificial intelligence is making profound impact on multimedia content creation and consumption. This position paper discusses a potential future model of these activities, where multimodal intelligent agents select, recommend, and generate content on behalf of content creators in response to consumer input and behaviors. We identify technical challenges for building and evaluating such agents, design considerations for user interfaces and agent behaviors, and social implications for content creators. We hope this piece can open up conversations about one possible path of user generated content with the tempting goal of maximizing scalability and adaptability.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: Generative AI, User Generated Content, Agent

1 INTRODUCTION

User generated multimedia contents, such as videos, images, music, and podcasts, have gained immense popularity over the past twenty years. At the same time, online digital media are in constant evolution. While personalized content recommendation has seen great success in matching information with users [6], the recommended contents themselves are fixed after production. The emergence of generative artificial intelligence presents an opportunity to rethink the current rigid form of online media [8, 19]. The ability of these models to understand, manipulate, and generate visual, auditory, and textual information suggests the possibility to create malleable media that adapt to their consumers' needs by breaking a linear narrative and even the boundaries of a single 'episode' of content.

In this position paper, we envision a new form of online media delivered by agents that curate, recommend [12], and create [7, 18] contents in a dynamic and interactive manner on behalf of individual content creators. These agents understand explicit natural language requests and possibly implicit states of the audience [27], and respond by selecting or generating multimedia content based on existing work and behavior control commands from the content creators. In this sense, the agent acts as a *creative expression proxy* for the content creator. For example, the agent of a photographer may present new work of this artist, recommend their previous work based comments from viewers (e.g. "I really like the lighting here"), and generate 'possible work' upon requests (e.g. "I'd like to see the same composition but in a desert"). In another example, the agent of a vlog maker can compose vlogs customized for a viewer's interest by mixing clips from the maker's existing videos and synthesized, 'possible' vlogs.

While current online media, such as videos, images, and music recordings, present each piece in a static and isolated manner, a creative expression proxy agent represents the creative history and capacity of a content creator as a whole, and lends itself to interaction with the audience and interaction-informed adaption. Such sharp deviation from the traditional path warrants a closer look at the technical, design, and social factors around this potential form of media. The rest of this paper will focus on discussing the potential research opportunities in these three aspects.

2 TECHNICAL CHALLENGES

Recent advances in language, image, and video generative models have made much of the anticipated capabilities of a creative expression proxy agent possible, such as language/image/video understanding and generation [4, 5, 11]. However, it is still not clear how to capture the rich,

Author's address: Jiannan Li, jiannanli@smu.edu.sg, Singapore Management University, 80 Stamford Road, Singapore, 178902.

nuanced, and ever-evolving styles of content creators [1]. Although diffusion-based models work well for learning styles from curated data, they often ignore the fact that an artist's or other kind of creative's style changes over time and they may apply several distinct styles when treating different subjects [5]. As a result, such methods cannot model the creative agency and expression drive that a content creator may exercise when choosing one style over another [23]. They would also fail to capture how a creator's style could undergo subtle transformation every time they practice their craft.

While not seeing AI-generated work supersedes a content creator's own expressive power, we do anticipate future agents to possess the ability to emulate not just the overall styles but their human partner's potential creative decisions at any moment. To perform this alignment, the agent may regularly probe the content creator with generated work to gauge whether their partner would make something similar for the same topic [2, 10]. The agent could gather additional information by studying the improvements that their partner would apply to the generated work.

A relevant challenge is in evaluating agents' performances in emulating human content creators. We do not yet know definitely what constitutes 'good' performances, as the end goals of content creators and consumers when working with these future agents are still ambiguous. For example, a content creator may expect an agent to closely follow their styles, or rather to supplement their work, that is, offering stylistic or content elements different from but coherent with their work. While a technical problem on the surface, we believe such ambiguity could be examined through a social lens as well, as it is rooted in people's varying and evolving expectations of their relationship with AI [8, 21].

3 DESIGN CONSIDERATIONS

Involving an intelligent agent suggests rethinking the design for many, if not all, steps in the communication process between content creators and their audience. Below we discuss two design considerations for further exploration.

