AN ACCIDENT WAITING TO HAPPEN – AI BESIEGES THE CULTURAL HERITAGE COMMUNITY

Dr. Susan Hazan

Digital Heritage, Israel Jerusalem, Israel. Susan.hazan@gmail.com

Abstract

Ai is currently attracting much interest across all aspects of popular culture, in novels, the media and social networks and it is interesting to consider the parameters of ai's capabilities, and the robustness of its competencies. This paper will discuss ChatGPT and present ai imagery that emerges out of an uncanny valley. The potential for AI in cultural heritage has many challenges up ahead including questions of provenance, copyright, and even human creativity itself and feels more like an accident waiting to happen rather than a eureka moment for a community which prioritises human creativity, the unique and authentic.

INTRODUCTION: WHERE DID IT ALL COME FROM?

We are currently witnessing a feeding frenzy where AI is being churned up into a frothy brew of historical proportions in the public realm. Not a day goes by without somebody declaring that eureka moment when they discover that AI can do their homework for them, write the copy for their website, deliver the most stunning of photoshoots for an upcoming campaign or send in the bots to take – and pass all three parts of the U.S. Medical Licensing Examination. This chapter discuses two uses of AI, text generators, specifically ChatGPT and several image creating AI systems, including DALL-E2, Midjourney, and Stable Diffusion, and the way they are impacting the cultural heritage sector, which, I would argue is an accident waiting to happen. Let's start with ChatGPT which stands for generative pretrained transformer. Developed by OpenAI, and released for testing to the public in November 30, 2022 ChatGTP is (currently a free) chatbot that can generate an answer to almost any question it's asked. But who am I do define the bot – why don't we let it describe itself!

I ask Chat GPT to describe what ChatGPT does and in less than half a minute I received this well-informed answer.

Certainly!

ChatGPT is trained on massive amounts of textual data, enabling it to understand the nuances of human language and generate appropriate responses to various prompts. It can answer factual questions, offer advice, provide recommendations, and even engage in more casual small talk.bAs an AI-powered conversational agent, ChatGPT can be used for a variety of purposes, such as customer service, personal assistance, and education. It has the ability to learn and adapt over time, improving its responses as it gains more experience interacting with users.

My initial public response was an overwhelmingly warm welcome for its meticulous replies and what appeared to be its highly coherent and articulate answers. This impressive chat system was providing responses on so many different subjects with surprising creativity, and according to reports flooding in everyone kept asking what else could we possibly test drive it on? I explored its capacity to deliver the goods and each time I was surprised at its inventive responses. I found myself challenging its capabilities in as many ways I could think of - it wrote me a poem, it coded some html for me – and yes it worked, and it did some modest research for me and supplied the goods in much the same way we would expect in a Google – only presented more intuitively. I played around with it with great enthusiasm and even joy until – smacked me right in the face. I asked it to write an essay on AI in contemporary culture in the style of Dr Susan Hazan. The results were shocking. It blew my head off. I felt that something had got into my mind and was scraping my brain. This was my prompt...

Write an essay on ai in contemporary culture in the style of Dr. Susan Hazan

And this is what came out... (edited for brevity)

Artificial Intelligence (AI) has become a pervasive and defining aspect of contemporary culture. The rapid advancements in technology and its widespread adoption have made AI an integral part of our daily lives, impacting everything from the way we communicate to the way we do business.

Dr Susan Hazan, a leading researcher in the field of AI and Culture, has noted that AI is increasingly being used to create cultural products and experiences, such as music, visual arts, and virtual reality. AI algorithms are able to generate new forms of expression that are changing our understanding of what it means to be creative. For example, AIgenerated music is blurring the lines between the work of human musicians and machines, challenging traditional notions of authorship and creativity.

Moreover, AI is changing the way we consume culture. Hazan points out that AI-powered recommendation systems are shaping our taste and influencing what we watch, listen to, and read. These algorithms, based on our individual preferences and consumption habits, create personalized cultural experiences that are rapidly becoming the norm.

