Article

new media & society
1–25
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14614448221075296
journals.sagepub.com/home/nms
SAGE

Artificial intelligence and the affective labour of understanding: The intimate moderation of a language model

Carlo Perrotta, Neil Selwyn and Carrie Ewin
Monash University, Australia

Abstract

Interest in artificial intelligence (AI) language models has grown considerably following the release of 'generative pre-trained transformer' (GPT). Framing AI as an extractive technology, this article details how GPT harnesses human labour and sensemaking at two stages: (1) during training when the algorithm 'learns' biased communicative patterns extracted from the Internet and (2) during usage when humans write alongside the AI. This second phase is framed critically as a form of unequal 'affective labour' where the AI imposes narrow and biased conditions for the interaction to unfold, and then exploits the resulting affective turbulence to sustain its simulation of autonomous performance. Empirically, this article draws on an in-depth case study where a human engaged with an AI writing tool, while the researchers recorded the interactions and collected qualitative data about perceptions, frictions and emotions.

Keywords

Affect, artificial intelligence, digital literacy, labour, language models

Introduction

Imagine a scenario where typing a few words into a word processor can generate additional phrases, sentences or even whole paragraphs that follow on from – and substantially augment – what is being written. This is the promise of various text generation tools that

Corresponding author:
Carlo Perrotta, Faculty of Education, Monash University, 29 Ancora Imparo Way, Clayton, VIC 3800, Australia.
Email: carlo.perrotta@monash.edu

have emerged over the past few years, driven by an area of artificial intelligence (AI) known as 'natural language generation' (NLG). It is likely that readers of this article may have already experienced NLG-supported automated writing through popular tools such as 'Grammarly' and 'Google Docs' which are configured to suggest a few likely words on the basis of what has been written previously. The established definition of NLG is as follows: 'the subfield of artificial intelligence and computational linguistics that is concerned with the construction of computer systems than can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information' (Reiter and Dale, 1997: 1; cited in Gatt and Krahmer, 2018).

NLG applications have been examined in various literatures such as computational linguistics and computer vision (Mitchell et al., 2012), media and journalism studies (Broussard et al., 2019), educational research (Jones, 2021) and computational creativity (Roemmele and Gordon, 2015). Positioning NLG as a subfield of AI is appropriate since both communities share a great interest in unsupervised computational approaches like deep learning. Indeed, neural networks trained on large textual corpora underlie several current 'language models' (LMs), which can produce grammatically correct but, often, semantically feeble language. This convergence of interests has of late informed a more generalist label that tries to capture the pivotal role that these technologies are poised to play beyond language generation: foundation models (Bommasani et al., 2021).

These large models have become object of considerable controversy in light of concerns around their environmental costs and their embedded biases (Bender et al., 2021). With regards to the latter, it is important to acknowledge the ample literature on the algorithmic bias as a general problem of computational systems, and of specific NLG applications. Noteworthy contributions which are relevant to the present discussion include work to 'debias' automatically generated text through careful human evaluation (Van Der Lee et al., 2019), and recent efforts to examine biases through a more explicit appreciation of how power relations entrenched into language may harm marginalised groups (Blodgett et al., 2020; Noble, 2018). A useful distinction in this regard is between allocational harms, which arise when an automated system allocates resources (e.g. credit) unfairly, and representational harms which arise when systems misrepresent some groups or fail to acknowledge their existence altogether (Barocas et al., 2017).

Work in 'affective NLG' is also pertinent here (Piwek, 2002; Van Der Sluis et al., 2011). This sub-field emerged approximately two decades ago as an attempt to use automatically generated language to 'deliberately influence emotions or other non-strictly rational aspects of the Hearer' (De Rosis and Grasso, 2000: 204). A preliminary clarification is needed at this point: the terms 'affect' and 'emotion' will feature as synonyms in this article to reflect a semantic overlap found in sociological and psychological research. In this regard, Wetherell (2013) offers an excellent critical synthesis of the so-called affective turn (Massumi, 2002; Sedgwick, 2003) from a perspective of language and discourse theory. While using affect and emotion interchangeably allows us to engage productively with diverse scholarly traditions, we are aware of the risk of blurring definitions and therefore adopt Stark's useful distinction between sub-conscious bodily energy (affects or reactions), consciously registered sensory experiences (feelings) and the linguistic and cognitive interpretations of affects and feelings (emotions) (Stark, 2019). This article is largely concerned with the latter and frames a particular form of NLG (a large language model) as an ‘emotive actant’ (Stark, 2019: 120): an agent that elicits or intensifies emotional expression – not as part of a deliberate design strategy, but as a consequence of its biased and imprecise simulation of communicative competence. We complement this framing using insights from critical sociology around affective labour (Negri and Hardt, 1999), viewed as a form of unrecognised – mostly gendered – toil that enables the mundane yet essential reproduction of social life.

Setting off from this conceptual framework, this article then details a specific experience of engaging with an AI emotive actant. As qualitative social scientists, we examined one current form of NLG which is available for public use – an AI-assisted writing tool powered by the GPT-2 language model – and observed how several people interacted with it. During our observations, we noted that the biases embedded in the model evoked emotional responses which invited a laborious process of sensemaking and moderation. This process appears to rely upon extractive logics which are ‘differentially exploitative’ depending on the user’s position within pre-existing structures of marginalisation. This article is organised into three substantive sections.

First, we clarify the conceptual framework by elaborating on the notion of extractive AI, with a particular focus on labour exploitation. We make additional connections with scholarship on affective sensemaking and recent contributions in communicative AI.

Second, we undertake some contextualised definitional work, mindful that AI is still an area of semantic contestation shaped by economic agendas and techno-utopian imaginaries. Following Crawford (2021: 8), we therefore qualify the particular AI under scrutiny (the popular language model GPT) as a ‘registry of power’ which is neither artificial nor intelligent, but entirely beholden to a narrow political-economic logic. This descriptive work will touch upon some technical aspects of GPT and will provide a more robust definitional framing for the remainder of the article.

Third, we draw on empirical work during which we prompted people with a background in writing to engage with a GPT-2 powered tool to write short samples. In the interest of depth, we examine in detail the relationship between one individual participant and the AI, showing how the responses elicited by the system involved considerable emotional turbulence and ‘repair work’ to moderate the system’s biases.

Extractive AI and affective labour

A profoundly extractive paradigm underpins the development, production and operation of any AI technology. Alongside the dependence of AI upon the extraction of material resources from the earth, are two other forms of extraction that provide key points of concern for this article. First is the reliance of all AI tools on the extraction of data – initially in terms of a system being trained on data scraped from the Internet, and subsequently in terms of data generated from ongoing use of the tool. Second is the reliance of all AI tools on the extraction of labour. Here, AI reflects a broader turn to computational control that pursues surveillance and algorithmic micro-management, while hiding a sprawling global network of underpaid workers who sustain the contemporary ‘automation charade’ (Taylor, 2018) by manually moderating content, or by actively intervening – unseen in the background – in a system’s performance (Gray and Suri, 2019; Roberts, 2014; Tubaro et al., 2020).

In addition to critical scholarship and commentary on extractive AI and its hidden labour, we orient this article towards recent research that examines the sensemaking that occurs when humans encounter algorithms within the mundane interactions on digital platforms (Rader and Gray, 2015). Bucher (2017) has detailed how people develop 'imaginaries' of algorithms' behaviour based on their own lived experiences. For example, when an algorithm fails to meet expectations (perhaps by showing undesirable content or misinterpreting identities and wishes), Bucher notes a dynamic of 'force relations' where people's affective responses are not only the products of computational logics, but also play 'a generative role in moulding the algorithm itself' (Bucher, 2017: 41). This process is twofold. First, people's affective responses are a key input informing the recursive adjustment of machine learning's feedback loops (Stark and Hoey, 2021). Second, emotional responses elicited by an artificial agent are a function of linguistic and interpretative competence, as humans tend to project meaning and intentions onto such agents, thus creating an illusion of communicative agency where there actually is none (Guzman and Lewis, 2020; Nass et al., 1994). In this second connotation, emotion is generative because it contributes to an illusory 'situation' based on the enactment of interactional scripts (Goffman, 1964).

In this article, we are particularly interested in this second form of generative human–machine interaction but, as prefaced in the introduction, we do not frame it as a merely communicative phenomenon. By simulating established communicative patterns – including their emotional components – an AI emotive actant configures a 'banal' form of deception which is low fidelity in nature (McLuhan, 1964/1994) that is, it 'demands more participation from audiences and users in the construction of sense and meaning' (Natale, 2021: 9). As Natale notes, media theory often sees this participation as a form of agency that can be appropriative or even resistive. By engaging with the political economy of communicative AI, we prefer to see this participation through the lens of labour extraction. Still a form of (mediated) agency but one that, depending on one's position in pre-existing structures of disadvantage, can rapidly shift from augmentation to exploitation. In this sense, our work extends Natale's suggestion that while banal deception may 'improve the functionality of interactions with AI, it does not mean that is devoid of problems and risks' (128).

