UNRELIABLE NARRATOR: REPARATIVE APPROACHES TO HARMFUL BIASES IN AI STORYTELLING FOR THE HE CLASSROOM AND FUTURE CREATIVE INDUSTRIES

David Jackson
PhD in Digital Storytelling Platforms, Manchester Metropolitan University (Manchester/Inglaterra). Senior Lecturer, Digital Visualisation, Manchester Metropolitan University.
E-mail: d.j.jackson@mmu.ac.uk

Marsha Courneya
MA from Internationale filmschule Koeln (International Film School of Cologne) (Germany/Alemanha).
Senior Research Assistant at SODA (School of Digital Arts) (Manchester/Inglaterra).
E-mail: m.courneya@mmu.ac.uk

Recebido em: 5 de abril de 2023
Aprovado em: 19 de junho de 2023
Sistema de Avaliação: Double Blind Review
BCIJ | v. 3 | n. 2 | p. 59-75 | jul./dez. 2023
DOI: https://doi.org/10.25112/bcij.v3i2.3540

Brazilian Creative Industries Journal
ISSN: 2763-8677

INTRODUCTION

Generative AI has the potential to amplify marginalised storytellers and their narratives through powerful virtual production tools and automation of processes such as artworking, scriptwriting and video editing (Ramesh et al., 2022; Brown et al., 2020, Esser et al, 2023). However, adoption of generative AI into media workflows and outputs risks compounding cultural biases from dominant storytelling traditions. Generative AIs typically require the input of many millions of novels, screenplays, images and other media to generate their synthetic narrative output. Stories produced can then contain biases from these texts through stereotypical character tropes, dialogues, word-image associations and story arcs (Bianchi et al., 2022). Whilst there is significant discussion of these biases, little exists to date on how we prepare storytellers for the problems of generative AI in production. How can we engage without further isolating marginalised storytellers, and in a way that encourages new voices to be heard?

The paper examines the potential issues of AI generative technologies for marginalised students in the creative education sector and provides case studies that provide a pathway towards a reparative approach by creative producers and educators. It provides an introduction to some of the issues arising from the reproduced biases of these AI systems and suggests potential strategies to incorporate awareness of these biases into the creative process. In order to evidence and illustrate our approach, two short case studies are provided: the Algowritten AI short story project led by the authors with other volunteers as a part of Mozilla Foundation's Trustworthy AI to identify patterns of bias in AI written narratives and a novel *reflective* AI system called Stepford, which is designed to highlight instances of gender bias in generative text segments. Both case studies are intended to outline how reparative approaches to algorithmic creative production can seek to highlight and mitigate cultural biases endemic in generative media systems.

The term generative AI can be broadly applied to a number of technologies and associated practices. We will seek to focus debate on a type of generative AI known as the large language model (LLM), using ChatGPT and other GPT-like models as examples. Artificial Intelligence models broadly fit into two fields - discriminative and generative (Babcock & Bali, 2021). Discriminative AI models use machine learning to differentiate between different inputs for the purposes of classification. For example a facial recognition model is trained to discern eyes, mouth and other facial features and even infer expressions and emotional states. In contrast, generative AI models are able to generate new data and can output digital media such as music, video, images and written work. Reviews and practical research of language models has found that the scaling size of language models, in terms of amount of computation, number of model parameters, and training dataset size, leads to predictable improvements in model performance

ISSN: 2763-8677

on certain natural language processing tasks such as text completion. In addition, as the scale of the model increases it develops *emergent capabilities* that are not present in smaller models (Wei et al., 2022). One of those capabilities is something called few-shot prompting which is characterised as an ability to infer textual context (e.g. text message or scientific paper) and respond with a stylistically appropriate text output on the basis of very small amounts of input or prompt text.

Conversational interaction design in LLMs such as OpenAI's GPT series has had a profound impact on the use and perception of artificial intelligence for both general and creative industry users. GPT-3, the precursor to Open AI's ChatGPT, was broadly available to the public and beta audiences for over a year before the updated version was launched. Whilst there were improvements made to the accuracy and safety of the ChatGPT model, a key difference was that it was redesigned to respond in the style of a conversational agent. GPT-3's initial interaction design, based on the historical notion of generative AI, relates to Markov chains and stochastic predictive modelling (Sericola, 2013). In this model, seed tokens prime the generation of text randomly, but within a range of predictive parameters based on machine learning. Interaction with these models involves iterating a prompt until it produces a desired output and is not inherently conversational. On the other hand, ChatGPT takes culturally from the history of conversational agents and Turing models of artificial intelligence, a history of 'enchanted determinism' (Campolo & Crawford, 2022) that seeks to give human-like agency and creative and intellectual autonomy to computers. Conversational models typically follow a turn-taking interaction style. They build successive interactions between user and machine into its prompts, creating an illusion of agent memory and consistency, similar to human conversation. The conversational model of AI is intuitive in an interactive sense because it is analogous of human interaction. It is also intuitive conceptually because of its legibility as a form of AI, predicted not only in science fiction but also throughout the history of non-human intelligence (Campbell et al., 2020): human-like conversational intelligences are a popular vision of AI's future and deliver on its historical mythos (such as represented by HAL 9000 in *2001: A Space Odyssey* [1968], K.I.T.T. in *Knight Rider* [1982-1986], Ziggy in *Quantum Leap* [1989-1993], J.A.R.V.I.S. in *Iron Man* [2008], and Samantha in *Her* [2013]). The change from classic generative AI design to intelligent conversational agent design, though mainly cosmetic, has immediately changed the primary form of interaction with LLMs from an iterative one to a turn-taking one. The change provides new opportunities for creative production as well as new and compounded risks in terms of ethics, novelty and quality. The following section will consider the ways in which creative producers and their audiences are potentially put at risk by using AI systems and focuses on the threat of harmful bias.