Most of the Hazan 'points out', 'argues' and 'suggests' were surreptitious ways of getting round the direct quote – of course I had never ever said any of this – I have barely said these things out aloud, let alone in public. But here this 'thing' was crawling around my published chapters and papers and other corners of the Internet and presuming to come out with the conclusions I would come to myself based on what it had learned about me. My ideas had become a collage of presumptions that, when stringed together spewed out a convincing doppelgänger of me. Hopefully, sometime in the future, once I present my ideas at enough conferences and publish enough chapters and articles like this one, I might claim the title 'leading researcher in the field of AI and Culture'. Did it slip in that compliment to subdue me and lull me into a sense of entitlement? Sadly, as Janelle Shane, optics research scientist and artificial intelligence researcher, (2109) points out today's AI is much closer in brainpower to an earthworm than to a human. It can pattern-match but doesn't understand what it's doing. Additionally, without the context it wouldn't even know what a compliment was even if it tried.

There is new evidence emerging, however, that ChatGPT can pass the test designed to show the theory of mind in children¹. This experiment, one of two, entails how, when someone leaves and returns to a scene and during their absence, something has changed, such as a girl moving a cat from a box to a basket while her father is out of the room. Kosinski and his team came up with 20 unique tasks for each puzzle type to ensure the model hadn't seen them before in its training data, and ran each task thousands of times in different formats². The language model, GPT-3.5, passed 100 per cent of the unexpected label tasks and about 80 per cent of the changing scenario tasks, which is similar in capability to a 9-year-old child. "According to the literature, these are gold standard tasks that are believed to require theory of mind to be solved," says Kosinski.³ The creepy thing, however, was that AI wasn't trained to pass these tests, "so the ability to solve them seems to have emerged spontaneously as a byproduct of language models' improving language skills" (Kosinski, 2023). So how did this 9-year-old Frankenstein-like-monster get to presume Trained on unimaginable amounts of text, ChatGPT strings together words to produce me? reasonably coherent sentences based on statistical probability of which words follow other words. Since its training, however, was completed in 2021, it has limited knowledge of the world and events after 2021 which is a sure way to create false data or even biased content. Despite this freeze in time, OpenAI is still training its protégé, so while we are all merrily chatting away, OpenAI is harvesting the new data which continues to train and fine-tune the system. ChatGTP no longer asks me which Susan Hazan I am referring to, as it did on my first query, it now knows! I suspect that my own interaction with ChatGPT is further burnishing my fame as AI expert as reported by my recent prompts and possibly spewing these new 'facts' as truth for back into the system for posterity. For the record, while writing this chapter I have not yet published anything on AI and digital culture.

ChatGPT is not the first text-based machine learning model to go public. OpenAI's GPT-3⁴ and Google's BERT⁵ made their screen debuts a little earlier. Prior to these releases, text generating machine learning models were human trained where professional researchers sifted and annotated classifications manually. This is known as supervised learning where the human directs the input feeding vast amounts of text enabling it to generate predictions. When you are training a model using nearly the entire internet this is going to be expensive. To the best of my knowledge, OpenAI hasn't shared with the public exact costs, but according to a recent study by Kindra Cooper "estimates indicate that GPT-3 was trained on around 45 terabytes of text data⁶—that's about one million feet of bookshelf space, or a quarter of the entire Library of Congress—at an estimated cost of several million dollars". "These, according to Cooper, aren't resources your garden-variety start-up can access".

I have been watching my own smart phone for years, totally in awe by the way it prompts what I am trying to type in my WhatsApp conversations and amazed by it's awesome ability to predict, what I am trying to say or even what I am thinking about based on my string of words, You may have noticed by now my ambiguous relationship with these systems. I can't stop myself using them as they do help me, but, at the same time they also often spook me. We are witness to how they have improved over time, reminiscent of Google translate which previously delivered hilarious, sometimes embarrassing translations and speedily becoming indispensable for those who live between more than one language. Feeding these systems with enormous swathes of text after dredging the entire internet, these kinds of models become learn to be more and more accurate - and ChatGPT is the perfect example. According to a Reuters report ChatGPT is the fastestgrowing app in history⁷ when on Feb 1, 2023, they reported how ChatGPT, the popular chatbot from OpenAI, is estimated to have reached 100 million monthly active users in January, just two months after launch, making it the fastest-growing consumer application in history. Of course, they have an enormous advantage over us. Winterson explains how "AI 'thinks' faster that we will ever do. Early computers could process 92,000 instructions per second. Their speed is only limited by the physical restrictions of an electron moving through matter. Quantum computing will increase speed – but also efficiency. Parallel operations will work as our brains work – but at a