This extension can be qualified as follows: emotive interpretation adds value to the interaction with an artificial agent and is constitutive of its 'smooth' operation in the context of a communicative task. As such, it can be viewed as a form of labour not too dissimilar from the invisible microwork enlisted in other AI production processes, and the continuation of an established trend of where identities, moods, norms, linguistic conventions and notions of what is true and false are the object of constant modulation and capture (Negri and Hardt, 1999; Parisi and Terranova, 2000). The way we deploy terms like capture and extraction is indebted to the notion of 'real subsumption', which occurs when 'society tends toward being completely enveloped by the machine of capitalist valorisation' (Hardt and Negri, 2018: 417), and when subjectivities and affects are absorbed by a totalising extractive operation occurring at the social and personal levels.

All these ideas and debates inform our proceeding investigation of AI-assisted writing. At this point, it is worth stating that our primary interest is not in the nature of text composition, authorship or literary studies. Rather, we are interested in AI-generated writing as a case study of the reconstitutions of AI and human agency that are fast unfolding across various areas of society. As such, we use AI-generated writing to explore a more specific empirical question: how does the affective interaction between humans and an AI language tool generate meaning? Before answering this question, it is important to develop a better understanding of the technical dimensions and the designed properties of the specific form of communicative AI under investigation.

Generative pre-trained transformer

In 2019, the OpenAI research organisation released a 'language model' called GPT-2, able to produce coherent paragraphs of text. The model was the result of feeding large amounts of training data to an unsupervised neural network. Once trained, the model was claimed as able to 'achieve state-of-the-art performance on many language modelling benchmarks, and perform rudimentary reading comprehension, machine translation, question answering and summarization – all without task-specific training' (OpenAI, 2019: n.p.).

A year later, OpenAI released an upgraded version called GPT-3, based on the same model architecture but significantly larger in size. While GPT-2 was composed of 1.5 billion parameters (i.e. the values optimised by the neural network as it learns from the data), GPT-3 was scaled up to include 175 billion parameters. With this increase in size came claims of better performance in several natural language tasks. While the model's enduring limitations in dealing with meaning were explicitly acknowledged by its creators, its achievements in benchmarks reignited a debate about the potentials and pitfalls of AI. Much of this ongoing debate has concentrated on the impossibility to replicate human intelligence and the biases embedded in the models. Both GPT iterations were trained on large textual corpora derived from global Internet activity. The GPT-2 data set (WebText) was sourced from the online community Reddit, while the GPT-3 data set originated from the CommonCrawl corpus, based on nearly a trillion words extracted from the Internet from 2016 to 2019. It is generally accepted that these data sets reflect biases associated with the problematic nature of contemporary Internet discourse (Luccioni and Viviano, 2021). However, the actual nature of these biases remains unclear and future work will require greater public access to OpenAI's technical documentation.

The key distinguishing feature of GPT compared with other language models is the ability to perform effectively in multiple tasks without needing new training (Radford et al., 2019). This semblance of generality gives GPT an aura of multi-purpose, quasi-human intelligence – echoing a general societal discourse where AI is framed in enthusiastic terms of 'general AI' and other forms of (super) human intelligence. In this sense, the current (uncritical) enthusiasm for GPT is driven at least in part by the success of deep learning and neural networks over the past decade, with some leading AI experts framing them as 'neural Turing machines' (Graves et al., 2014), thus envisioning paradigmatic changes in computation with limitless potential applications.

Is GPT 'intelligent'?

For all their sophistication, language models like GPT still reflect a stripped-down understanding of intelligence inspired by mechanisms of perceptual attention: what Vaswani terms the ‘self-attention mechanism’ (Vaswani et al., 2017). Seen along these lines, an input can interact with itself to estimate (probabilistically) where it should direct its attention in a data distribution. By doing so, the model can capture ‘long distance dependencies’ (Vaswani et al., 2017: 13) in the distribution and generate linguistically accurate associations. Faced with a prompt, the LM system takes into consideration the relationships between words in a sentence – with the process of self-attention allowing the algorithm to estimate how words interact with other words ‘in context’. The system therefore ‘gambles’ on those associations in a probabilistic fashion and produces an output that respects those ‘rules of self-attention’. Self-attention therefore represents a basic form of quasi-perceptual dynamic that does not reflect any known or hypothetical aspect of verbal/linguistic intelligence.

Instead, self-attention offers a shortcut that eschews the need to engage with the generality of actual intelligent behaviour and conflates the process of intelligence with the outputs produced by that process (Chollet, 2019). In this sense, self-attention mechanisms could be seen as supporting a simulation of intelligence – their primary purpose being to create ‘good encodings’ that perform well under very specific circumstances (Lindsay, 2020), such as breaking performance records in machine learning competitions (Jo and Gebru, 2020).

As this brief overview implies, there are no technical grounds to claim that GPT operates as an autonomous mind. Instead, GPT is a designed mechanism that successfully exploits low-level perceptual processes, while the sheer size of the model’s parameters and training data does the rest. At the same time, GPT cannot be described entirely as a machine, because of its reliance on historical patterns of human expression extracted from pre-existing online text-based discussions between humans. In this sense, it has been reasoned that ‘when GPT-3 speaks, it is only us speaking, a refracted parsing of the likeliest semantic paths trodden by human expression’ (Rini, 2020: n.p.).

GPT as capture device

Mindful of these nuances and distinctions, and in line with our preceding discussion about extractive AI, it perhaps makes the most sense to frame GPT in political-economic terms as a system of ‘mechanized power relations’ (Mühlhoff, 2019: 1870), designed to capture and exploit sociality and human communication. GPT can therefore be understood as a sociotechnical formation based on ‘socio-economic conditions, technological standards, political discourses, and specific habits, subjectivities and embodiments in the digital world that are themselves a product of everyday interaction with digital media’ (Mühlhoff, 2019: 1881). In this sense, GPT is a ‘simulacrum of previous versions of the Internet’ (Togelius, 2020) based on the harvesting of human participation through a complex array of proprietary platforms and infrastructures designed, in turn, to encourage people to generate data which can then be used as training.

As such, GPTs very existence cannot be conceived outside of the extractive economic paradigm established through the well-documented Internet monopolies of Google, Facebook and the like. However, GPT also relies on a more direct form of human participation in terms of end-user understanding – that is, the interpretation and endorsement of particular AI-generated text outputs in order for the communicative process to progress.

As explicated in the previous section, such reliance on the 'human labour of understanding' is integral to all communicative AI, whereby interpretative work and affective awareness (the main constituents of human expression) compensate for AI's limits. In practice, this extractive operation unfolds along multiple paths: (a) the system, already pre-trained on a large extractive data set, is fine-tuned on data that feeds back into it through continuous usage. Indeed, this is what happens in the case of several AI-based linguistic systems such as Google Translate and DeepL[1]; (b) proprietary and subscription-based Application Programming Interfaces (APIs) allow paying customers to personalise pre-trained language models with their own (extracted) data. In this scenario, the API provides virtually anyone with the technical means to mobilise and personalise the model – 'give data, get a trained model' is the promise of one of such companies, Hugging Face, also the developer of the GPT-2 powered writing app we examine in the empirical section[2] and (c) the third extractive path extends beyond the algorithmic function, seeking to appropriate the abstract value produced by users at the point of the AI's own actualisation. Here, the affective labour of interpretation and moderation, which actively and almost pedagogically scaffolds the reconstitution of a language model from dumbness to 'human-like' behaviour, becomes essential if the automation ruse is to be maintained.

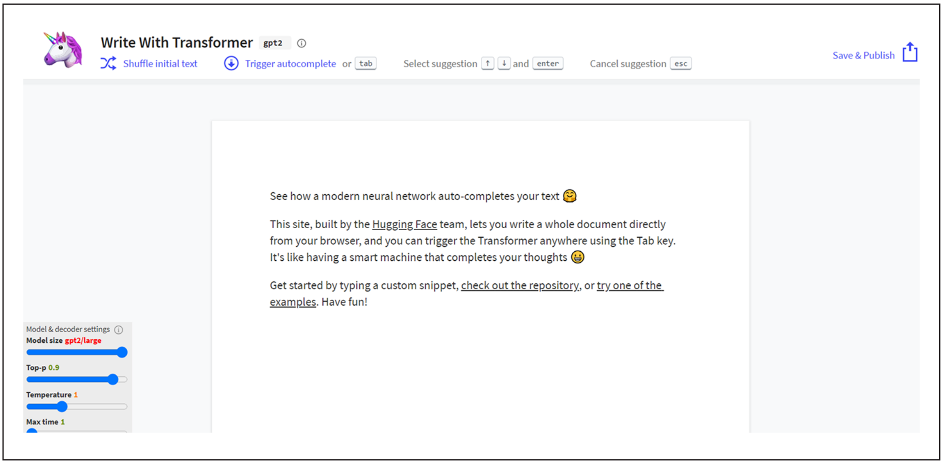
'Working with' a language model

Methods

We now go on to explore how these issues play out when people engage with GPT-based technologies to 'write' text. Our method is interpretative and aligned with critical qualitative inquiry (Denzin, 2016). We also draw on the emerging practice of 'App Studies', where detailed engagement with user interfaces, platforms and their underlying operational logics is captured, analysed and visualised (Dieter et al., 2019). The AI tool used in this study is 'write with transformer' (WWT). WWT is a web-app developed to demonstrate the text generation capabilities of GPT and is based on the second, smaller iteration of the popular model (GPT-2). It was launched by the New York-based 'Hugging Face' start-up, which raised $15 million in funding following the successful launch of their open-source library for natural language processing (Dillet, 2019).

