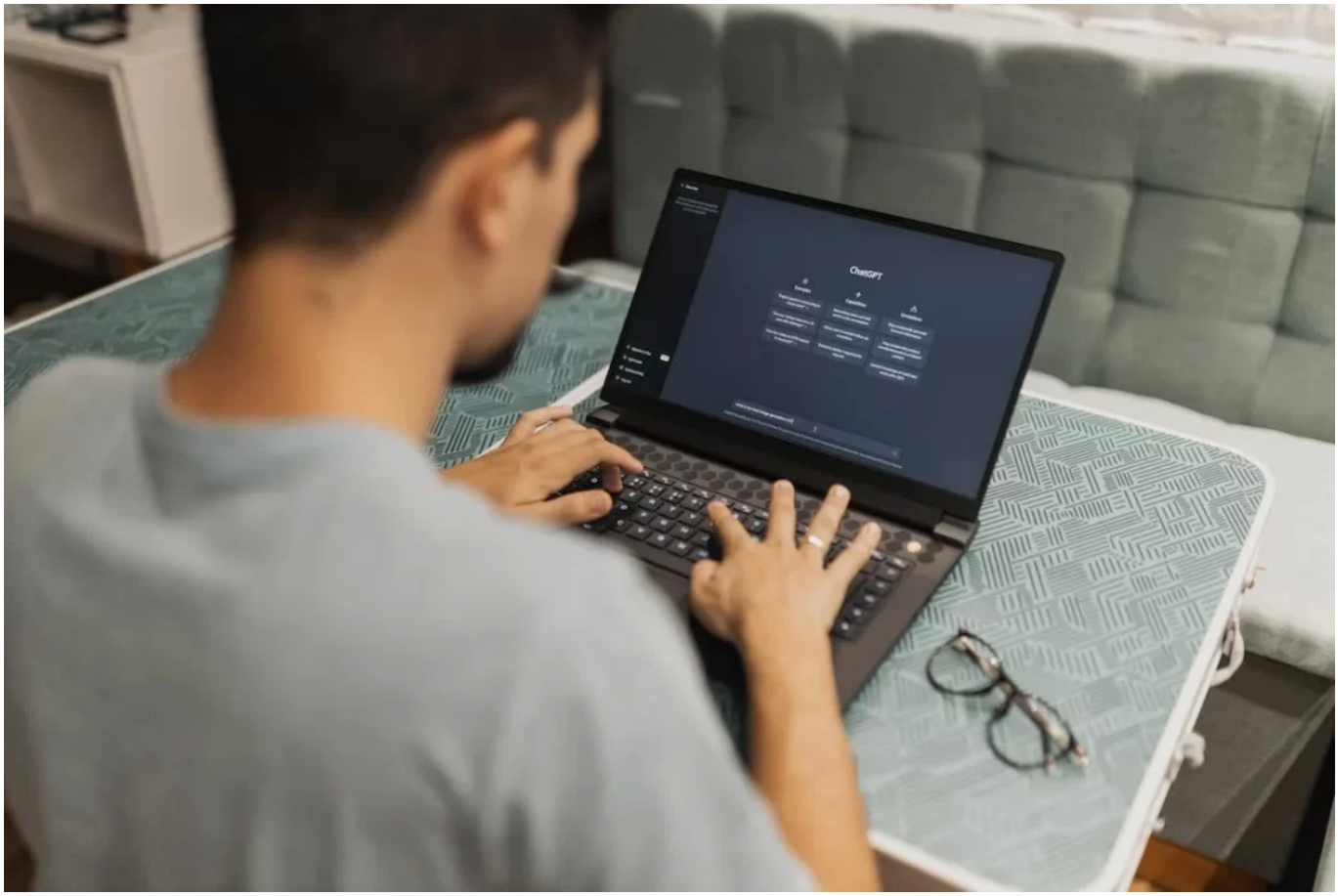
How do we cite ChatGPT?



Chat GPT and other examples of Generative Artificial Intelligence (AI) are capturing our collective imagination. We are excited, startled, and bewildered by the type of responses we receive from these AI applications. With the rapid trialling and adoption of these applications, we are left intrigued and perplexed about the possible capabilities and work applications of AI. We are trying to figure out how do we relate to them. Can we see ChatGPT, for example, as an intentional unit or author? Can we reference it as such? Do we rather reference the team who designed and developed the AI model? Or do we need to rethink what AI is and what an author is?