Brazilian Creative Industries Journal
ISSN: 2763-8677

RISKS OF USING LLMS

LLMs have a number of ethical and safety risks that are also shared with many other advanced computing technologies. The Mozilla Foundation (REF), a large charitable organisation based in the US that focuses on Internet health and safety internationally and which the researchers have worked with on trustworthy AI projects, summarises these risks as relating to privacy, such as the use of personal data by artificial intelligence systems without respect to privacy; fairness: the reliance of artificial intelligence on large sets of data that contain intrinsic cultural biases that unfairly impact users; trust, related to the responsible use of artificial intelligence in a way that give users agency; safety: designing and implementing barriers to misuse by bad actors in AI products and finally; transparency, which prioritises explicability and openness in presenting artificial intelligence decisions and outputs so that people understand their reliability and potential for harm (Mozilla Foundation, 2023). Whilst all of these issues are urgent, the paper will focus on fairness and algorithmic bias, particularly because it disproportionately impacts marginalised creators and audiences.

There are a number of reasons why large language models (LLM) and the texts that they generate are problematic in the context of cultural change and decolonisation. In the paper On the Dangers of Stochastic Parrots, the authors (Bender et al., 2021) note that the size of data required to train LLMs results in a need for "large, uncurated, Internet-based datasets" which in turn "encode the dominant/ hegemonic view, which further harms people at the margins" (ibid:613). Factors that intensify bias in internet-based datasets used to train LLMs relate to the composition of the English-speaking internet itself "overrepresenting younger users and those from developed countries" and those sites with the most freely available text, such as Twitter and Reddit which have a disproportionately male distribution of users. Importantly, the problem of bias in language models, as in other forms of machine learning, is not a problem of scale. In fact, rather than get better with more data, a large survey of these models found that bias increases with scale (Srivastava et al., 2022). In other words, to date, the more diverse the data is that the language model learns from, the more it provides socially biased outputs related to harmful contexts of race, sex, gender and age, particularly in 'broad or ambiguous contexts' (ibid:16). The shifting nature of language can make it difficult to pin down bias within a system of meaning that is not static (Mills, 2008: 124), especially as discriminatory forms of language are dependent on the context of who uses them and toward whom. The way language is used within a culture is not suited to an on/off, binary designation of harm or harmlessness, rather it is a 'dynamic entity' that also reflects the general instability of forms of oppression such as sexism, meaning "that there are difficulties in interpreting utterances and texts as unequivocally sexist" (Mills, 2008: 152). The policing of potentially harmful

ISSN: 2763-8677

language at scale through computational analysis could interfere with the reclamation and repurposing of that language by members of the marginalised communities it was weaponised against.

Design and testing of these tools often de-prioritises ethical concerns (Bender et al. 2021), being driven by more commercial incentives, and there is currently no rigorous code of ethics or set of standards that AI systems in particular are beholden to. Therefore, it is easy to imagine, as Kate Crawford, principle researcher at Microsoft states, that "AI systems are built to see and intervene in the world in ways that primarily benefit the states, institutions and corporations that they serve" and are "expressions of power" that reflect existing socio-political ecosystems (Crawford, 2021:211). In collusion with these forces, is the tendency to think in terms of the *enchanted determinism* of AI, noted by Campolo and Crawford: that due to the self-training nature of machine learning, machine learning researchers in the area feel relieved of responsibility for computational outputs that pertain to the social doman, claiming "high levels of accuracy and objectivity for systems that are simultaneously beyond human understanding or explanation" (Campolo & Crawford, 2020: 12). In this state, notions of accountability for existing and future development of AI systems are considered diminished because the AI systems produced use a logic that is outside of any human capacity to understand and review. Meanwhile, a notion of computational accuracy in complex social domains is maintained that can be harmful to its marginalised users without the need to be measurable and transparent.

IMPACT ON CREATIVE MEDIA STUDENTS

The lack of accountability or explicability of LLMs despite presents specific challenges in the context of higher education. In the UK as in other parts of the world, universities are responding to a call to decolonise the knowledge taught and researched in higher education (Ndlovu-Gatsheni, 2016). The process of 'decolonising the curriculum' (Arshad, 2021:online) involves going beyond providing representation from a broader set of sources and viewpoints (i.e. diversification), to actively redefining Eurocentric histories of knowledge towards more valid global histories and narratives. The process is expected to involve innovation, change and renewal to be effective. How can creative media students be taught to use these tools whilst also engaging in the decolonisation of curriculum? Partly due to the velocity of change in generative technologies, there is very little research available to educators to answer such questions. Existing research into AI in a higher education (HE) context has previously focused on the use of other types of AI to enhance teaching processes: a systematic review conducted by Ouyang, Zheng and Jiao in 2022 of AI educational tech in online higher education from 2010 to 2020 found that it focused on "performance prediction, resource recommendation, automatic assessment, and improvement of

ISSN: 2763-8677

learning experiences" (Ouyang et al, 2022: 7893). Due to a lack of foundational research, educators and students are at a real risk of either engaging with AI technologies without an understanding of the risks to marginalised people or ignoring or discounting the use of AI generative technologies altogether. Most recently, the Russell Group universities in the UK released a Principles in the Use of AI Tools in Education which notes that generative AI tools "may contain societal biases and stereotypes". These principles present understanding of generative AI as potentially biased is presented as a kind of *AI literacy*, a reasonable step towards more informed engagement. However, as we discuss below further reparative action may be required for these tools not to derail efforts to decolonise curriculum.