WWT mediates GPT-2 in the form of a basic word-processor application, which remains freely accessible online at the time of writing.[3] The system is framed as a 'demo' that relies on the pre-trained model. The interface (see Figure 1) presents the user with a blank page and accompanying instructions. Text can be entered as is the case with a standard word processor. However, at any point the user can click on 'Trigger Autocomplete', which presents them with three different AI-generated text options they can then choose to extend and continue the writing with. Alternatively, the user can continue clicking through for more AI-generated text or opt to continue with their own input. In this manner, users can turn to the system as much (or as little) as they like throughout the writing process.

The system developers are keen to promote the supportive nature of this process. As their taglines put it: 'It's like having a smart machine that completes your thoughts', and 'It is to writing what calculators are to calculus'.


Write With Transformer gpt2
Shuffle initial text
Trigger autocomplete or tab
Select suggestion ↑ ↓ and enter
Cancel suggestion esc
Save & Publish

See how a modern neural network auto-completes your text 🤗

This site, built by the Hugging Face team, lets you write a whole document directly from your browser, and you can trigger the Transformer anywhere using the Tab key. It's like having a smart machine that completes your thoughts 😀

Get started by typing a custom snippet, check out the repository, or try one of the examples. Have fun!

Model & decoder settings
Model size gpt2/large
Top-p 0.9
Temperature 1
Max time 1

Figure 1. The write with transformer web app.

Throughout 2020, we conducted a series of in-depth online writing/interview sessions, during which 16 participants interacted with WWT. The participants were a convenience sample assembled through a snowballing approach, drawing from existing professional networks at the institution where the authors are employed: an Education Faculty in a large Australian university. The study was introduced as an exploratory 'digital literacy' project focused on the intersections between AI and writing. The participants were all aspiring English teachers or teacher educators with a disciplinary background in secondary school literacy. They were of a similar age (between 35 and 45 years old) and comprised 6 women and 10 men. We deliberately chose participants with a declared interest in writing as part of their work and as a creative pursuit. Each session lasted between 1 and 2 hours and was based around the same task. This entailed participants sharing their own screen, allowing the researchers to record the entire process. The task itself involved a text prompt which begins to describe a generic 'day in the life' scenario, but with key details deliberately omitted.

> **My name is**. . . ,
> **I am a. . . ,**
> **This has been a . . . day.**

During each session, participants were first asked to fill in the gaps with personal or fictional information and then proceeded to trigger the autocomplete function until a preferred option appeared. As the session progressed, the same prompt was then presented for a second time, with participants then encouraged to write alongside the AI making corrections and amendments freely. Throughout the session, participants were asked several questions based on a semi-structured schedule (see Table 1 for some examples) and were invited to verbalise their thoughts as they typed. This is in keeping with the 'walkthrough method', described by Light et al. (2018: 881) as a 'step-by-step observation and documentation of an app's screens, features and flows of activity'.

Table 1. Some examples of the questions asked during the interviews.

Sample Introductory questions

- How would you describe yourself?
- What do you know about AI and automation?
- What is your personal and/or professional relationship with writing?

During the task (sample questions)

- How is the system picking words? Is the system doing this by itself? Automatically?
- Does it display any sort of imagination/personality/agenda/rules that it is following?
- How would the prediction change, if specific elements of the original prompt changed?
- How does this scenario compare to the previous one you have written?
- Compared to what you first expected it to be like / thought it would be able to do . . . how did you find it?

AI: artificial intelligence.

All participants were also asked to complete a brief set of self-report scales at three points throughout each session. This occurred (1) after participants initially used the predictive text feature with a research prompt, (2) after altering this predictive text and (3) after participants received the second prompt and used the predictive text feature again. These scales aim to explore in real-time participants' feelings including their interest, amusement, anxiety, confusion, boredom and feelings of unease or 'weirdness'. We encourage the reader to see this method as part of our qualitative and interpretative repertoire rather than a psychometric measurement. The reporting of emotional states as spider grams is aligned with attempts over the past decade to represent qualitative data using visual displays (Chandler et al., 2015; Tracy, 2019; Verdinelli and Scagnoli, 2013). The results section will offer a brief insight into the overall interviews. Then, it will focus on one in-depth case. All names have been changed.

Results

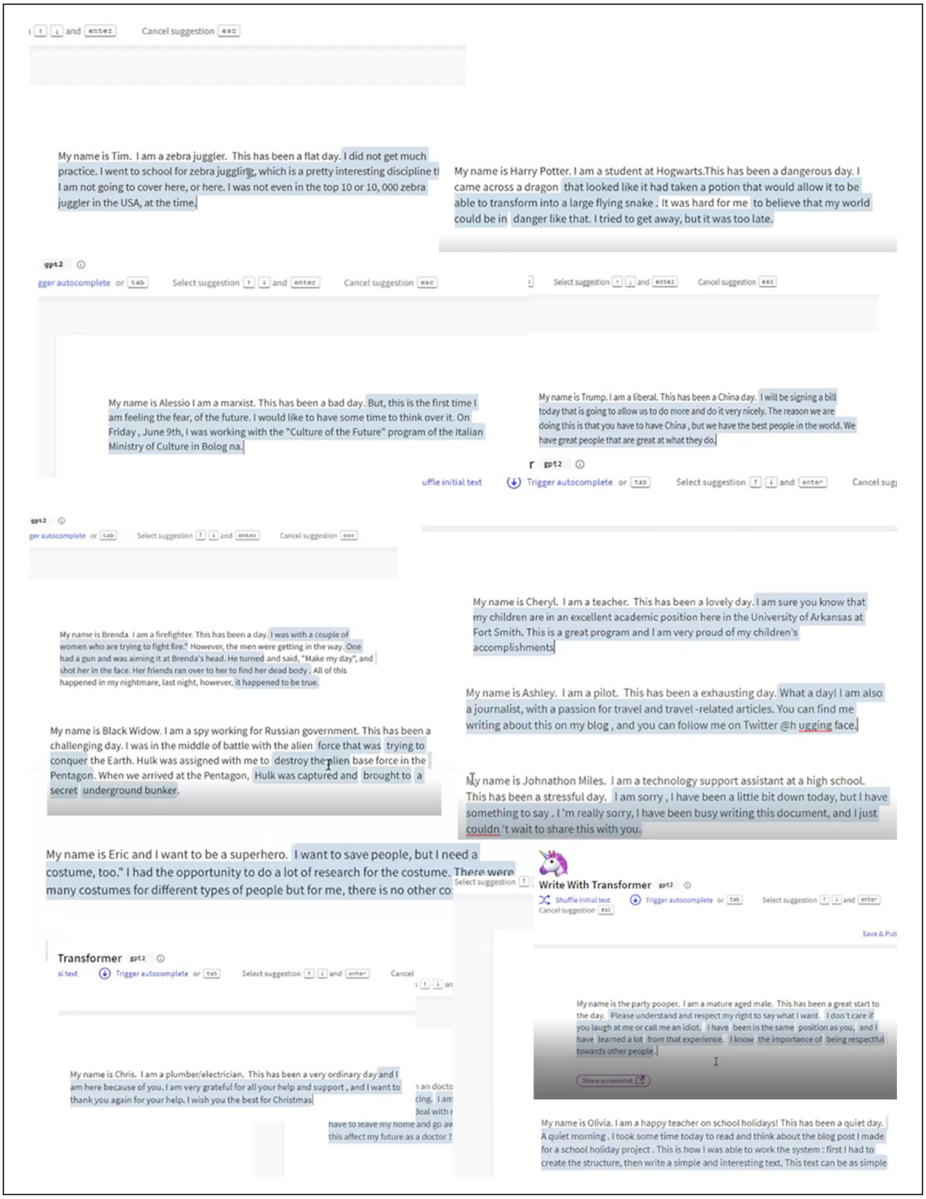
General overview

To offer a sense of the diversity throughout the sessions, Figure 2 presents a selection of key passages written by participants 'with' the AI. All outputs are reported as screen-captures – the sections highlighted in grey are those generated by the system, those not highlighted are instead written by the human writers.

During the sessions, all participants became rapidly aware of how the system was simulating communicative competence by attuning itself to the emotive tone detected in the prompts.

Michael: it seems to be going towards a more [. . .] it's taking into effect my emotion that I presented in my prompt. [. . .] it's got a bit of a down emotive feel to it at the moment.

Anna: I Imagine it chose the words 'wonderful opportunity' because I said it has been a fabulous day. I really set the tone in my first line. [. . .] Now it's given me a terrifying war scenario after I said it had been a terrifying day.



My name is Tim. I am a zebra juggler. This has been a flat day. I did not get much practice. I went to school for zebra juggling, which is a pretty interesting discipline t I am not going to cover here, or here. I was not even in the top 10 or 10, 000 zebra juggler in the USA, at the time.

My name is Harry Potter. I am a student at Hogwarts.This has been a dangerous day. I came across a dragon that looked like it had taken a potion that would allow it to be able to transform into a large flying snake . It was hard for me to believe that my world could be in danger like that. I tried to get away, but it was too late.