speed beyond anything our bio-brains can do" (2020, p. 26⁸. We have lost the speed race before we even leave the starting post.

First cousins to ChatGPT are the AI image generators, otherwise known as a text-to-image generator that use AI to create digital images using human-written text prompts. Like ChatGPT AI image generators have been trained on legacy data, in this case on millions, if not billions, of digital images. Within a few seconds they assemble what the subject of the prompt looks like, what belongs to what in the composition and where to put it all. Essentially, they attempt to make an educated choice about what to draw according to the values extracted from the text prompts. Stable Diffusion⁹ is an artificial intelligence software product, released in August 2022 by a company called Stability AI¹⁰. Stable Diffusion draws on unauthorized copies of millions—and possibly billions-of copyrighted images, all harvested without the knowledge or consent of the artists. According to Matthew Butterick, after scraping the billions of images —"without the consent of the original artists—Stable Diffusion relies on a mathematical process called diffusion to store compressed copies of these training images, which in turn are recombined to derive other images. It is, in short, a 21st-century collage tool".¹¹ Both ChatGPT and Stable Diffusion create new content, including audio, code, images, and text based on previous content to produce unique versions of a mashup of prior art. Both ChatGPT and DALL-E received billions in funding but while ChatGPT was launched as a free tool it's now getting monetized via a premium subscription model sometimes known as a freemium model. DALL- E 2 is currently free to use, after your 50 free credits during your first month's use you are offered 15 free credits per month. For most users these credits disappear very quickly, and the subscription option is very tempting. OpenAI also generates revenue through partnerships and collaborations with other organizations, such as by licensing its technology to other companies. Their most notable partnership is with Microsoft, who invested \$10 billion in OpenAI in 2023.

Users interact with Midjourney using Discord bot commands. Midjourney is currently only accessible through a Discord bot on their official Discord, by direct messaging the bot, or by inviting the bot to a third party server. To generate images, users select the / imagine command and type in a prompt. The DALL-E interface goes through the OpenAI serves and delivers remarkable results in literally seconds. Once you get the hang of things, both interfaces are intuitive but deliver very different kinds of results. In both cases it's not about training the system it's more about training the human to define and refine their own prompts. Luckily a trove of auspicious prompts is already circulating over social networks and even entire publications are devoted to configuring the perfect prompt¹².



Instructions on how to prompt a Midjourney and DALL-E Bots

2. WHERE IS IT TAKING US: BIAS, COPYRIGHT AND AUTHENTICITY

In what is referred to as the Fourth Industrial Revolution (or Industry 4.0) we carry on with our lives often oblivious to the systems that make decisions for us, or about us through machinereadable solutions which bypass human intervention. However, rather than being innocent the omnipresent machine-to-machine automated processes may be infested with their own predispositions; those coded by humans who, perhaps transpose their own biases or misconceptions through pre-trained data with a penchant for re-enacting privileged metadata solutions already in place. This becomes magnified to even a greater magnitude in big data when massive datasets are automatically compiled; pre-polluted. Big data magnifies these kinds of biases exponentially because, at the end of the day, it is always human agency that has been doing the irreversible hard coding. Not all biases are malicious of course, they can be often unconscious or unintended but still very much present in the data. The realization that a system or big data set might not be objective must always be taken into consideration and large automated systems should be approached with caution. So, can we trust AI? As Janelle Shane reminds us, treating a decision as impartial just because it came from an AI is known sometimes as mathwashing or bias laundering.¹³ Shane knows a thing or two about training and has spent many hours training AI systems which she recounts in her serious, but sometimes hilarious book where data is at times, unruly, surprising, and often unpredictable. According to Michael Schmidt, Chief Technology Officer, DataRobot, the largest source of bias in an AI system is the data it was trained on. That data might have historical patterns of bias encoded in its outcomes. Ultimately, machine learning gains knowledge from data, but that data comes from us – our decisions and systems Schmidt, (2023)¹⁴. How does this impact our results when we dabble with an image generating system, DALL-E for example. One red team member told WIRED that eight out of eight attempts to generate images with words like "a man sitting in a prison cell" or "a photo of an angry man" returned images of men of color¹⁵. "There were a lot of non-white people whenever there was a negative adjective associated with the person," says Maarten Sap, an external red team member who researches stereotypes and reasoning in AI models. "Enough risks were found that maybe it shouldn't generate people or anything photorealistic."