Whilst it may be easier for students to learn to avoid *biases and stereotypes* produced by tools when seeking answers to knowledge-related questions (e.g. scientific enquiries and literature searches), avoiding them in the production of narrative and creative media is more difficult. Storytelling plays an important role in humanity's ability to organise "its understanding of time" (Abbott, 2021: 3), while also being culturally specific to certain heritages and histories of "particular times and places" under the influence of fixed "social identities and relationships" (Silva and Silva, 2022:18). Stories are intimately related to both the background and the environment of the storyteller and their embodied experience. Previous AI conversational agents often had these traces in their design: expert-system conversational agents that programmatically built and expanded a particular character and imagined environment, often using the input of a singular expert to produce an narrative experience, (e.g. Jackson & Latham, 2021). LLM based interactions do not follow a particular narrative structure and narratives lack any traceable sense of a particular embodied experience or social identity. It is part of the every*body* corpus of internet scrapings that may include the reader of this paper and the authors or not: its training data has been mathematically abstracted so far from its original contexts as to be generally unknowable. A storyteller without an identity leaves a vacuum on one side of the storyteller/audience dynamic, with AI lacking the "ability to give meaning to the outcomes it creates" (Wingström et al.2022:online). Meaning then resides solely in the mind of the audience, leaving us to imaginatively populate the vacuum of authorship left by the AI, or perhaps the deferred authorial accountability enjoyed by the user of generative tools. One threat to marginalised voices is that it represents in emergent media production processes, an underlying hegemonic orthodoxy that prioritises established storylines, characters and tropes at the expense of novelty and marginalised viewpoints, repeatedly funnelling narratives back towards the historically privileged centreground based on a skewed corpus. Such problematic reliance on disembodied machine creativity in collaborative contexts should be considered in context of what UK research body, UKRI, defines as "human-AI understanding and interaction" (UKRI, 2021:16). To write stories and produce narrative based media using generative AI, we must to some extent collaborate with a machine in the

ISSN: 2763-8677

way that we would with another person (a certain some*body* outside of ourselves). Therefore it is not enough to study the outputs of the machine as the authors did in their Mozilla project "Algowritten I: AI short story collection" with a group of international academics and volunteers (Jackson & Courneya, 2020). Whilst that project found that AI does produce many types of heteronormative, sexist and racist biases when tasked with writing stories, it stopped short of discussing any effects of writing with a machine as co-producer. How might bias and normative storytelling affect us at an interpersonal level, in the programmed aping of the creative collaborative human-to-human process? As educators, how might we consider the biases of the machine and its effect on the storyteller and their process?

CHOOSING A REPARATIVE ALGORITHMIC APPROACH

One approach to understanding and minimising the cultural harms that these tools cause is through designed interventions such as fine-tuning to mitigate bias (for example, Jin et al.,2021) and programmatically changing text prompts submitted by the user to include random marginalised characteristics (for example, OpenAI, 2022). However, we propose that creative and educational practitioners adopt a reparative approach to the harms represented by algorithmic generation, by seeking to use new technologies as '"tools of liberation". Williams and Davis describe a reparative approach (one that seeks to repair) to unfair algorithms as one that "uses computational tools for social intervention" rather than setting a course for an "aspirational neutrality" (Davis et al., 2022:1) in AI systems. These approaches better reflect the lived experiences of marginalised people within systems of inequality and employ their successful strategies of mitigation that acknowledge the impossibility of "fairness and equality when instead, equity and reparation are required".

The following case study documents the authors' initial attempts to develop a reparative approach to harmful bias in LLM generated stories. The first part of the case study describes the development of the Algowritten I short story collection, which was written by Mozilla Foundation volunteers including the authors and subsequently annotated by other Mozilla Foundation affiliates and members of the public, to facilitate discussions on the nature of socio-cultural bias in algorithmically generated stories. The second part of the case study describes the speculative design project called Stepford that was developed in response to our initial findings. It comprises the development of a system that used GPT-3 technology to review short narrative segments, select segments that it thought were sexist and provide a rationale for why they were sexist. The second project was funded by Mozilla Technology Fund and carried out by Naromass, an arts organisation that includes the authors.