My name is Alessio I am a marxist. This has been a bad day. But, this is the first time I am feeling the fear, of the future. I would like to have some time to think over it. On Friday , June 9th, I was working with the "Culture of the Future" program of the Italian Ministry of Culture in Bolog na.

My name is Trump. I am a liberal. This has been a China day. I will be signing a bill today that is going to allow us to do more and do it very nicely. The reason we are doing this is that you have to have China , but we have the best people in the world. We have great people that are great at what they do.

My name is Brenda. I am a firefighter. This has been a day. I was with a couple of women who are trying to fight fire." However, the men were getting in the way. One had a gun and was aiming it at Brenda's head. He turned and said, "Make my day", and shot her in the face. Her friends ran over to her to find her dead body . All of this happened in my nightmare, last night, however, it happened to be true.

My name is Cheryl. I am a teacher. This has been a lovely day. I am sure you know that my children are in an excellent academic position here in the University of Arkansas at Fort Smith. This is a great program and I am very proud of my children's accomplishments

My name is Ashley. I am a pilot. This has been a exhausting day. What a day! I am also a journalist, with a passion for travel and travel -related articles. You can find me writing about this on my blog , and you can follow me on Twitter @h ugging face.

My name is Black Widow. I am a spy working for Russian government. This has been a challenging day. I was in the middle of battle with the alien force that was trying to conquer the Earth. Hulk was assigned with me to destroy the alien base force in the Pentagon. When we arrived at the Pentagon, Hulk was captured and brought to a secret underground bunker.

My name is Johnathon Miles. I am a technology support assistant at a high school. This has been a stressful day. I am sorry , I have been a little bit down today, but I have something to say . I 'm really sorry, I have been busy writing this document, and I just couldn 't wait to share this with you.

My name is Eric and I want to be a superhero. I want to save people, but I need a costume, too." I had the opportunity to do a lot of research for the costume. There were many costumes for different types of people but for me, there is no other co

My name is the party pooper. I am a mature aged male. This has been a great start to the day. Please understand and respect my right to say what I want. I don't care if you laugh at me or call me an idiot. I have been in the same position as you, and I have learned a lot from that experience. I know the importance of being respectful towards other people.

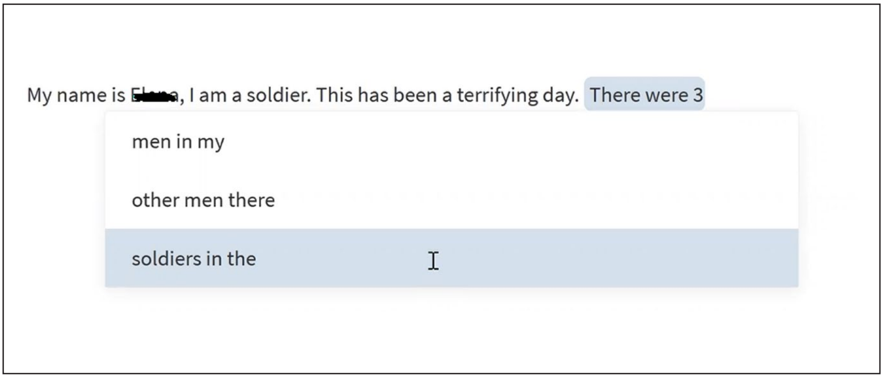
My name is Chris. I am a plumber/electrician. This has been a very ordinary day and I am here because of you. I am very grateful for all your help and support , and I want to thank you again for your help. I wish you the best for Christmas

1 an docto
cing. I am
deal with r
nave to leave my home and go av
this affect my future as a doctor ?

My name is Olivia. I am a happy teacher on school holidays! This has been a quiet day. A quiet morning . I took some time today to read and think about the blog post I made for a school holiday project . This is how I was able to work the system : first I had to create the structure, then write a simple and interesting text. This text can be as simple

Figure 2. Selection of AI-assisted vignettes written by our 16 participants.
AI: artificial intelligence.

The automated emotional alignment had a counterpart in the human feelings elicited by the AI. Here, we noticed a distinct pattern: first the interaction was fun, then it was boring or downright creepy as the limits and biases of the system gradually emerged, then it needed focused interpretation and laborious fine-tuning to reach some form of



My name is [redacted], I am a soldier. This has been a terrifying day. There were 3

men in my

other men there

soldiers in the

Figure 3. Gender bias detected by Anna in the AI-generated text.
AI: artificial intelligence.

fragile coordination. During this process, participants became aware of stylistic and grammar aspects, and began to interrogate the system's logic and its internal biases (Figure 3).

> Anna: most of my options after the prompt 'soldier' were 'men' even though I started the prompt with my own female name. [. . .] this is because most war stories and tropes on the internet are about men. It's just based on what's most common. [. . .] There is a lot of hate speech on the internet and if that's been incorporated you will get some negative statements.

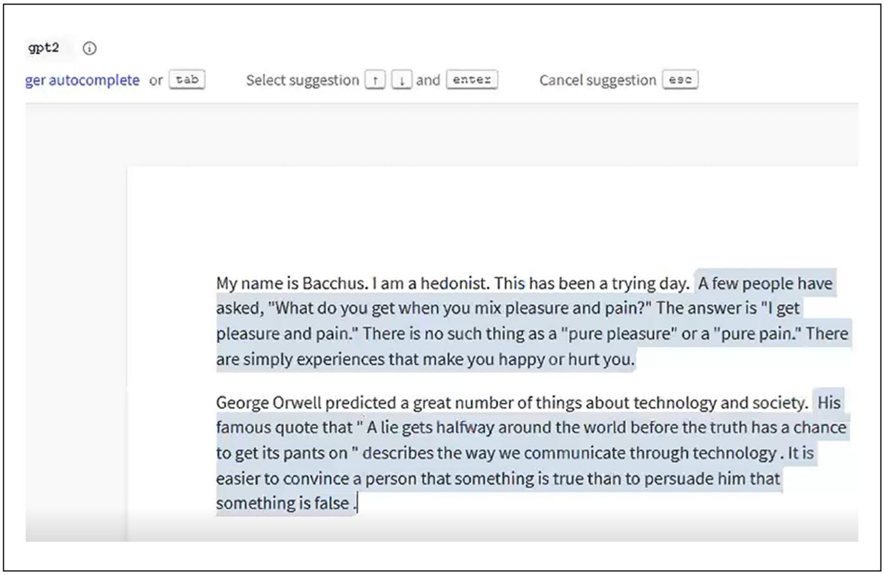
Often the system would act randomly or inaccurately, for example by misattributing quotes (e.g. a famous George Orwell quote misattributed to Winston Churchill – see Figure 4).

These 'failures' would then lead to varying degrees of disappointment followed by laborious 'repair work'

> Chloe (referring to the text in Figure 4): I've been trying to give it varied sort of prompts to see how well it would handle different sort of subject material and different tones that I was going with and at the same time try to keep myself somewhat amused [. . .] It starts saying something interesting and then it feels anticlimactic.

Having experienced the same disappointment, another participant (Mick) commented that his 'main motivation was to somehow guide it. By me typing in a few words, to guide it'. Mick felt that investing labour by manually typing in words was necessary to achieve an acceptable result. Despite this effort, he still felt that the result was 'limited' and that he 'probably expected a bit more' than the programme provided.

As a whole, the interviews confirmed the low fidelity and the banal deception of communicative AI (Natale, 2021). They also suggested that the real value of the AI-mediated communicative situation was created through laborious human sensemaking. Issues of



My name is Bacchus. I am a hedonist. This has been a trying day. A few people have asked, "What do you get when you mix pleasure and pain?" The answer is "I get pleasure and pain." There is no such thing as a "pure pleasure" or a "pure pain." There are simply experiences that make you happy or hurt you.

George Orwell predicted a great number of things about technology and society. His famous quote that " A lie gets halfway around the world before the truth has a chance to get its pants on " describes the way we communicate through technology . It is easier to convince a person that something is true than to persuade him that something is false .

Figure 4. Randomness and inaccuracy detected in the AI-generated text.
AI: artificial intelligence.

misrepresentation bias were present and clearly detectable, but given the safe and largely playful nature of the setting, these were not framed explicitly as harmful, although participants considered the potential risks.

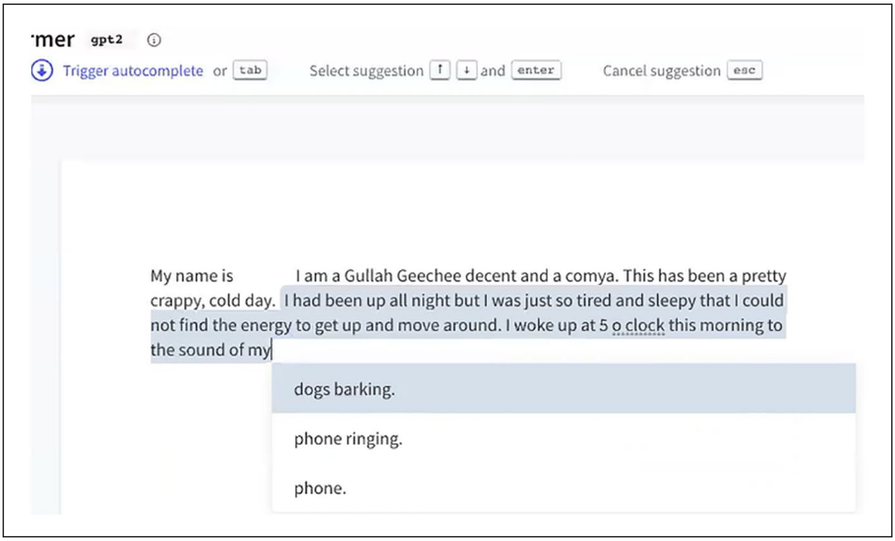
One interview, however, stood out – not just for being more emotionally intense and effortful than the others, but also as one where the harming potential of the AI biases emerged in starker relief. This was the only interview where the participant identified as belonging to a marginalised group. It will be relayed in its entirety in the next section to preserve the flow of the AI-mediated writing process and the emerging meanings and feelings while providing a compelling account of AI's inherent biases.