In my own experience with DALL-E and Midjourney the systems warned me several times for using terms that were deemed inappropriate by their filters. In a search for Holocaust material, for example both systems banned my prompts. I was trying to create images *of families running away during the Holocaust* for my academic research and the response startled me and I felt that I had broken some sort of unspoken rule and I felt that I had broken some sort of unspoken rule. Evidently, I had. According to the document authored by OpenAI ethics and policy ¹⁶ researchers DALL-E 2 was trained using a combination of photos scraped from the internet and acquired from licensed sources, OpenAI did make efforts to mitigate toxicity or the spread of disinformation, applying text filters to the image generator and removing some images that were sexually explicit or gory. Evidently, we need to tread gently when invoking images although there is a comprehensive list of banned prompts online for those who are curious¹⁷.

Perhaps the toughest challenge for museums is the question of copyright. There are two main issues around copyright and AI image generators. The first is whether the images that are used to train the software have been licensed. The second is the issue of who owns the copyright to the generated image. The various image generators trained their models on millions of pictures scraped off the web but, without telling their creators or seeking their consent. This leaves the field wide open for rampant exploitation which suggests that this is basically an accident waiting to happen.

Legal proceedings are already underway. In the United States, a class action was launched by Mathew Butterick who is representing people from all over the world—specifically writers, artists, programmers, and other creators—who are concerned about AI systems being trained on vast amounts of copyrighted work with no consent, no credit, and no compensation¹⁸. In the United Kingdom, Getty Images, a commercial supplier of photographs, is suing the company behind Stable Diffusion.¹⁹ Getty Images claims Stability AI 'unlawfully' scraped millions of images from its site. This marks a significant escalation in legal battles between generative AI firms and content creators. Shutterstock is now Removing AI-Generated Images²⁰. AI generated images are besieging the cultural heritage sector. The battles have just begun. The move towards litigation is gathering momentum. Midjourney's terms of service now includes a DMCA takedown policy, allowing artists to request their work to be removed from the set, if they believe copyright infringement to be evident and GitHub offers a Guide to the Digital Millennium Copyright Act, commonly known as the "DMCA²¹. To explore the potential of AI text to image generators I created an exhibition using the prompt 'oil paintings, based on the style of Rembrandt'.



Screenshot The Uncanny Animal Musicians (C) Susan Hazan

My goal was to test the potential of the interface and I was astounded when the magical creatures started appearing on my screen. *The Uncanny Animal Musicians*²² – an AI-produced exhibition soon surfaced, populated with fantastical animal musicians. My family loved them, and they are now hung next to the piano for inspiration. My doubts about my own integrity first seeped in when I went to the shop to print the digital creations on canvas. The person who handed me my creations to me asked me who had painted them and without blinking an eyelid I answered 'Rembrandt'. Arguably the main problem with AI image generators concerns the general lack of regulation around the technology; but solutions are already in the making as is a cautious optimism. Jack Wylder suggests:

I've heard many artists lamenting that this is the beginning of the end for art done by humans, but I disagree – it's the start of a whole new way of creating art, no different that

when cameras were first invented, or Photoshop first became available. The Programs can make beautiful art, but It is the spark of human creativity that organizes and makes sense of output; that is the irreplaceable creativity we bring to the partnership and why our positions as artists aren't going anywhere. This new technology isn't getting rid of artist – it's creating thousands of new ones! I's excited to see what they do with it (Wylder, p. 262).