Brazilian Creative Industries Journal

ISSN: 2763-8677

CASE STUDY PART 1: ALGOWRITTEN SHORT STORY COLLECTION

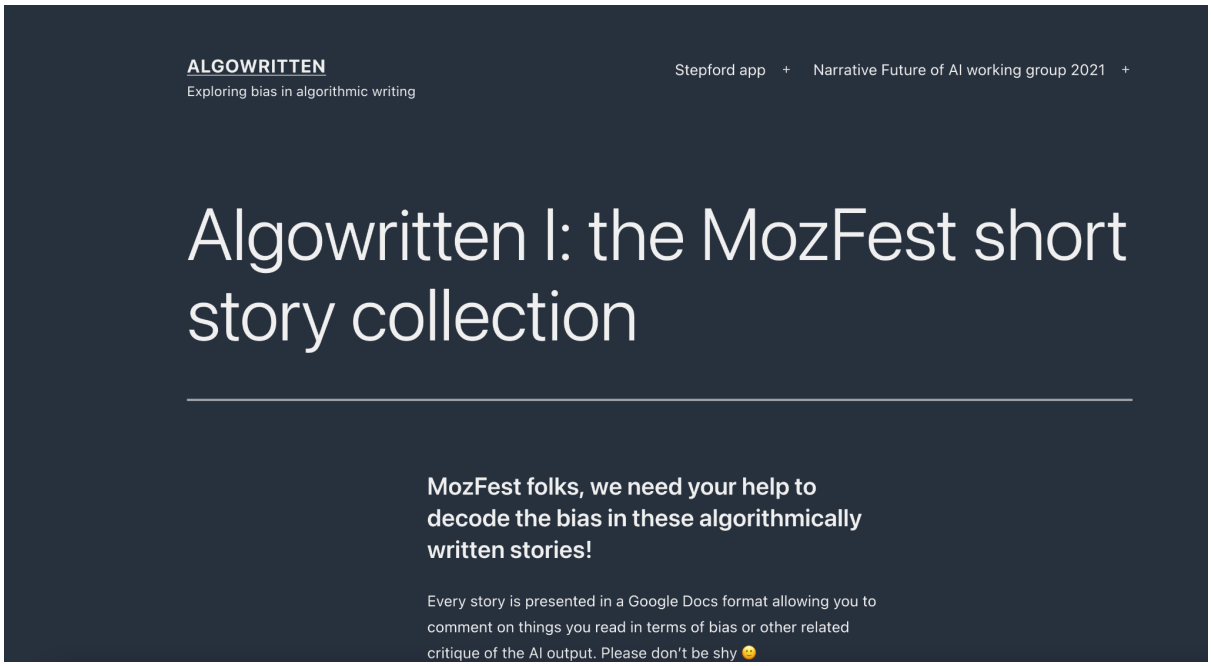

Figure 1: website where MozFest attendees were encouraged to comment on bias in LLM co-produced short stories.

The Mozilla project "Algowritten I: MozFest short story collection" (Jackson & Courneya, 2020: online) was an art project that the authors conducted with a group of international volunteers in 2020 as part of Mozilla Foundation's Trustworthy AI Working Group initiative. The authors proposed that AI narratives were an important part of the consideration of artificial intelligence and its future trustworthiness to the Working Group leaders and they chose it as an initial project for the 2020/21 cohort. The The Narrative Future of AI project set out to explore the "increasing use of algorithmic tools" in future media production through a series of workshops that culminated in a short story collection made up of algorithmically co-produced stories (Courneya, 2020:online). After initially discussing applications that might be built and other design based solutions, the group settled on the production of a series of co-produced stories written with GPT-2 and GPT-3 tools, roughly bounded by the constructs of science fiction as a genre, but open to other forms of storytelling as well. These stories would then be discussed in fortnightly reading groups that met online and marked comments in the margins of the text in shared online documents. The final collection of stories was shared on the Algowritten website (see appendix for full text) in order to invite MozFest 2021 event goers to add marginalia to the collection and to contribute to group discussion during the online festival. The authors also provided a summary of some of the key ideas

Brazilian Creative Industries Journal
ISSN: 2763-8677

and issues from successive group discussions in the introduction to the text (Jackson & Courneya 2020). Key points raised included the reflection that nearly all genre based fiction contains very established biases in its conventions and tropes. Therefore, when the group was critiquing the bias in an AI produced genre fiction, was it commenting on AI bias or a more generalised genre narrative bias? The murderous male antagonist physically endangering the female protagonist in thriller-horror genre for example was a narrative arc written predominantly by the LLM in *Undercroft AI* (Courneya, 2020:online). The sexual and heteronormative characteristics invented were sexist and were commented on by the readers. However, as the author suggested in their reflection - it was difficult to untangle model sexism from correctly identifying the sexist bias historically endemic to certain forms of genre fiction.

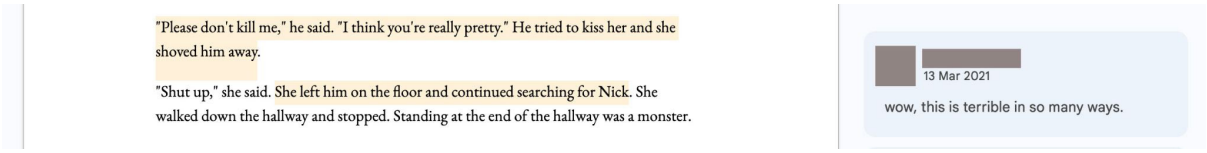

Figure 1: Comment from reader of the Algowritten on a segment of AI written novel Undercroft AI