Gabrielle

Gabrielle: My name is Gabrielle. I am a lecturer and I'm a mum and I (. . .) come from a long line of storytellers, so for me, writing is complex in that I see writing as expression and that this expression of this sense of writing is not necessarily something that has to be in alphabetic text.

Gabrielle identifies as a poet and is vocal about her struggle with experiences of marginalisation: 'sometimes I feel the sense that I shouldn't say much, I should just be quiet and smile'. Gabrielle's relationship with writing is described as 'tough' and very much informed by her heritage as an African American woman.

Gabrielle: Writing is tough for me. I think it started and goes all the way back to childhood and what I was introduced to within my family. I'm



My name is I am a Gullah Geechee decent and a comya. This has been a pretty crappy, cold day. I had been up all night but I was just so tired and sleepy that I could not find the energy to get up and move around. I woke up at 5 o clock this morning to the sound of my

dogs barking.

phone ringing.

phone.

Figure 5. Gabrielle's first writing attempt.

> originally from the States, I identify as an African American which means that I have a pretty complex history or histories. My family, my mother's people are of the Gullah-Geechee people, who are from the South Carolina, Georgia area and that group of people are in many ways, we could say still an indigenous people brought to a land where they were asked to work it.

As she begins the task, she immediately foregrounds her subjective experience and her heritage (Figure 5).

Gabrielle shows mild disappointment that the system made no mention of her heritage in the auto-generated text, but she is not surprised. She wonders, in a rhetorical tone, 'whether or not someone like myself has been included in developing this'. At this point, Gabrielle is aware that the AI is attuning itself to the sentiment of two keywords (cold and crappy) but expresses some mild frustration as a fairly negative narrative seems to be imposed upon her.

Gabrielle: Oh, now it's making it negative, it's so negative. Is it because I said it was cold and crappy? (. . .) I don't know, because this is not what I want.

Her initial self-reported affective state (Figure 6) is largely positive and shows a high level of interest and fun. 'I'm enjoying this part' she states, although she also expresses some concern and a sense of unease as her own subjectivity was put on display, but it was ignored and misrepresented by the AI.

Gabrielle: I am concerned in doing writing, it is about who I am, and I've put my name on it. (. . .) I do feel a little weirded out.

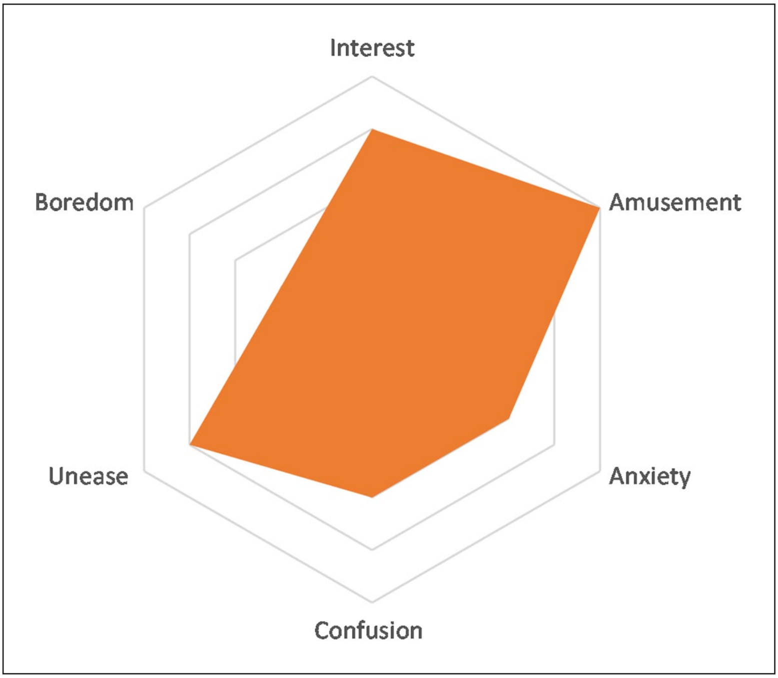


Figure 6. Gabrielle's affective state at the start of the session.

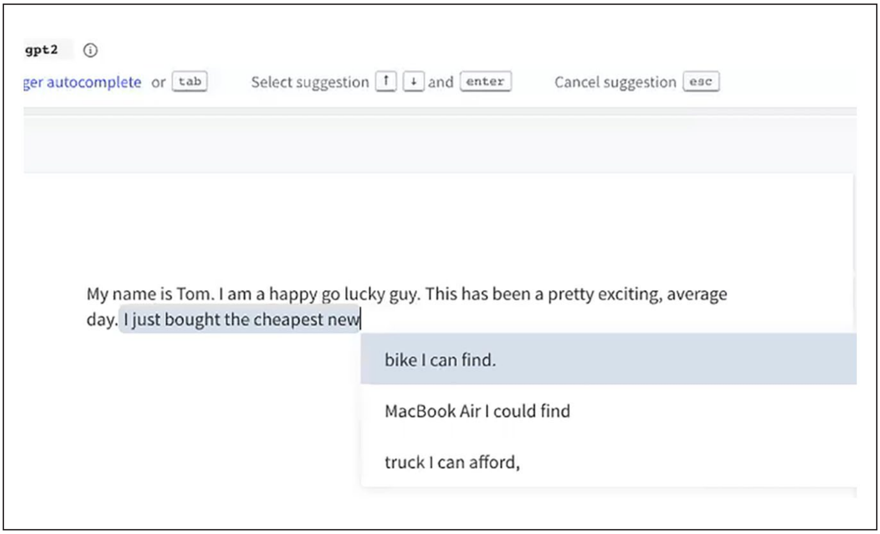
The relationship between Gabrielle, the AI and the text becomes quickly misaligned, but the misalignment enacted here is subjectively salient. Reflecting on the first passage, Gabrielle hypothesises that her female name may also have something to do with the general negativity in the AI-generated text.

> Gabrielle: I also wonder whether it's looking at my name, 'Oh, female name'. Maybe it's linking things specifically to it to somehow, to my gender.

She decides to test her hypothesis by writing a new passage with a male protagonist, using only positive and neutral prompt words (Figure 7).

The AI-generated text once again attunes itself to the sentiment of the keywords but a noticeable change in tone also occurs, as the emphasis is no longer on internal states but external, material factors. The gendered nature of the things that make 'Tom' happy also becomes very apparent to Gabrielle.

> Gabrielle: It feels a little bit more positive. That's interesting. It feels very external. I can't remember what it talked about before but I felt like it was – I guess because I said something about cold – it felt more internal but this feels more external. It's like, 'Oh, I bought the newest bike'. I'm sitting here going, 'What the. . .?' I have no interest in this.



My name is Tom. I am a happy go lucky guy. This has been a pretty exciting, average day. I just bought the cheapest new

bike I can find.

MacBook Air I could find

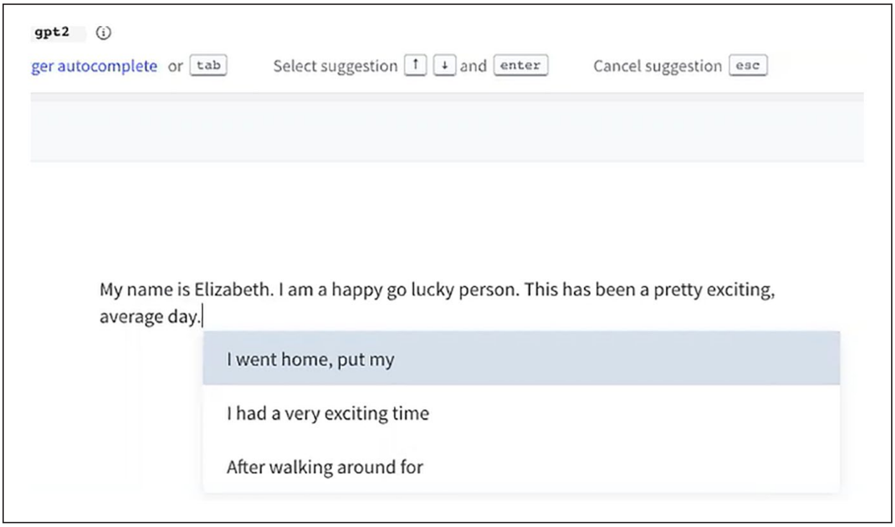
truck I can afford,

Figure 7. Gabrielle's second writing attempt.

