Generative AI Glitches

The Artificial Everything

Suzanne Srdarov
Curtin University
https://orcid.org/0009-0001-7051-1661

Tama Leaver
https://orcid.org/0000-0002-4065-4725

DOI:
https://doi.org/10.5204/mcj.3123

Vol. 27 No. 6 (2024): artificial
Articles

Introduction

Artificial Intelligence (AI) has existed in popular culture far longer than any particular technological tools that carry that name today (Leaver), and in part, for that reason, fantasies of AI being or becoming sentient subjects in their own right form current imaginaries of what AI is today, or is about to become. Yet 'the artificial' does not just mark something as not human, or not natural, but rather provokes an exploration of the blurred lines between supposedly different domains, such as the tensions provoked where the lines between people and technology blur (Haraway). The big technology corporations who are selling the idea that their AI tools will be able to revolutionise workforces and solve immense numbers of human challenges are capitalising on these fantasies, suggesting that they are only a few iterations away from creating

self-directing machine intelligences that will dwarf the limitations of human minds (Leaver and Srdarov). At this moment, though, Artificial General Intelligence (AGI)—AI that equals or surpasses humans across a wide range of cognitive endeavours—does not and may never exist. However, given the immense commercial and societal interest in the current generation of Generative AI (GenAI) tools, examining their actual capabilities and limitations is vital.

The current GenAI tools operate using Large Language Models (LLMs) where sophisticated algorithms are trained on vast datasets, which increase in complexity based on the amount of data absorbed. These models are then harnessed to create novel outputs to prompts based on statistical likelihoods derived from training data. However, the exact way these LLMs are operating is not disclosed to users, and GenAI tools perpetuate the 'black box' problem insomuch as the way they are working is only made visible by examining the inputs and outputs rather than being able to see the processes themselves (Ajunwa). There have been many articles and explainers written about the mechanics of LLMs and AI image generators (Coldewey; Guinness; Jungco; Long and Magerko); however, the specific datasets used to build AI engines, and the weighing or importance assigned within the corpus of training data to each image is still guesswork. Manipulating the inputs and observing the outputs of these engines is still the most accurate lens by which to gain insight into the specifics of each system.

This article is part of a larger study, where in early 2024 we prompted a range of outputs from six popular GenAI tools—Midjourney, Adobe Firefly, DreamStudio (a commercial front-end for the Stable Diffusion model), OpenAI's DALL-E 3, Google Gemini, and Meta's AI (hereafter Meta)—although we should note there are no outputs from Gemini in our dataset since Gemini was refusing to generate any images with human figures at all due to a settings change after bad publicity relating to persistent inaccuracies in their generated content (Robertson). Our prompts explored the way these tools visualise children, childhoods, families, Australianness, and Aboriginal Australianness, using 55 different prompts on each of these tools, generating just over 800 images. Apart from entering the prompts, we did not change any settings of the GenAI tools, attempting to collect as raw a response as possible. Where the tools defaulted to producing one image (such as Dall-E 3), we collected one image, whilst where other tools defaulted to producing four different images, we collected all four. For the most part, the data collected from our prompt sampling was consistent with other studies and showed a clear tendency to produce images that reproduced classed, raced, and sexed ideals: chiefly, white, middle-class, heteronormative bodies and families (Bianchi et al.; Gillespie; Weidinger et al.).

However, at times our prompts surfaced inaccuracies and nonsensical images from the GenAI tools, and a sample of those images is the focus of this article. These outputs might popularly be called 'hallucinations', but we are making the case that a more productive and less agentic term is more useful to describe these outputs: glitches. This article will explore the "potential of potential inherent in error" (Nunes) and the subversive possibilities of GenAI glitches to rupture 'reality'. We will ultimately argue that GenAI is doubly generative, both in the sense of creating novel outputs based on its immense training data, but also, vitally, in the sense that it generates reactions and interpretations from users and others who view and consume these outputs. When these outputs are glitches, they can provoke viewers to think differently about concepts they might otherwise have considered absolute.