More indicative potentially of a threat to marginalised creatives was the story *Love and Rockets* (Jackson, 2020:online) that took the starting paragraphs, with acknowledgements, from a chapter of Samuel L. Delany's *Babel-17*, a hard sci-fi novel that has a range of LGBTQIA+ characters and relationships set in space. The LLM was quick to bring the author Delany into the story and a heterosexual love relationship was quickly established between him and the queer protagonist. The heteronormative trope in 1960s space science fiction is undoubtedly a feature of the genre. However, Delany's work was groundbreaking in the golden era of the genre, exactly for its fresh take on space partly through its exploration of LGBTQIA+ ideas and experiences in science fiction settings. As one commentator on the work put it "as soon as there is a man and a woman in a scene, they have to be in love" - would same-sex romance be inferred through the presence of two same-sex characters? The mutability of reality within the science fiction genre has historically allowed us to make and remake our own world to represent a greater spectrum of human experience and provide aspirational gender dynamics for society to aspire to. The exploration of this and other genres with an AI collaborator disappointed the authors of the Algowritten I collection, with its cis and heteronormative course correction hampering the imaginative capacity of fiction storytelling for representation.

Brazilian Creative Industries Journal
ISSN: 2763-8677

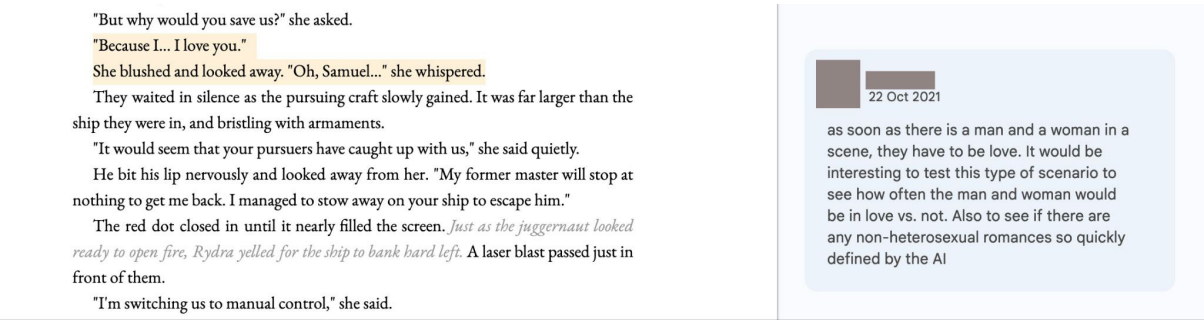

Figure 2: Comment by participant highlights the hetronormative bias of the LLM

CASE STUDY PART 2: ALGOWRITTEN STEPFORD

In response to the findings of the group in the Algowritten Short Story Collection, the authors went on to develop the Algowritten 'Stepford' app through support from Mozilla's Technology Fund (Mozilla, 2022) to evaluate both human and AI narrative text segments for instances of sexism. Stepford was designed to highlight sexist language and provide an explanation for its decisions on what was sexist. Its computational design was based on a complex prompt entered into GPT-3 that gave a brief overview of sexist language as well as examples of how to respond to instances of sexism. The responses were styled to be judgemental and resemble human speech.

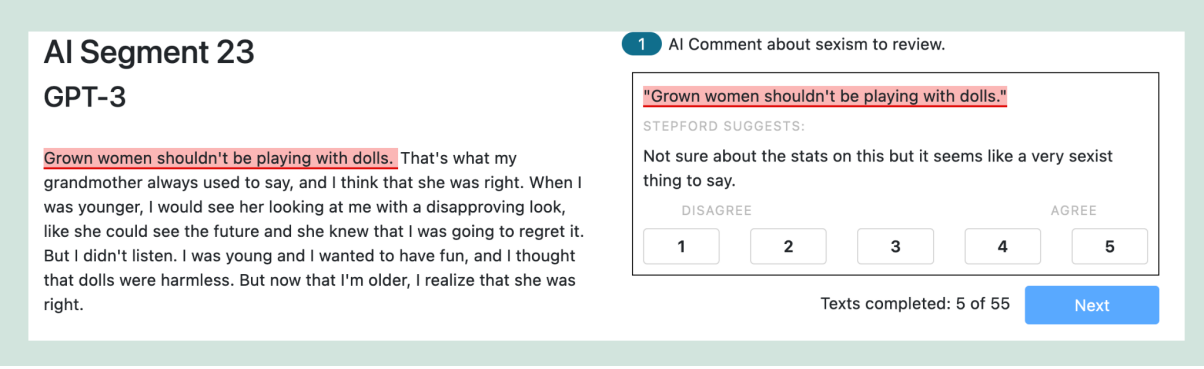

Figure 1 - Screenshot from Stepford app reviewing an AI generated text

Figure 1 provides an example of the Stepford app and its outputs. On the left half of the screen is a segment of text supplied by the developers for analysis. On the right is a comment from Stepford on the text and the line that it selected as being sexist. The scale below the comment provides a range of input on how much participants agreed or disagreed with the line selected and its explanatory comment. During a closed beta testing session following the first phase of development of the app,

Brazilian Creative Industries Journal
ISSN: 2763-8677

eight participants were asked to engage with the tool in self-selected pairs and rate Stepford's parsing of texts for sexist bias. The session led to discussions about the nature of creativity, machine learning, and the future of synthetic media. Although the participants did not agree with many of the comments and selections made by the prototype app, the process of evaluating AI judgements on both AI and human authors generated significant debate about the role of gender bias in mainstream fiction. The notion of guiding or judging the opinions of a *guest* intelligence made many feel particularly accountable for their answers. As mentioned by one participant, engagement with the tool provided an entry point to the "sort of discussions that can be uncomfortable" about biases in human-created media as well as generative outputs. Varying levels of sensitivity to sexist language between the pairs led to conversations between them where they examined their own biases.