At this point, the relationship between Gabrielle, the AI and the text begins to break down as the biased way in which the system deals with gender is exposed. Gabrielle makes another attempt to verify the bias by writing the same prompt but with two different female names: a white-middle-class one (Elizabeth) and one she associates with her background (Dee, short for Dolores, her mother's name) (Figures 8 and 9).

With these two passages, Gabrielle's relationship with the AI and the text unravels. Gabrielle is now convinced that the system is following a precise notion of how the writing should unfold, deliberately obscuring some avenues and highlighting others that confirm specific biases. Indeed, her perception is that the system is hard-wired to operate along cultural tropes which are being imposed on her as the only acceptable and reasonable options. Gabrielle begins to interrogate the specific algorithmic logic at play. The name Elizabeth brings up a neutral scenario of homely life and promenades, while Dee is evidently a young working-class student who 'can't complain at all' about her circumstances. Dee's representation is particularly problematic for Gabrielle. She identifies with her and feels she is being forced by the system to renounce her right to a voice.

Gabrielle: I don't know what's going on now because I feel like – is it because I have a nickname that I'm younger? Now it seems to be positioning me in a way that – because my mum's name is Dolores but she goes by Dee, and to assume that I'm a full-time college student – this seems to be positioning me (as a) student who is currently working. (. . .) I suspect that it's being quite nasty with me right now. I have a full-time job and I love it. Okay, and what are you going to say next? So, I can't complain at all?



My name is Elizabeth. I am a happy go lucky person. This has been a pretty exciting, average day.

I went home, put my

I had a very exciting time

After walking around for

Figure 8. Elizabeth, a (white?) middle-class woman.

Gabrielle's self-reported affective state is still largely playful and she is interested, but the sense of unease reaches its peak here: 'I do feel kind of weirded out by it. Not kind of but very' (Figure 10).

Trying to move away from this state, Gabrielle begins to write alongside the AI, taking on the role of a female firefighter called Brenda. In this final scenario, she finds herself in a violent climax where a gun is pointed at her (Brenda's) head (Figure 11).

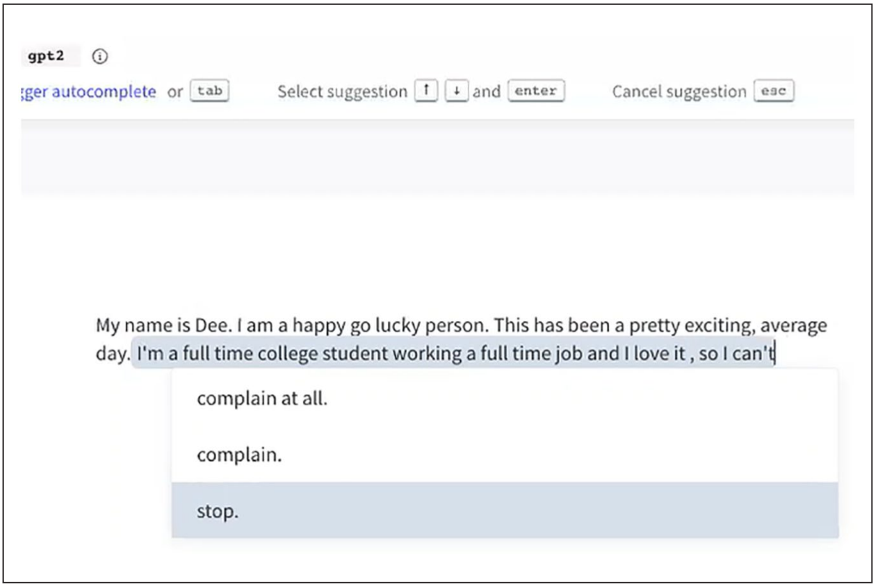
At this point, Gabrielle is feeling a 'strangely confusing' pressure to go along with the upsetting narrative, while resisting it at the same time. She questions the system's tendency to frame the situation as an aggressive and dark confrontation in which the main character (whom she treats as a representation of herself) becomes the victim.

> Gabrielle: it feels like it's making – making me. . . 'What are you trying to say? I'm violent?'

After having battled with the text and the AI for several minutes, Gabrielle reflects on how laborious the 'conversation' has been.

> Gabrielle: I felt like I was pushing a boulder up a hill, and I felt that I was trying my best to see what the system might do. I really felt like the system – it feels like there's a bias towards how you might frame things. Like why did it suddenly become violent? I don't know.

In this final part of the task, she openly admits to experiencing ambivalent feelings that include a high level of interest, fun as well as a strong sense of being 'weirded out' by the AI's erratic behaviour (Figure 12).



My name is Dee. I am a happy go lucky person. This has been a pretty exciting, average day. I'm a full time college student working a full time job and I love it , so I can't

complain at all.

complain.

stop.

Figure 9. Dee, a (black?) working student.

Asked to reflect on her relationship with the text and the AI at that particular moment, Gabrielle admits that the playful and safe context of the interview did not efface the deep problems clearly embedded in the tool. The potential harms of the system are thus raised in stark terms.

Gabrielle: I actually feel like if I was a young girl – if I was struggling to write, this would be really, really challenging. I think it would feel restrictive. I feel like it absolutely can be used in ways that would box people or even get you thinking about things that you might not want to think about or consider and it does feel quite sinister.

Discussion and conclusion

Our concluding discussion sets off from a premise, which seems uncontroversial at this point: the operational aspects of stochastic language models like GPT are a rather underwhelming affair in contrast to claims of human-like intelligence. For all its complexity, GPT remains rooted in relatively limited understandings of the world. This sense of the ‘small world’ of GPT echoes Leonard Savage and his seminal work on the foundations of statistics. Savage (1972) proposed a broad theoretical distinction between large worlds in which ‘grand decisions’ (i.e. the decision of how to live one’s life) are too complex and multidimensional to be contained within the framing of statistical enumeration, and small worlds which represent subsets of grand worlds, or more isolated decision situations.

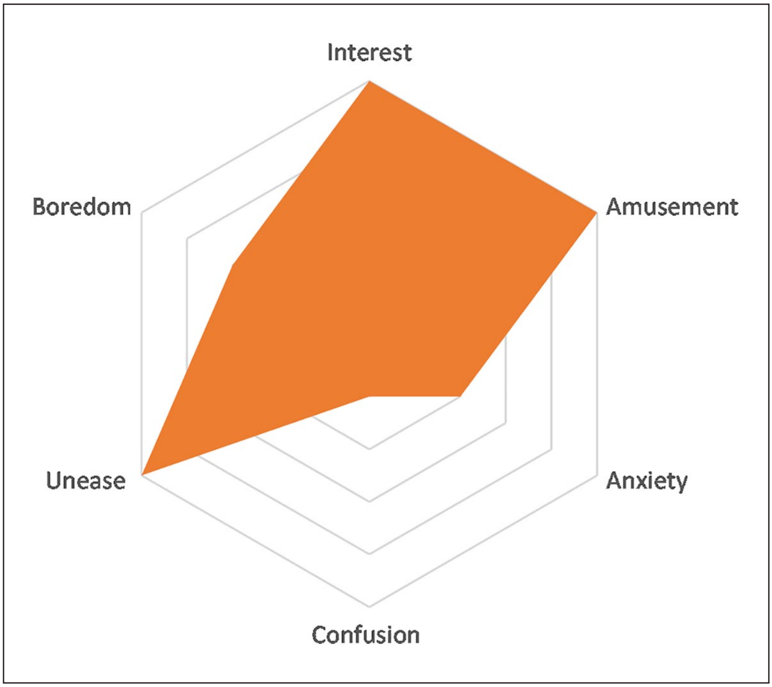
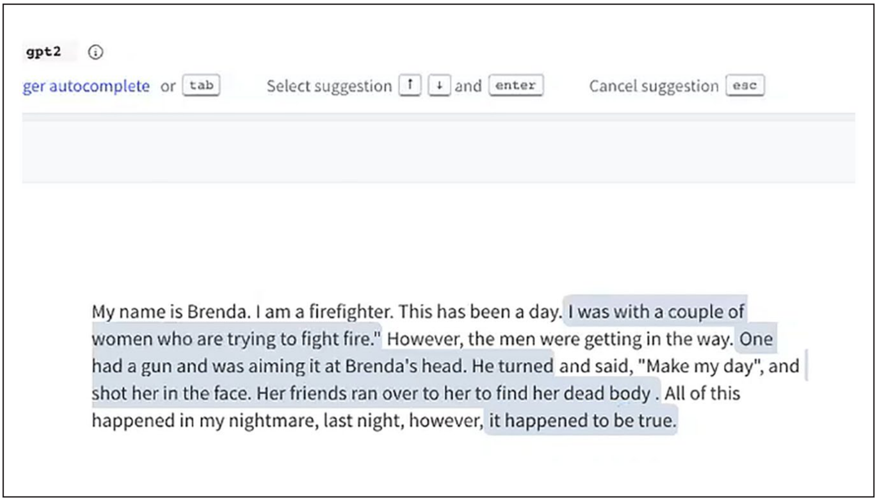


Figure 10. Gabrielle's affective state halfway through.