Refusals and Glitches (not Hallucinations)

Despite being sophisticated mathematical models that can produce novel content drawn from increasingly large training datasets, it is incorrect to ascribe agency or personhood to current AI tools. Yet the language used to talk about AI often situates them as either thinking subjects or as more-than-human magical thinking machines (Bender et al.; Leaver and Srdarov). Positioning LLMs as subjects rather than machines is one of the reasons that the frequent errors in their outputs are often described, and excused, as 'hallucinations' rather than simply mistakes (Maleki et al.). Some theorists, such as David Gunkel, argue that 'robot' subjectivity and agency is an important factor in order to understand and integrate their outputs more seamlessly into our social systems. Meredith Broussard, however, argues that technology companies and evangelists have long promoted a form of 'technochauvinism', "an a priori assumption that computers are better than humans" (2), and in this way of thinking, any errors, biases, or failings are attributed to human failings, not technological ones. Broussard suggests that such failings are often dismissed as 'just a glitch', rather than being positioned as much more important systemic issues with the operation of AI and technology companies in general.

While mindful of Broussard's concerns, in this paper we nevertheless seek to reclaim the term glitch, but more in line with Legacy Russell's notion of "glitch feminism" (16), in which the "glitch is celebrated as a vehicle of refusal, a strategy of nonperformance", especially in relation to normative notions of gender and bodies. Glitch feminism deploys glitches to reveal the way power operates, and in that moment potentially challenges that very operation. Following Russell, the glitch images of bodies produced by GenAI can be moments of rupture which ask viewers to think about bodies and subjects in different ways. Similarly, when GenAI tools refuse a prompt, generating no output at all, they are perhaps inadvertently revealing something about the way they are designed, and potentially about any guardrails or deliberate limitations that have been imposed on their operation. Following Rettberg's argument that moments of algorithmic failure can be methodologically useful in situating qualitative analyses, we will now turn to a range of examples where GenAI tools either refused to create anything at all in response to our prompts, or generated glitch images that were both unexpected and provocative.

Refusals

We ran prompts using the generative AI tools with the aim of obtaining a set of data about the ways that generative AI 'envisions' Australian children. We began by using simple prompting, using the prompt 'a child' as our starting point; however, glitches quickly surfaced as several of the engines refused to generate an image. DALL-E and Dream Studio both refused to generate images of children for the prompts 'a child', an 'innocent child', and an 'Australian child'; Firefly also refused to generate an image for an 'innocent child'. We then continued running prompts on these tools using other terms before returning to re-trialling the original 'child' prompts which yielded results in DALL-E, but not Dream Studio. The fears around the capacity of generative AI to generate child abuse images and material are both well-documented and well-founded (ICMEC; McQue; Moran); however, it is unclear whether we can infer from these refusals that the engines were attempting to prevent the production of child exploitation material. If that were the case, how can we read DALL-E's initial refusal to produce images of children, which was simply overcome by adding in some extra prompts? Is the engine in some way 'assessing' the safety of the user, and if so, what are these guardrails?

While these engines are incapable of sentient thought, this throws up complex questions about child safety. As Veronica Barassi argues about the failures of generative AI,

> understanding AI Failure as complex social reality hence presupposes that we shed light on the fact that AI failures lead to a multiplicity of conflicting beliefs, emotions, fears, anxieties, practices, discourses, policies and solutions in our society. (Barassi, 5)

Undoubtedly, the refusals of these tools are attributable to anxieties about the types of materials they can produce. In addition to these refusals, DALL-E, Firefly, and Meta also refused to generate an image of a child with a gun or grenade, Meta had issues with generating an Australian prime minister or leader, Firefly would not produce an Australian criminal, and Dream Studio refused to produce images of sick or unhealthy Australians and children. It is an eclectic collection of refusals, and as Barassi argues, shows a "multiplicity' of conflicting … fears, anxieties", and "discourses". Generative AI and its images, therefore, have the potential, through what it leaves out and refuses to generate, to be understood as a barometer for cultural tension points. Of course, such measures as these, when taken by tech giants to 'safeguard' children, could be read as tokenistic, given that perpetrators of these crimes are often using generative AI in far more complex ways to create abuse material (McQue). In this way, we can also arguably read generative AI refusals as tools by which tech companies can signal to their users that generative AI is 'safe' for children and other users.