As AI has evolved over the decades it has challenged or tested how we see ourselves as sentient beings and, relatedly, what we see or define as intelligence, intentionality, the self, authorship, and accountability for example. It also raises questions on *how* and *who* we attribute thinking, feeling, learning, reflecting, judging, and evaluating to – whether it be other humans, animals, or machines. Consider the histories of discrimination, exclusion, and who was accorded 'rights' over different time periods. And consider the biases that are 'baked into' the AI model through the biases in the data used to train the AI and in the design team that developed the AI model.

With ChatGPT's conversational format, where we can interact with the AI model in a simulated conversation, it appears to mimic our day-to-day interactions with other human beings. So, we find people referring to or engaging ChatGPT as a *fellow being*, *interlocutor*,or *author*:

"ChatGPT says…", "ChatGPT's answer to me is…", or "ChatGPT's advice is…".

We attribute – implicitly or explicitly – intelligibility and intentionality to ChatGPT and we assume we are having a meaningful conversation with it. Some are referencing ChatGPT as an authoritative source and author in academic and practitioner articles. But what is 'it'? And does it provide the same response to the same prompt or question at different times? What informs its responses? And can we say that it 'authors' or 'signs off' these responses?

As I discussed in the March 2023 SABPP Fact Sheet, generative AI is a class of machine learning models comprising of artificial neural networks and algorithms (or set of computing instructions or procedures to solve problems and complete tasks). ChatGPT is a Large Language Model (LLM) type of generative AI. The description 'large language model' refer to the large text datasets the model is trained on and the internal configuration (or the parameters of the artificial neural network) of the model that can be changed over time. The training data usually is public text accessed and drawn from the world wide web. The GPT in ChatGPT means the Model is based on a generative pre-trained transformer. The transformer or machine learning model is pretrained on large public data sets. It comprises of *mathematical techniques* to track *relationships* in the *data* and the claim is that it can thereby analyse and respond to context and the variable meanings of words in these contexts. The question though is: are these mathematical techniques *mimicking or simulating* meaning-making, intelligibility, and intentionality that we normally attribute to human speech, writing, and interaction? Even if it is sophisticated mimicking to the extent that we cannot easily tell that it is, it is still mimicry after all. However, perhaps we need to step back and reconsider that our perspectives are limited to what we think is humanly possible – where we use human beings as the benchmark for intelligent and sentient beings – and, therefore, we are not open to new emergent or other forms of sense-making, intelligibility, and interaction?

These are difficult philosophical questions to be wrestled with and there are no clear-cut answers. So, who and how do we cite when use ChatGPT in the meantime? The APA style team have made some suggestions – where APA refers to the popular American Psychological Association's style and referencing guide. McAdoo (2023) describes one aspect of their suggestions below:

"Unfortunately, the results of a ChatGPT "chat" are not retrievable by other readers, and although nonretrievable data or quotations in APA Style papers are usually cited as **personal communications**, with ChatGPT-generated text **there is no person communicating**. Quoting ChatGPT's text from a chat session is therefore more like *sharing an algorithm's output*; thus, **credit the author of the algorithm** with a reference list entry and the corresponding in-text citation" (bold, italics, and underline added) (Source: https://apastyle.apa.org/blog/how-to-cite-chatgpt).

The recommendation is not to cite ChatGPT as an author or 'person communicating'. It is authors of the algorithm of the AI model, the version of the AI model, and the specific description of the AI model (in the case of ChatGPT it will be "large language model") that we cite. Full text outputs from ChatGPT or other AI is also suggested for an appendix section, given that the AI is changing parameters over time and that

it does not provide the same response to the same prompt at different times. McAdoo states, it is "particularly important to document the exact text created because ChatGPT will generate a unique response in each chat session, even if given the same prompt."

As indicated in the beginning, this was written by a human and by convention it is referenced as a work of an author; and by tradition we assume intentionality and intelligibility of the author and the work. Being human though means that the article is infused with the limitations, foibles, and folly of a human being in a particular setting. Both the human and the work need to be situated in a particular context and time.

So, going back to the beginning of the article, where to from here? Who will we 'cite' for the future that awaits us? Adapting the phrasing from popular sci-fi films, we can certainly say that we will boldly and misguidedly go and construct new horizons and worlds as we and machines/AI evolve together. The key question is how we manage the journey into the future that we create. Who will author that future?