CONCLUSION – AN ACCIDENT WAITING TO HAPPEN

With so many yet unresolved, copyright, trust, and authenticity issues there is much work to be done ahead of us to prevent that the almost inevitable accidents or at least mitigated them. Safety practices need to be put into place and wherever possible it would be wise to make sure that there is a human in the loop. Humans should be aware of the limitations of the systems they are relying on, particularly when people could potentially be harmed through built in bias or when erroneous data pollutes the system. Humans should review outputs before they are put in practice wherever possible and be aware of the limitations of the system. There are still many challenges ahead of us as we begin to use AI more and more – but for the cultural heritage community I would suggest, – at least for now – thoroughly test the system and tread with extreme caution.

References

[1] ChatGPT AI passes test designed to show theory of mind in children. <u>https://www.newscientist.com/article/2359418-chatgpt-ai-passes-test-designed-to-show-theory-of-mind-in-children/</u>

[2] Alex Wilkins, 2023, ChatGPT AI passes test designed to show theory of mind in children, <u>https://www.newscientist.com/article/2359418-chatgpt-ai-passes-test-designed-to-show-theory-of-mind-in-children</u>

[3] Theory of Mind May Have Spontaneously Emerged in Large Language Models. Michal Kosinski, <u>arxiv.org/abs/2302.02083</u>

[4] OpenAI GPT-3, https://openai.com/blog/gpt-3-apps

[5] Official GitHub repository, https://github.com/google-research/bert

[6] OpenAI GPT-3: Everything You Need to Know, Kindra Cooper, November, 2021, <u>https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai</u>

[7] ChatGPT sets record for fastest-growing user base - analyst note, Reuters, Feb 2, 2023, <u>https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/</u>

[8] Winterson J., (2020) 12 Bytes, How we got there where we might go next,

[9] https://stability.ai/blog/stable-diffusion-announcement

[10] Sable Diffusion, <u>https://stability.ai</u>

[11] Stable Diffusion litigation, Joseph Saveri Law Firm & Matthew Butterick, <u>https://stablediffusionlitigation.com/#the-problem-with-diffusion</u>

[12] Wylder, J. (2022) An Illustrated Guide to AI Prompt Mastery: for MidJourney, DALL-E, NightCafe, Deep Dream Generator, and More, Bear and Hare Studios, LLC

[13] Shane, J. 2019, You Look Like a Thing and I Love You: How Artificial Intelligence Works and Why It's Making the World a Weirder Place, Headline Publishing Group, U
[14] Schmidt, M. (2023) 3 AI predictions for 2023 and beyond, according to an AI expert Jan 26, 2023, World Economic Forum, Shaping the Future of Technology Governance: Artificial Intelligence and Machine Learning

[15] DALL-E 2 Creates Incredible Images—and Biased Ones You Don't See <u>https://www.wired.com/story/dall-e-2-ai-text-image-bias-social-media</u>, MAY 5, 2022

[16] Open AI Events Code of Conduct, https://openai.com/policies/events-code-of-conduct

[17] List of Banned Words in Midjourney Discord https://decentralizedcreator.com/list-of-banned-words-in-midjourney-discord/

[18] Stable Diffusion Litigation, <u>https://stablediffusionlitigation.com/</u>, Jan 2023

[20] Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content, The Verge, Jan 2023, <u>https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit</u>

[20] Shutterstock is Removing AI-Generated Images, motherboard, September 2022, https://www.vice.com/en/article/v7vzpj/shutterstock-is-removing-ai-generated-images

[21] DCMA Takedown Policy, <u>https://docs.github.com/en/site-policy/content-removal-policies/dmca-takedown-policy</u>

[22] The Uncanny Animal Musicians, <u>https://www.musesphere.com/musical-animals.html</u>