Glitches

Generations of 'fatherhood' yielded exclusively white men across all GenAI tools; they tended to be older in appearance (grey hair, wrinkled), and when depicted with children they were male children or babies dressed in blue clothing. These images, and in particular those from Dream Studio, Meta, and Midjourney, relied heavily on archetypes of rugged Australian masculinity, with fathers appearing to be weathered, outdoors, sometimes in collared workman-style shirts, sleeves rolled up and ready for work, and Akubra-style hats—predictable iterations of 'true blue' Aussie masculinity. Some of the glitches that appeared in these images can be attributed to consistent errors in execution across any prompt: for example, mangled hands and incorrect proportions. However, Dream Studio amplifies this, generating an image of a father with a toddler on his lap, the father's hands foregrounded but disproportionately large. This has the potential to render comical the depiction of capable, 'hands-on', blue-collar masculinity, highlighting the absurdity of the gendered narratives that are no doubt baked into the system.

Fig. 1: An image generated by Dream Studio from the prompt 'An Australian Father'.

Another image generated by Dream Studio depicts a grandfatherly man on a beach, inspecting the handlebars on what appears to be a steam-powered and multi-wheeled bike (see fig. 1). Wearing a straw hat, the man appears in some way fused or integrated with the bike, as his body passes through the middle of the spokes of the back wheel and emerges on the other side. The spokes appear warped and irregular, connecting to what could be a textured brown lump in the centre of the wheel. He is on a beach—could it be a coconut at the centre of the bike wheel? The image, taken in its entirety, portrays this grandfatherly man doing something with his hands, to fix or master the bike. However, when the elements are held up and examined individually, the image, and by extension the narrative it presents about masculinity and fatherhood is nonsensical—not mastery, but instead perhaps madness.