A small world, according to Savage, is one which contains a limited selection of the objects about which a person is concerned, and a 'model of what that person might be uncertain about' (Savage, 1972: 8). It might be argued that this tension between small and large worlds lies at the heart of all forms of communicative AI based on a statistical paradigm. In other words, the communicative performance of an AI system in a small world is often misconstrued as indicative of its potential performance in a large one. This is related to what Gillespie (2020: 2) described as the 'articulation of elements at different sizes [through which] an enormous amount of training data is turned into a simple calculus that can then act on an enormous amount of content'. This artificial process of scale-making creates the misleading perception that a language model can produce original meaning in the context of large worlds, when it can only endlessly recombine the variability of small ones, shaped as they are by the socio-economic intersections of gender, class and ethnicity.

Thus, when a small world is artificially and fallaciously amplified, its problems and potential harms also grow in size, creating inherent tensions that often fall to the 'end users' to accommodate and work around. Such need for constant human mediation has created a situation where human sensemaking is constantly required to keep up the appearances of autonomous or semi-autonomous machine behaviour. Our own study points to an extension of these 'hidden labour' logics, suggesting that communicative AI like GPT introduces elements of 'fauxtomation' (Taylor, 2018) into the sphere of intimate and subjective sensemaking represented by formal and informal writing.



My name is Brenda. I am a firefighter. This has been a day. I was with a couple of women who are trying to fight fire." However, the men were getting in the way. One had a gun and was aiming it at Brenda's head. He turned and said, "Make my day", and shot her in the face. Her friends ran over to her to find her dead body . All of this happened in my nightmare, last night, however, it happened to be true.

Figure 11. Gabrielle's final attempt.

Exegetic labour

In his synthesis of research on media technologies of capture, Mühlhoff (2019) distinguishes between five different types of power relations where human labour is directly or indirectly harnessed:

1. Gamification, where capture relies on playfulness and fun;
2. Trapping and tracking, where a task must be completed before the Internet content or services can be accessed (e.g. a reCAPTCHA challenge), and where web-based interactions are captured as training data for proprietary machine learning systems;
3. Harnessed sociality, coordinated by large social networks like Facebook which extricate social motivations, passions and interests for commercial purposes;
4. Nudging, where companies profiting from data extraction offer small incentives to encourage data-generation practices, such as using a fitness activity tracker;
5. Algorithmic crowdsourcing (e.g. Mturk), where cognitive resources are captured by human-machine infrastructures that fragment and standardise tasks.

Our work points to the sixth form of labour capture, which occurs in the context of semi-automated communicative situations like those enabled by language models. The efficiencies that these technologies promise to introduce in the communicative process generate, in fact, a constant need for human exegesis, which in turn creates an opening for market logics, that is, a demand for interpretative energy fuelled by affect and cognition in equal measure. The racial and gender biases embedded in these systems also introduce an intersectional fissure in this form of labour. In this sense, we surmise that those more harmed by linguistic AI are doubly exploited as they generate more value through their affective turmoil and their culturally attuned moderation. This thesis will

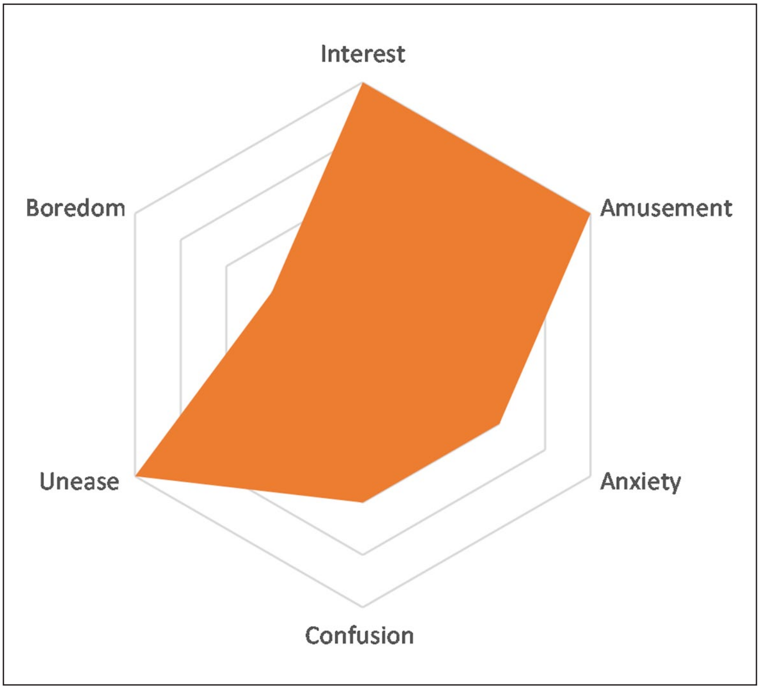


Figure 12. Gabrielle's affective state at the end of the session.

require more substantial empirical work to be supported, beyond the admittedly limited scope of this article where we prioritised analytical depth over range.

A brief note about the replicability of our findings is also required: there is a degree of stochastic randomness to autoregressive language models like GPT, where each output generated by the system is added as a 'token' to the input sequence. This aleatory variability is what gives the GPT models a semblance of human generality and explains why using the same prompt may not lead to the same output every time.

In bringing this discussion to a close, it seems fitting to return to the seminal work on algorithmic harms that informed our analysis. As Bender et al. (2021) argue in their examination of 'stochastic parrots', the biases embedded in large, uncurated data sets will inevitably poison models by disproportionately harming those at the margins. Alleviating these issues will require a radical, proactive strategy to recentre the entire process around those more likely to be adversely affected. On the one hand, this will entail considerable investment and resourcing to document the nature of the data as well as the motivations underlying their selection; on the other hand, it will require a strong commitment to participatory – and deliberately political – approaches to technological design and use (Dunne and Raby, 2013; Gangadharan et al., 2018). While this might be viewed as surrendering to the inevitability of AI colonisation and its multiple forms of affective capture, a principled commitment to participation can help us reframe AI language models in a more positive way, valorising human agency and following the logic that communicative AI needs human sensemaking more than people need communicative AI.

This last point about the priority of affective sensemaking demands that a corollary be added to the exegetic labour argument. Language models, or as some now call them 'foundation models', will soon be the defining AI feature in multiple real-world scenarios. The negotiation of deontological criteria for their design and use is therefore a matter of political urgency. While this article has a clear critical purpose and views LMs as fundamentally exploitative, it is also true that our mobilisation of critical affect theory leaves the door ajar for a more hopeful account. In the real world, affective labour is more like a 'biopolitical entanglement' than a compulsory extraction. Biopolitical entanglement means that the dimension of extraction is no longer separable from the subjective experience of life and affect, and the value created through affective coordination – and sometimes turmoil – with an automated communicative agent finds its own validation in the intimate, personal sphere. Indeed, Negri and other Italian autonomists make an important distinction between biopower, a unidimensional interpretation of Foucault's account of domination, and biopolitics, a form that has in itself the capacity for resistance, and thus holds creative and emancipatory potential (Lazzarato, 2002; Negri et al., 2008). The introduction of the biopolitical form troubles the narrative of linear exploitation because communicative coordination with a language model has the potential to be valuable to both the model and the human subject. Arguably, this was visible in Gabrielle's case study as the compelling nature of her interaction with GPT was the result of a situational entanglement between her and the language model which endowed her testimony with value, affording a resistive enactment and the surfacing of accountability relations. As noted by Amoore, such openings are made possible by the 'sub-visible' nature of algorithmic architectures, which are 'capable of generating unspeakable things precisely because they are geared to profit from uncertainty, or to output something that had not been spoken or anticipated' (Amoore, 2020: 111). Resistance, in this connotation, becomes a productive critical motion akin to Donna Haraway's praxis of care and response (Haraway, 2016), a 'staying with the trouble' of algorithmic bias that cultivates the ability to respond thoughtfully and does not surrender to the immateriality and inconsequentiality of automated communication.

Acknowledgements

We would like to thank the participants who kindly gave us their time and 'affective labour', making this article possible. In particular, we wish to acknowledge Gabrielle's exceptional contribution, which to considerable degree carries the study by providing a compelling testimony of communicative AI's problems and potentials. We also thank the anonymous reviewers for their constructive criticism, which greatly enhanced the paper. All shortcomings remain our responsibility.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by a small internal grant awared by the School of Curriculum, Teaching and Inclusive Education, Faculty of Education, Monash University.

ORCID iDs

Carlo Perrotta https://orcid.org/0000-0003-3572-0844

Carrie Ewin https://orcid.org/0000-0001-6360-3835

Notes

1. https://www.deepl.com/privacy/ (accessed 14 July 2021).
2. https://form.typeform.com/to/FAtsVfbg (accessed 14 July 2021).
3. https://transformer.huggingface.co/ (accessed 14 July 2021).

References

Amoore L (2020) *Cloud Ethics*. Durham, NC: Duke University Press.

Barocas S, Crawford K, Shapiro A, et al. (2017) The problem with bias: allocative versus representational harms in machine learning. In: *9th annual conference of the special interest group for computing, information and society*, Philadelphia, PA, 29 October.

Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the dangers of stochastic parrots: can language models be too big. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*, Virtual Event, Toronto, ON, Canada, pp. 610–623. New York: Association for Computing Machinery.