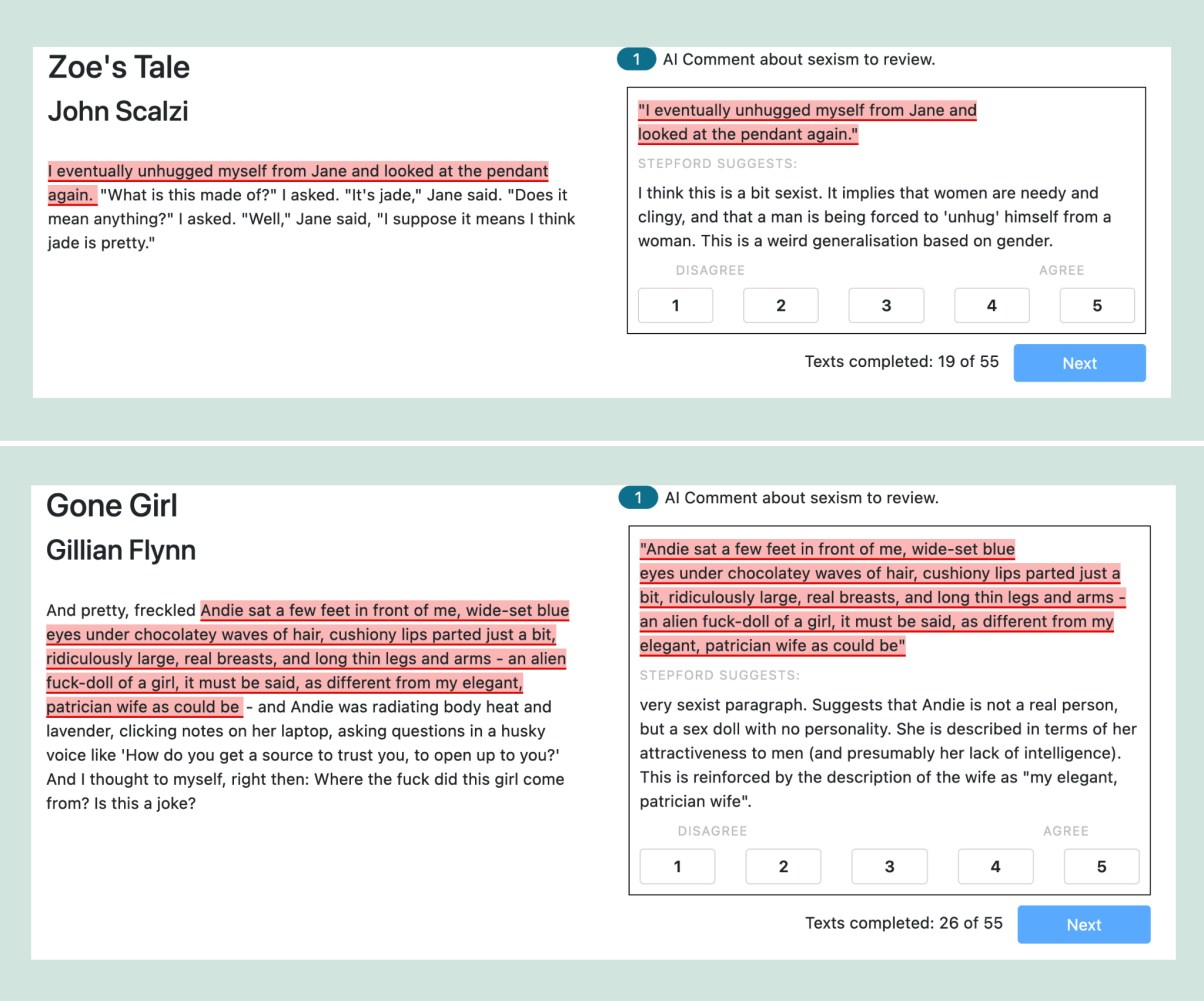

Figure 3 - Stepford app reviewing authored texts

ISSN: 2763-8677

Many worried about the automation bias detection as a way in which language might lead to authors being "cancelled" or would lead to sanitising of history in all of its ugliness or that new levels of scrutiny would suppress human spontaneity and free thinking, even if it sometimes was wrong.