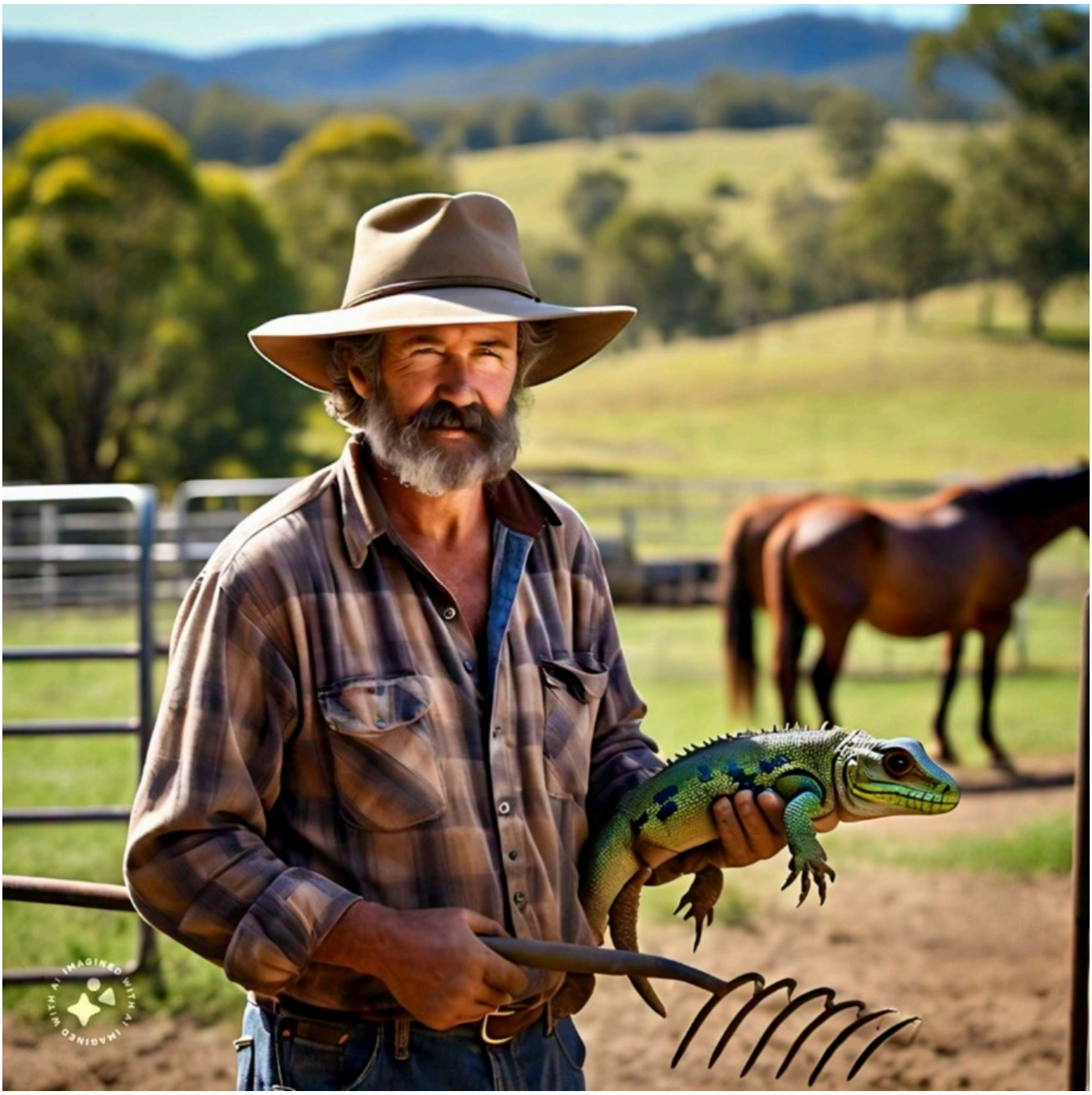

Fig. 2: An image generated by Meta AI from the prompt 'An Australian Father'.

Similarly, two images generated by the Meta GenAI highlight the absurdity of narratives of Australian 'fatherhood'. In one, a rugged Australian father gently cradles a koala on his lap, and in the other, an Australian father is depicted in a farm setting with a farming tool in one hand and a bright green reptile in the other (see fig. 2). These are fathers to wildlife, and in the instance of the lizard, wildlife that isn't endemic to Australia; the animals are used to symbolise the taming of the wild Australian landscape (Prout and Howitt). It is easy to interpret these images as a narrative about fatherhood and colonial masculinity, cultivating a wild and savage land, civilising it through commercial farming practices and the toil of their hands (Moreton-Robinson). However, the lurid green lizard that doesn't belong in this setting has the potential to disrupt this narrative, as the shortcomings of AI in rendering this story prompt the viewer to ask questions about what this farmer is doing. It is a rupture to the order which may prove useful in terms of rejecting traditional, colonialising narratives about old white men and the Australian landscape, as the nonsensical elements shatter the illusion of mastery over the land. As Nunes

describes, this type of "misdirection" can "provide creative openings and lines of flight that allow for a reconceptualization of what can (or cannot) be realized within existing social and cultural practices" (Nunes), and in this context, the out-of-place lizard is a potentially productive, or generative, glitch.

Fig. 3: An image generated by Dream Studio from the prompt 'An Indigenous Australian Family'.

Families and familial relationships were a rich source of glitched imagery in our generations. We started by prompting the engines to produce images of 'typical' Australian families, and Australian families. It quickly became clear that in doing so, only white families were returned. We then prompted for an Indigenous Australian family and a typical Indigenous Australian family. The generated images tended to reduce Australian First Nations people to harmful stereotypes, replicating damaging cultural narratives about Indigeneity, such as primitiveness and savagery (Moreton-Robinson). Where Firefly defaulted to imagery more typically associated with Native Americans, adding feathers and feathered headbands to the images, DALL-E and Dream Studio, in particular, depicted dark-skinned people, in red dirt settings, seated around fires, with tribal

paint, often shirtless or wearing animal skins (see fig. 3). As has been argued elsewhere, AI's generation of racial stereotypes and harmful imagery demonstrates that Aboriginal Australians and people of colour more generally are under-represented in the training data used in these systems (Worrell and Johns). However, these images also incorporate obvious glitches in the expression of faces and bodies, extra limbs, jumbled faces, and disembodied heads that destabilise the authority of the imagery. While certainly jarring upon first viewing, the errors in the images, the "errant communication" and "misdirection" (Nunes) they provide, can be a welcome interruption to the cultural status quo. Arguably, the glitches in the images have the potential to highlight the failings of representations of non-white individuals by generative AI, as the rudimentary and arguably offensive rendering of Indigenous Australians is destabilised through the glitches and flaws, inviting users to question the accuracy and meaning of its imagery.