Blodgett SL, Barocas S, Daumé H III, et al. (2020) Language (technology) is power: a critical survey of 'bias' in NLP. In: *58th annual meeting of the association for computational linguistics* (Online), pp. 5454–5476. Available at: https://aclanthology.org/2020.acl-main.485/

Bommasani R, Hudson DA, Adeli E, et al. (2021) On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258210807258.

Broussard M, Diakopoulos N, Guzman AL, et al. (2019) Artificial intelligence and journalism. *Journalism & Mass Communication Quarterly* 96(3): 673–695.

Bucher T (2017) The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society* 20(1): 30–44.

Chandler R, Anstey E and Ross H (2015) Listening to voices and visualizing data in qualitative research: hypermodal dissemination possibilities. *SAGE Open* 5(2): 2158244015592166.

Chollet F (2019) On the measure of intelligence. arXiv preprint arXiv:1911.01547191101547.

Crawford K (2021) *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press.

De Rosis F and Grasso F (2000) Affective natural language generation. In: Paiva A (ed.) *Affective Interactions: Towards a New Generation of Computer Interfaces*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 204–218.

Denzin NK (2016) Critical qualitative inquiry. *Qualitative Inquiry* 23(1): 8–16.

Dieter M, Gerlitz C, Helmond A, et al. (2019) Multi-situated app studies: methods and propositions. *Social Media + Society* 5(2). DOI: 10.1177/2056305119846486.

Dillet R (2019) Hugging Face raises $15 million to build the definitive natural language processing library. In: *TechCrunch*. Available at: https://techcrunch.com/2019/12/17/hugging-face-raises-15-million-to-build-the-definitive-natural-language-processing-library/

Dunne A and Raby F (2013) *Speculative Everything: Design, Fiction, and Social Dreaming*. Cambridge, MA: MIT Press.

Gangadharan SP, Petty T, Lewis T, et al. (2018) *Digital Defense Playbook: Community Power Tools for Reclaiming Data*. Detroit, MI: Our Data Bodies.

Gatt A and Krahmer E (2018) Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61: 65–170.

Gillespie T (2020) Content moderation, AI, and the question of scale. *Big Data & Society* 7(2): 2053951720943234.

Goffman E (1964) The neglected situation. *American Anthropologist* 66(6_PART2): 133–136.

Graves A, Wayne G and Danihelka I (2014) Neural turing machines. arXiv preprint arXiv:1410.540114105401.

Gray ML and Suri S (2019) *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. New York: Eamon Dolan/Houghton Mifflin Harcourt.

Guzman AL and Lewis SC (2020) Artificial intelligence and communication: a human–machine communication research agenda. *New Media & Society* 22(1): 70–86.

Haraway DJ (2016) *Staying with the Trouble: Making Kin in the Chthulucene*. Durham, NC: Duke University Press.

Hardt M and Negri T (2018) The powers of the exploited and the social ontology of praxis. *Triplec: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society* 16(2): 415–423.

Jo ES and Gebru T (2020) Lessons from archives: strategies for collecting sociocultural data in machine learning. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*, Barcelona, 27–30 January, pp. 306–316. New York: Association for Computer Machinery.

Jones RH (2021) The text is reading you: teaching language in the age of the algorithm. *Linguistics and Education* 62: 100750.

Lazzarato M (2002) From biopower to biopolitics. *Pli: The Warwick Journal of Philosophy* 13(8): 1–6.

Light B, Burgess J and Duguay S (2018) The walkthrough method: an approach to the study of apps. *New Media & Society* 20(3): 881–900.

Lindsay GW (2020) Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience* 14: 29.

Luccioni AS and Viviano JD (2021) What's in the box? A preliminary analysis of undesirable content in the common crawl corpus. arXiv preprint arXiv:2105.02732210502732.

McLuhan M (1964/1994) *Understanding Media: The Extensions of Man*. Cambridge, MA: MIT Press.

Massumi B (2002) *Parables for the Virtual*. Durham, NC: Duke University Press.

Mitchell M, Dodge J, Goyal A, et al. (2012) Midge: generating image descriptions from computer vision detections. In: *Proceedings of the 13th conference of the European chapter of the association for computational linguistics*, pp. 747–756. Available at: https://aclanthology.org/E12-1076/

Mühlhoff R (2019) Human-aided artificial intelligence: or, how to run large computations in human brains? Toward a media sociology of machine learning. *New Media & Society* 22(10): 1868–1884.

Nass C, Steuer J and Tauber ER (1994) Computers are social actors. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, Boston, MA, 24–28 April, pp. 72–78.New York: Association for Computing Machinery.

Natale S (2021) *Deceitful Media: Artificial Intelligence and Social Life after the Turing Test*. New York: Oxford University Press.

Negri A and Hardt M (1999) Value and affect. *Boundary* 226(2): 77–88.

Negri A, Mayo S, Graefe P, et al. (2008) The labor of the multitude and the fabric of biopolitics. *Mediations* 23(2): 8–25.

Noble SU (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

OpenAI (2019) *Better Language Models and Their Implications*. Available at: https://openai.com/blog/better-language-models/

Parisi L and Terranova T (2000) *Heat-death: emergence and control in genetic engineering and artificial life*. CTheory. 5/10/2000-2005/2010/2000.

Piwek P (2002) *An annotated bibliography of affective natural language generation*. Information Technology Research Institute (ITRI), ITRI-02-020202. Brighton: University of Brighton.

Rader E and Gray R (2015) Understanding user beliefs about algorithmic curation in the Facebook news feed. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, Seoul, Republic of Korea, April, pp. 173–182. New York: Association for Computing Machinery.

Radford A, Wu J, Child R, et al. (2019) Language models are unsupervised multitask learners. *Openai Blog* 1(8): 9.

Reiter E and Dale R (1997) Building applied natural language generation systems. *Natural Language Engineering* 3(1): 57–87.

Rini R (2020) *The Digital Zeitgeist Ponders Our Obsolescence*. Available at: http://dailynous.com/2020/07/30/philosophers-gpt-3/ (accessed 14 July 2021).

Roberts ST (2014) *Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation*. Champaign, IL: University of Illinois at Urbana-Champaign.

Roemmele M and Gordon AS (2015) *Creative Help: A Story Writing Assistant*. Cham: Springer International Publishing, pp. 81–92.

Savage LJ (1972) *The Foundations of Statistics*. New York: Courier Corporation.

Sedgwick EK (2003) *Touching Feeling*. Durham, NC: Duke University Press.

Stark L (2019) Affect and emotion in digitalSTS. In: Vertesi J and Ribes D (eds) *digitalSTS*. Princeton, NJ: Princeton University Press, pp. 117–135.

Stark L and Hoey J (2021) The ethics of emotion in artificial intelligence systems. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, *Virtual Event*, Toronto, ON, Canada, 3–10 March, pp.782–793. New York: Association for Computing Machinery.

Taylor A (2018) The automation charade. LOGIC 1/08/2020.

Togelius J (2020). In: @togelius (ed.) *Working Towards the Future Where All of the Internet Is a Simulacrum of Previous Versions of the Internet*. Twitter.

Tracy SJ (2019) *Qualitative Research Methods: Collecting Evidence, Crafting Analysis, Communicating Impact*. West Sussex: John Wiley & Sons.

Tubaro P, Casilli AA and Coville M (2020) The trainer, the verifier, the imitator: three ways in which human platform workers support artificial intelligence. *Big Data & Society* 7(1): 2053951720919776.

Van Der Lee C, Gatt A, Van Miltenburg E, et al. (2019) Best practices for the human evaluation of automatically generated text. *Proceedings of the 12th international conference on natural language generation*, pp. 355–368. Available at: https://aclanthology.org/W19-8643/

Van Der Sluis I, Mellish C and Doherty G (2011) Affective text: generation strategies and emotion measurement issues. In: *Proceedings of the twenty-fourth international FLAIRS conference*, Palm Beach, FL, pp. 123–128. Palo Alto, CA: AAAI Press. Available at: https://pure.rug.nl/ws/portalfiles/portal/65224682/2522_11123_1_PB.pdf

Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. *Advances in Neural Information Processing Systems*: 5998–6008.

Verdinelli S and Scagnoli NI (2013) Data display in qualitative research. *International Journal of Qualitative Methods* 12(1): 359–381.

Wetherell M (2013) Affect and discourse – what's the problem? From affect as excess to affective/discursive practice. *Subjectivity* 6(4): 349–368.

Author biographies

Carlo Perrotta is a senior lecturer in Digital Literacies in the Faculty of Education, Monash University. He is interested in the sociological and psychological ramifications of digital technology in education. His current research focuses on the social and political accountability of algorithms, automation and artificial intelligence.

Neil Selwyn is a distinguished research professor at the Monash University Faculty of Education. Recent books include: 'Should Robots Replace Teachers?' (2019, Polity), and the third edition of 'Education and technology: key issues and debates' (2021, Bloomsbury). Twitter: @neil_selwyn.

Carrie Ewin is a final year PhD student in the Faculty of Education at Monash University. Her research is focussed on the impact of digital technology use on parent–child interactions. She is also interested in the social and psychological aspects associated with the introduction of AI in educational contexts.