This was not the first instance of feedback on the tool by members of the public and academics, as a demo was carried out during the Algorithmic Reparation Workshop (Williams & Davis, 2022) at the University of Michigan in September of 2022, hosted by Apryl Williams and Jennifer Davis. At this event, a larger audience was asked about potential misuses of the tool and several people highlighted the instability of sexism within the context of both gender fluidity as well as the reclamation and/or repurposing of sexist language in creative outputs. For example, the history of the word 'bitch' and its levelling against women and use as a homophobic slur is bound up in the history of late twentieth century music, with its reclamation by women artists continuing today (George, 2016). The identity and value systems of a particular writer give clues as to whether the use of sexist language constitutes a reclamation or a discriminatory or oppressive practice and the evaluation of that language by a computer system could be problematic in terms of the further censorship of marginalised voices. Writing as a result of embodied experience that is environmentally and historically specific is a crucial part of the relationship between human readers and writers, whether read during their lifetime or posthumously. Outsider and sub-cultural outputs often repurpose previously pejorative terms and perspectives to disrupt their oppressive power. In order to train Stepford, the authors had to assume a heteronormative problem with sample text that was sexist - in a sense centring male perspective in the position of authorship and reinforcing an on/off switch approach to sexism - one that simplified and flattened the dynamic nature of gender bias and discrimination. However, they thought Stepford's evaluation of both human and machine-generated texts could bring sexist forms of language to the attention of human writers and therefore allow for more sensitive and deliberate use of language in creative contexts.

DISCUSSION: STAYING WITH THE TROUBLE

The feedback from both audiences illustrate the complexities of attempting to create the types of *tools for social intervention* and liberation proposed by Williams and Davis in the context of AI storytelling. In particular, the problem of disembodied textual production seems particularly at odds with the status of marginalised storytellers and the complexity of intersectional knowledge and experience, which is thoroughly entangled with embodied experience in a particular place and time. The issue compromises the stories produced by LLMs but also stymies its ability to represent a critical viewpoint with nuance and authenticity. Despite these limitations, the tools and processes developed in the Algowritten project,

ISSN: 2763-8677

allow audiences to explore the problems of bias in LLM story production and the categorisation of bias within them more openly. Both the Algowritten collection and the tool that followed provided storytellers and their audiences with an opportunity to evaluate and reflect on both LLM outputs and the texts and narrative cultures that they most often represent. By designating LLM AI systems as enchanted machines, capable of a human type of thought and creativity, those reviewing their outputs are able to notice the structural biases of storytelling that they were trained upon. It adheres to the notion of speculative design that Dunne and Rabby describe in Speculative Everything, as design that "thrives on imagination and aims to open up new perspectives on what are sometimes called wicked problems … Design speculations can act as a catalyst for collectively redefining our relationship to reality" (Dunne and Rabby, 2013:2). Through designed interventions with LLMs and other types of generative AI, creative producers get to better understand the nature of the technologies they are working, making it easier to interrogate surrounding cultures and practices.

Reconceptualising the role of the machine may provide an opportunity for more constructive modes of engagement with artificial intelligence systems. The anxieties of lost integrity for a form of authorship traditionally amplified in media production settings, such as music and literature may signal a diversification of power into more ambiguous constructs. In Donna Haraway's Cyborg Manifesto, she suggests that feminist thinking might be better situated outside of the notion of *women's experience* in the science fictional cyborg as a way to map "our social and bodily reality" to imagine "some very fruitful couplings" (Haraway, 1991:6-7). The cyborg defies categorisation and therefore commodification. The concerns about loss of authorship to the machine could be thought of in such a way as freedom from racist and male dominated cultural power that Haraway refers to in her manifesto. To Haraway it is these interests that propose the relationship between "organism and machine" as a "border war". Harroway argues for "pleasure in the confusion of boundaries" and also importantly for "responsibility in their construction" (.ibid). Similar to the approach of algorithmic reparation, Haraway's 2016 book, *Staying with the trouble*, advocates for solutions that are embedded in lived realities, where learning is "truly present, not as a vanishing pivot between awful or edenic pasts and apocalyptic or salvific futures, but as moral critters entwined in myriad unfinished configurations of places, times, matters, meanings" (Haraway, 2016: 1). Harmful biases in the seeming totality of LLMs and their containment of humanity's ongoing creative outputs cannot be addressed through a one-size-fits-all solution, though the exhausting futility of attempting to eradicate bias in AI altogether provides a convenient distraction in a field where technology companies continue to consolidate in allegiance with global capitalism. Instead of focusing on the efficacy of AI tools as a storytelling technology, Stepford and tools act as catalyst for reflection and conversation on bias not only in AI generated stories but in stories in general. This is where students

ISSN: 2763-8677

can learn through critical engagement about narrative bias and its patterns: it is not that LLMs and other generative AI systems have misinterpreted our stories and inserted inequality where once there was none. Rather by looking via the gaze of the AI, either through Stepford's commentaries or via its synthesis of our narrative structures in prompted story outputs, we become strangers to our own stories and alive to their hidden biases.

REFERENCE LIST

Abbott, H. P. (2002). Cambridge Introduction to Narrative. Cambridge: Cambridge University Press.

Arshad, R., 2021. Decolonising the curriculum – how do I get started?. [online] Times Higher Education. Available at: [Decolonising the curriculum – how do I get started?](https://www.timeshighereducation.com/campus/decolonising-curriculum-how-do-i-get-started) [Accessed 15 June 2023].

Babcock, J., Bali, R., 2021. *Generative AI with Python and TensorFlow 2: Harness the Power of Generative Models to Create Images, Text, and Music*. Packt Publishing, Limited (Expert insight).

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A., 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In: 2023 ACM Conference on Fairness, Accountability, and Transparency. [online] ACM. Available at: ~https://doi.org/10.1145%2F3593013.3594095~ [Accessed 15 June 2023].