Audience Interface. Traditional online media (e.g. YouTube, Instagram, and TikTok) interfaces center around the content itself and tone down the presence of other components such as comments to get users fully engaged. However, recent intelligent-agent-infused tools tend to reserve a significant portion of their interfaces for interaction with the agents [24–26], often in the form of a conversation bar, due to the key role such interaction plays in the operation of these tools. We anticipate the interface design for future agent-infused online media channels face the tension between agent and content presence, both possibly competing for audience attention. At one end of the spectrum, we can expect designs resembling current interfaces, where the content, such as the video or image, takes the center stage and interactions with the agent happen at the peripheral (e.g. a search bar). At the other end, the presence of the agent can take a dominant role, as with the popular conversational agent ChatGPT, and interactions with the agent (often through conversations) drive the flow of content consumption. Where the ideal spot along this spectrum lies will likely depends on the type of the content and the characteristics of the audience, for example, to what extend they would like to control the consumption experience.

Agent Persona. As an interactive proxy for the content creator, an agent will display certain behavioral patterns, such as their use of language in conversations with content consumers and their choices of content to recommend in response to user requests. These patterns will collectively form the perceived persona of the agent [22], which may have a strong influence on the content consumers' experiences. Research on conversational agents has already suggested that the use of language could significantly impact people's willingness to continue to engage with and their trust towards an agent [13, 15, 16]. Furthermore in online media distribution, agents may make additional important decisions about content presentation, such as which strategies to apply for

recommendation and whether search or generation is to be prioritized when responding to a particular query. For example, an agent can choose to always return the content creator's work that closely matches the user's interest, or occasionally suggest something novel to reveal another facet of the content creator. This is another instance of the exploitation-exploration trade-off in recommender systems [3]. We can also imagine an agent that adopts different 'personality' traits—reactive or proactive, terse or eloquent—with its audience depending on the persona profile the content creator has chosen.

In the end, an agent's persona would constitute one part, and possibly a significant portion, of the content creator's image and brand, which are usually crucial for their career. Therefore, content creators should have the power to fully customize and control the behaviors of their proxy agents. Even with such power, one may argue that generative-AI-based agents are ultimately limited in their ability to emulate humans and therefore distort the public image of the content creator.

4 SOCIAL IMPLICATIONS

Having an intelligent agent with generation capabilities as the intermediary would likely pose many questions about what it means to be a content creator, in terms of anticipated output, relationship with their audience, and others.

In an age when computer programs can generate work that a content creator could produce, one may question whether the goal of many creative activities becomes steering these programs with new work instead of producing the work itself. In this conceived relationship, content creators become secondary to the agent that they feed and risk losing control over their work [28]. While machine-generated content is still often associated with a loss of authenticity today [20], many artists have embraced generative, non-deterministic content [8]. We could anticipate the birth of a new view of creative expression, which sees the production of a generative model that aligns with one's expression intents, rather than individual pieces of content, as the final outcome of certain creative activities. The mission of the content creators who adopt this view then becomes keeping the behavior of the model in synchronization with their own minds, through producing new training data, new prompts, or other tuning methods.

While some content creators might shift their purpose from making content to making models (agents), they could still face the possibility of losing the social identity. Their social identity as content creators and often Internet influencers help them connect with their audience and commercial and creative partners [9, 14, 17]. However, an agent that automatically responds to audience and partners could significantly weakens the role content creators themselves play in these connections. Questions remain to be answered about whether and how content creators can be given total control over not just their work, but also their valuable social networks.

5 CONCLUSION

The emergence of generative AI and agent-based interaction calls for a re-examination of the possible futures of user generated content. This position paper envisions a new form of content production and consumption, where an intelligent agent interact with the audience while selecting, recommending, and generating contents on behalf of a content creator. While this model could take adaptive content delivery to a new level, it involves drastic deviations from current user generated content ecology and raises new questions to be answered if it is ever to be implemented.