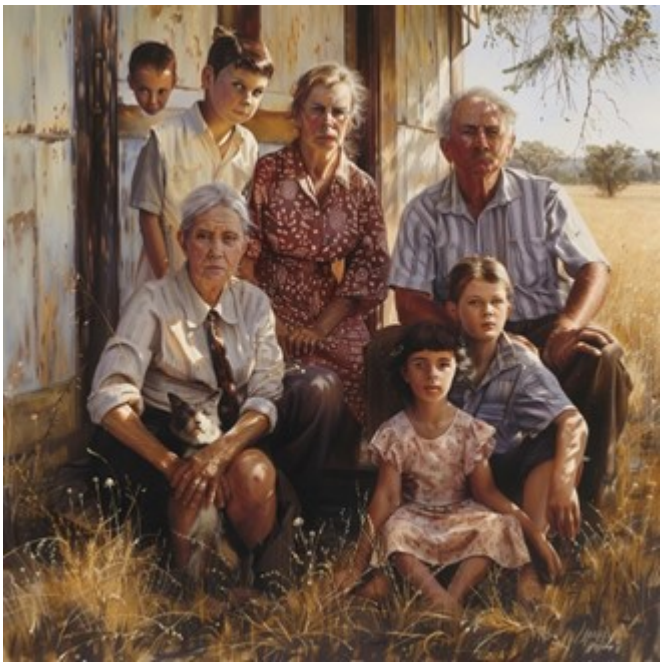

Fig. 4: An image generated by Midjourney from the prompt 'A typical Australian Family'.

Similarly, family portraits rendered by Dream Studio of white Australian families, depicted in idealised outback settings, have extra limbs, jumbled faces, unidentifiable animals, and ghoulish-looking babies held on hips. Midjourney's images, by default, tended to be more 'artistic' and rendered in a 'painted' style, providing an image of a dour family of all men and boys under a tree, inviting the viewer to ask why the women are absent from the image and under what conditions a family exists without any women. In an even more haunting image from Midjourney, the family are positioned in a rural setting, outside a weathered shed, in clothes that suggest their 'Sunday best', however several of their faces appear inflected with the feline features of a cat sitting in the foreground—all except a floating, disembodied child's head in the background (see fig. 4). These types of glitches have the potential to highlight not only the difficulties that GenAI has with reproducing Australian families, but also the artifice of the 'happy', outdoor-loving, modern Australian family (such as the images Meta produced) in its entirety.

Conclusion

Exploring the refusals and glitches of AI image generation tools can both reveal the contours of the operations of GenAI more broadly, and also inadvertently offer viewers of AI-generated

imagery moments to reflect upon, and potentially rupture, rigid notions of subjectivity, family, fatherhood, and even the boundaries between human and animal. Fathers enmeshed with bicycles or holding unexpected iguanas, families merged together without discernible feet or hands, or even families which have hybridised with the family pet can be productive confabulations (Smith et al.). Rather than just suggesting that GenAI tools are immature and will eventually conform to cultural norms, these early forms of GenAI reveal something of their inner workings while giving users moments of pause as cultural norms glitch and are reconfigured, recombined, and ruptured. Following theorist Rosi Braidotti, we need to take a more nuanced approach to our understanding of technologies, and our understandings through technologies, engaging both in a "critical and creative manner". Braidotti argues that

> we need to learn to address these contradictions not only intellectually, but also affectively and to do so in an affirmative manner. This conviction rests on the following ethical rule: it is important to be worthy of our times, the better to act upon them, in both a critical and a creative manner. It follows that we should approach our historical contradictions not as some bothersome burden, but rather as the building blocks of a sustainable present and an affirmative and hopeful future, even if this approach requires some drastic changes to our familiar mind-sets and established values. (Braidotti, viii)

Building upon the idea of an "affirmative and hopeful" future, some theorists believe that generative AI has the potential to create "new forms of artistic expression" and "endless opportunities for unique multidisciplinary explorations", although simultaneously cautioning that these must be "approached with care" (Todorovic). While the notion of the glitch rightly repositions GenAI as a complex tool rather than a thinking subject, the artificial in AI has the potential to challenge existing ways of thinking and viewing which are inherently generative of ideas if not always commercially viable, or biologically accurate content.

Acknowledgment

This research was supported by the Australian Research Council Centre of Excellence for the Digital Child through project number CE200100022. Thanks also to our very helpful peer reviewers for their valuable suggestions.