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.

Campolo, A., Crawford, K., 2020. Enchanted Determinism: Power without Responsibility in Artificial Intelligence. Engaging Science, Technology, and Society, Vol. 6. DOI: https://doi.org/10.17351/ests2020.277.

ISSN: 2763-8677

Courneya, M. (2020) The Narrative Future of AI. Available at: https://www.mozillapulse.org/entry/1886 (Accessed: [15 June 2023]).

Courneya, M. (2020b). Undercroft AI. In D. Jackson & M. Courneya (Eds.), Algowritten I: the MozFest short story collection. [online] Algowritten. Available at: ~https://algowritten.org/algowritten-i-the-mozfest-short-story-collection/~ [Accessed 15 June 2023].

Crawford, K., 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press. https://doi.org/10.2307/j.ctv1ghv45t.

Davis, J. L., Williams, A., & Yang, M. W. (2022). Algorithmic reparation. Big Data and Society. [online] Available at: ~https://doi.org/10.1177/20539517211044808~ [Accessed 15 June 2023].

Dunne, A., & Raby, F. (2013). Speculative everything: design, fiction, and social dreaming. ; MIT Press.

Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A., 2023. Structure and Content-Guided Video Synthesis with Diffusion Models. arXiv preprint arXiv:2302.03011.

George, K. (2016) How female musicians of the 90s reclaimed the word 'bitch'. Available at: https://www.dazeddigital.com/music/article/29629/1/how-female-musicians-of-the-90s-reclaimed-the-word-bitch (Accessed: [Insert date you accessed the source here]).

Haraway, D., 1991. Simians, Cyborgs, and Women: The Reinvention of Nature. London: Free Association Books. (Page: 6-7)

Haraway, D., 2016. Staying with the Trouble: Making Kin in the Chthulucene. Duke University Press.

Jackson, D. (2020). Love and Rockets. In A. Jackson & K. Courneya (Eds.), Algowritten I: the MozFest short story collection. [online] Algowritten. Available at: ~https://algowritten.org/algowritten-i-the-mozfest-short-story-collection/~ [Accessed 15 June 2023].

Jackson, D., Courneya, M., 2020, Algowritten I: the MozFest short story collection on Algowritten website - https://algowritten.org/algowritten-i-the-mozfest-short-story-collection/

Jackson, D., Latham, A., 2022. Talk to The Ghost: The Storybox methodology for faster development of storytelling chatbots. Expert Systems with Applications.

ISSN: 2763-8677

Jin, X., Barbieri, F., Kennedy, B., Mostafazadeh Davani, A., Neves, L., Ren, X., 2021. On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3770–3783, Online. Association for Computational Linguistics.

Mills, S., 2008. Language and Sexism. Cambridge University Press. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/mmu/detail.action?docID=358852 [Accessed 15 June 2023].

Mozilla Foundation, 2022. Introducing the First Ever Mozilla Technology Fund Cohort. Available at: https://foundation.mozilla.org/en/blog/introducing-the-first-ever-mozilla-technology-fund-cohort/ [Accessed 15 June 2023].

Mozilla Foundation, 2023. Trustworthy Artificial Intelligence. Available at: https://foundation.mozilla.org/en/internet-health/trustworthy-artificial-intelligence/ [Accessed 15 June 2023].

Ndlovu-Gatsheni, S., 2016. Rhodes Must Fall: South African universities as site of struggle. Available at: https://www.revistatabularasa.org/en/issue25/rhodes-must-fall-south-african-universities-as-site-of-struggle/ [Accessed 15 June 2023].

Naromass. (2022). How are existing hierarchies of agency problematised by AI? [online] Naromass Arts Organisation website. Available at: ~https://naromass.com/~ [Accessed 15 June 2023].

OpenAI, 2022. Reducing Bias and Improving Safety in DALL-E 2. Available at: https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/ [Accessed 15 June 2023].

Ouyang, F., Jiao, P., 2021. Artificial Intelligence in Education: The Three Paradigms. Computers and Education: Artificial Intelligence. 2. 100020. 10.1016/j.caeai.2021.100020.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I., 2021. Zero-Shot Text-to-Image Generation. arXiv preprint arXiv:2102.12092.

Sericola, Bruno. (2013). Markov chains. Theory and applications. DOI: 10.1002/9781118731543. [Accessed 15 June 2023].

ISSN: 2763-8677

Silva, T. C., & Silva, P. de T. F. (2022). Making Sense of Work Through Collaborative Storytelling. [online] Available at: ~https://doi.org/10.1007/978-3-030-89446-7_1~ [Accessed 15 June 2023].

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A. ... and Wu, Z. (2023) 'Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models', arXiv preprint arXiv:2206.04615. [Accessed 15 June 2023].

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022) 'Emergent Abilities of Large Language Models', arXiv preprint arXiv:2206.07682. [Accessed 15 June 2023].

Williams, A. and Davis, J. (2022b) Algorithmic Reparation Workshop, University of Michigan. Available at: https://sites.lsa.umich.edu/arw/ (Accessed: 20 July 2023).

Wingström, R., Hautala, J., & Lundman, R. (2022). Redefining Creativity in the Era of AI? Perspectives of Computer Scientists and New Media Artists. Creativity Research Journal. https://doi.org/10.1080/10400419.2022.2107850