We discussed technical challenges, including building agents that emulate a human's creative expression decisions, and the evaluation of such agents' ability. Other questions include design decisions to be made about the interface and persona of the agent as displayed to the content consumers. Finally, this potential future comes with deeper unknown social implications around content creators' purposes and identity if agents are to be part of their professional life. Generative

AI is already altering the landscape of content creation and delivery. Despite much uncertainty and possibly controversy, we would like to put forward this idea as one version of the generative AI multiverse for discussion.

ACKNOWLEDGMENTS

We acknowledge the support of Ministry of Education Tier-1 Grant 22-SIS-SMU-092.

REFERENCES

[1] James S Ackerman. 1962. A theory of style. *The Journal of Aesthetics and Art Criticism* 20, 3 (1962), 227–237.
[2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).
[3] Andrea Barraza-Urbina. 2017. The exploration-exploitation trade-off in interactive recommender systems. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 431–435.
[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[5] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
[6] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender systems leveraging multimedia content. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–38.
[7] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.
[8] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. 2023. Art and the science of generative AI. *Science* 380, 6650 (2023), 1110–1111.
[9] Jacob Gardner and Kevin Lehnert. 2016. What's new about new media? How multi-channel networks work with content creators. *Business horizons* 59, 3 (2016), 293–302.
[10] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems* 26 (2013).
[11] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in Neural Information Processing Systems* 35 (2022), 8633–8646.
[12] Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. 2023. Recommender ai agent: Integrating large language models for interactive recommendations. *arXiv preprint arXiv:2308.16505* (2023).
[13] Yuin Jeong, Juho Lee, and Younah Kang. 2019. Exploring effects of conversational fillers on user perception of conversational agents. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.
[14] Anne Jerslev. 2016. Media times| in the time of the microcelebrity: celebrification and the YouTuber Zoella. *International journal of communication* 10 (2016), 19.
[15] Ahmet Baki Kocaballi, Liliana Laranjo, and Enrico Coiera. 2019. Understanding and measuring user experience in conversational interfaces. *Interacting with Computers* 31, 2 (2019), 192–207.
[16] Rafal Kocielnik, Raina Langevin, James S George, Shota Akenaga, Amelia Wang, Darwin P Jones, Alexander Argyle, Callan Fockele, Layla Anderson, Dennis T Hsieh, et al. 2021. Can I Talk to You about Your Social Needs? Understanding Preference for Conversational User Interface in Health. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. 1–10.
[17] Chun Lo, Emilie De Longueau, Ankan Saha, and Shaunak Chatterjee. 2020. Edge formation in social networks to nurture content creators. In *Proceedings of The Web Conference 2020*. 1999–2008.
[18] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. 2024. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048* (2024).
[19] Lev Manovich, Lev Manovich, and Emanuele Arielli. 2023. AI Image and Generative Media. *Artificial Aesthetics: A Critical Guide to AI, Media and Design* (2023).
[20] Anna Notaro. 2020. State-of-the-art: AI through the (artificial) Artist's Eye. *EVA London 2020: Electronic Visualisation and the Arts* (2020), 322–328.
[21] Antonio Pošćić and Gordan Kreković. 2020. On the human role in generative art: a case study of AI-driven live coding. *Journal of Science and Technology of the Arts* 12, 3 (2020), 45–62.
[22] Alisha Pradhan and Amanda Lazar. 2021. Hey Google, do you have a personality? Designing personality and personas for conversational agents. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. 1–4.

[23] Stephanie Ross. 2003. Style in art. *The Oxford handbook of aesthetics* (2003), 228–244.
[24] Steven I Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D Weisz. 2023. The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 491–514.
[25] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. 2024. LAVE: LLM-Powered Agent Assistance and Language Augmentation for Video Editing. *arXiv preprint arXiv:2402.10294* (2024).
[26] Jingxuan Wei, Shiyu Wu, Xin Jiang, and Yequan Wang. 2023. Dialogpaint: A dialog-based image editing model. *arXiv preprint arXiv:2303.10073* (2023).
[27] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864* (2023).
[28] Martin Zeilinger. 2021. *Tactical entanglements: AI art, creative agency, and the limits of intellectual property.* meson press.