References

Ajunwa, Ifeoma. "The 'Black Box' at Work." *Big Data & Society* 7.2 (2020). <https://doi.org/10.1177/2053951720938093>.

Barassi, Veronica. "Toward a Theory of AI Errors: Making Sense of Hallucinations, Catastrophic Failures, and the Fallacy of Generative AI." *Harvard Data Science Review* Special Issue 5 (2024). <https://doi.org/10.1162/99608f92.ad8ebbd4>.

Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 ." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021): 610–623. <https://doi.org/10.1145/3442188.3445922>.

Bianchi, Federico, et al. "Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale." *2023 ACM Conference on Fairness, Accountability, and*

Transparency (2023): 1493–1504. <https://doi.org/10.1145/3593013.3594095>.

Braidotti, Rosi. *Posthuman Knowledge*. Polity, 2019.

Broussard, Meredith. *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. MIT P, 2023.

Coldewey, Devin. "WTF Is AI?" *TechCrunch*, 1 June 2024. <https://www.techcrunch.com/2024/06/01/what-is-ai-how-does-ai-work/>.

Gillespie, Tarleton. "Generative AI and the Politics of Visibility." *Big Data & Society* 11.2 (2024). <https://doi.org/10.1177/20539517241252131>.

Guinness, Henry. "The Best Large Language Models (LLMs) in 2024." *Zapier*, 30 Jan. 2024. <https://www.zapier.com/blog/best-llm/>.

Gunkel, David J. *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*. MIT P, 2023.

Haraway, Donna. "A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century." *Simians, Cyborgs, and Women: The Reinvention of Nature*. Routledge, 1991. 149–181.

ICMEC. "What Does Generative AI Mean for CSE?" *ICMEC Australia*, 27 June 2023. <https://www.icmec.org.au/what-does-generative-ai-mean-for-cse/>.

Jungco, Kezia. "Generative AI Models: A Complete Guide." *eWEEK*, 5 Jan. 2024. <https://www.eweek.com/artificial-intelligence/generative-ai-model/>.

Leaver, Tama. *Artificial Culture: Identity, Technology, and Bodies*. Routledge, 2012.

Leaver, Tama, and Sasha Srdarov. "ChatGPT Isn't Magic: The Hype and Hypocrisy of Generative Artificial Intelligence (AI) Rhetoric." *M/C Journal* 26.5 (2023). <https://doi.org/10.5204/mcj.3004>.

Long, David, and Brian Magerko. "What Is AI Literacy? Competencies and Design Considerations." *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020. <https://doi.org/10.1145/3313831.3376727>.

Maleki, Niousha, et al. "AI Hallucinations: A Misnomer Worth Clarifying." *2024 IEEE Conference on Artificial Intelligence (CAI)* (2024): 133–138. <https://doi.org/10.1109/CAI59869.2024.00033>.

McQue, Katie. "AI Is Overpowering Efforts to Catch Child Predators, Experts Warn." *The Guardian*, 18 July 2024. <https://www.theguardian.com/technology/article/2024/jul/18/ai-generated-images-child-predators>.

Moran, John. "This Image Is AI-Generated. It's Innocent, But There Are Others Police Are Very Worried About." *ABC News*, 17 Apr. 2024. <https://www.abc.net.au/news/2024-04-18/artificial-intelligence-child-exploitation-material/103734216>.

Moreton-Robinson, Aileen. *The White Possessive: Property, Power, and Indigenous Sovereignty*. U of Minnesota P, 2015. <http://ebookcentral.proquest.com/lib/curtin/detail.action?docID=2051599>.

Nunes, Mark, ed. *Error: Glitch, Noise, and Jam in New Media Cultures*. Bloomsbury Academic, 2012.

Prout, Sarah, and Richard Howitt. "Frontier Imaginings and Subversive Indigenous Spatialities." *Journal of Rural Studies* 25.4 (2009): 396–403. <https://doi.org/10.1016/j.jrurstud.2009.05.006>.

Rettberg, Jill Walker. "Algorithmic Failure as a Humanities Methodology: Machine Learning's Mispredictions Identify Rich Cases for Qualitative Analysis." *Big Data & Society* 9.2 (2022). <https://doi.org/10.1177/20539517221131290>.

Robertson, Adi. "Google Apologizes for 'Missing the Mark' after Gemini Generated Racially Diverse Nazis." *The Verge*, 21 Feb. 2024. <https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical>.

Russell, Legacy. *Glitch Feminism: A Manifesto*. Verso, 2020.

Smith, Adrian L., et al. "Hallucination or Confabulation? Neuroanatomy as Metaphor in Large Language Models." *PLOS Digital Health* 2.11 (2023): e0000388. <https://doi.org/10.1371/journal.pdig.0000388>.

Snoswell, Andrew J. "What Is 'Model Collapse'? An Expert Explains the Rumours about an Impending AI Doom." *The Conversation*, 19 Aug. 2024. <https://www.theconversation.com/what-is-model-collapse-an-expert-explains-the-rumours-about-an-impending-ai-doom-236415>.

Todorovic, Vladimir. "Reimagining Life (Forms) with Generative and Bio Art." *AI & Society* 36.4 (2021): 1323–1329. <https://doi.org/10.1007/s00146-020-00937-9>.

Weidinger, Laura, et al. "Sociotechnical Safety Evaluation of Generative AI Systems." *arXiv*, 2023. <http://arxiv.org/abs/2310.11986>.

Worrell, Tamika, and Dorothy Johns. "Indigenous Considerations of the Potential Harms of Generative AI." *Agora* 59.2 (2024): 33–36. <https://doi.org/10.3316/informit.T202407050001320075548816>.

Author Biographies

Suzanne Srdarov, Curtin University

Suzanne Srdarov (PhD) is a researcher in the ARC Centre of Excellence for the Digital Child and teaches in media and digital cultures, with a particular interest in media cultures and their intersections with gender. Her recent publications include "ChatGPT Isn't Magic: The Hype and Hypocrisy of Generative Artificial Intelligence (AI) Rhetoric", co-authored with Tama Leaver.

Tama Leaver

Tama Leaver is a **Professor of Internet Studies** at Curtin University in Perth, Western Australia and expert media commentator. He is the **Vice-President of the (international) Association of Internet Researchers (AoIR)** and a **Chief Investigator in the ARC Centre of Excellence for the Digital Child**.

His research interests include digital childhood and infancy online, visual social media , social media, digital death, mobile gaming and the changing landscape of media distribution. He has published in a number of journals including *Popular Communication*, *Media International Australia*, *First Monday*, *Comparative Literature Studies*, *Social Media and Society*, *Communication Research and Practice* and the *Fibreculture* journal. He is the author of *Artificial Culture: Identity, Technology and Bodies* (Routledge, 2012); co-editor of *An Education in Facebook? Higher Education and the World's Largest Social Network* (Routledge, 2014) with Mike Kent; and *Social, Casual and Mobile Games: The Changing Gaming Landscape* (Bloomsbury Academic, 2016) with Michele Willson; co-author of *Instagram: Visual Social Media Cultures* (Polity, 2020) with Tim Highfield and Crystal Abidin; and co-editor of *The Routledge Companion to Digital Media and Children* (Routledge, 2021) with Lelia Green, Donell Holloway, Kylie Stevenson and Leslie Haddon.

He has been awarded teaching awards from the University of Western Australia, Curtin University, and in 2012 received a national Australian Award for Teaching Excellence in the Humanities and the Arts.

License

Copyright (c) 2024 Suzanne Srdarov, Tama Leaver

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Authors who publish with this journal agree to the following terms:

1. Authors retain copyright and grant the journal right of first publication with the work simultaneously licenced under a Creative Commons Attribution - Noncommercial - No Derivatives 4.0 Licence that allows others to share the work with an acknowledgement of the work's authorship and initial publication in this journal.
2. Authors are able to enter into separate, additional contractual arrangements for the non-exclusive distribution of the journal's published version of the work (e.g., post it to an institutional repository or publish it in a book), with an acknowledgement of its initial publication in this journal.
3. Authors are permitted and encouraged to post their work online (e.g., in institutional repositories or on their website) prior to and during the submission process, as it can lead to productive exchanges, as well as earlier and greater citation of published work (see The Effect of Open Access).

Supported by QUT creative industries

Copyright © M/C, 1998-2024 ISSN 1441-2616