Should we have seen the coming storm? Transformers, society, and CC

Carolyn E. Lamb[1] and Daniel G. Brown[2]
[1] School of Computing, Queen's University, Kingston, Ontario, Canada
[2] Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada
cel4@queensu.ca, dan.brown@uwaterloo.ca

Abstract

Since ICCC 2022, transformer models easily usable through natural language prompts have changed the face of computational creativity. They raise disquieting social and legal issues. The ICCC community was to a great extent unprepared. We review some harms, dangers, and questions raised by transformer models and recommend that the CC community must more widely and urgently attend to the social impacts of CC.

Introduction

Large Language Models (LLMs), particularly based on transformers (Vaswani et al. 2017), have shown increasing success at generating text and other forms of media, such as visual art, based on natural language prompts. This success has brought us to a tipping point in the social role of computational creativity (CC). Suddenly, our field of research is no longer an obscure curiosity, but a major news item.

This increased prominence comes with increased ethical scrutiny and fear of harm. In fact, the release of high quality transformers that can be used by a human without any programming ability, like Midjourney and ChatGPT, has already significantly affected human artists. The level of ethical, societal, and legal concern caused by these new systems appears to have taken our community by surprise. While some may draw a technical distinction between CC and generative AI, the two fields are sufficiently similar that public opinion and social impact are not likely to distinguish between them.

We argue that the CC community must pay urgent attention to LLMs' ethical issues. Otherwise we risk losing relevance as public perceptions and concerns regarding CC systems shift out from under us. Transformers risk the "malevolent creativity" described by Cropley et al (2008), where a system's output is novel and valuable to *someone*, but harmful to society at large.

We summarize a few of these ethical issues and perform a brief literature review showing that, while the CC community has been increasingly aware of creative possibilities of LLMs, analysis of their societal and legal effects has lagged. We finish with recommendations for how the CC community should address these issues.

Transformer models and society

Ethics of LLMs have been discussed since their inception. Bender et al (2021) provide an important summary. Unintended bias, toxicity in training sets, or deliberate misuse of these models all result in harmful output. The energy required for training and using these large models causes environmental harm. Below, we highlight some further issues relevant to CC.

Plagiarism and human replacement

The enormous training sets used by transformers contain copyrighted text and artwork by humans, collected by web crawling without consent (Zirpoli 2023). They can generate an unlimited amount of new work, including text or art "in the style of" a particular human, which can be used in place of the human's work. Art made by these models has already been used in contexts where traditionally a human would be employed (Schaub 2022). Generative AI may thus enable corporate groups to produce endless content for audiences without compensating rightsholders, destroying prospects for humans in creative careers (Sobel 2017).

In the US, "fair use" protects use of copyrighted materials for certain purposes, and might protect AI training (Lemley and Casey 2020), though other countries' laws may differ (Brown, Byl, and Grossman 2021). Sobel's (2017) review suggests a double bind: if AI training is not fair use, scientific progress is hindered, but if AI training *is* fair use, writers and artists suffer. The fair use claim is being tested in court via lawsuits against Midjourney and Stable Diffusion (Zirpoli 2023) for training on living artists' copyrighted work without consent. Getty Images has also sued Stability AI for training Stable Diffusion on Getty's photos without a license, sometimes producing output so close to training images that it contained the same watermark (Belanger 2023).

Professional writers also are concerned. For instance, the science fiction magazine Clarkesworld recently closed submissions due to a rush of AI-generated submissions (Xiang 2023b). Hundreds of low-quality AI-generated books have also appeared in Amazon's Kindle store (Bensinger 2023). Since Kindle Unlimited distributes proceeds between all participating authors, these AI authors are siphoning income from human authors (Scalzi 2023). In journalism, the use of AI to replace humans is also increasing (Sweney 2023), despite the LLMs' factual errors (Farhi 2023), and

in screenwriting, one of the issues raised in the recent WGA writer's strike is the potential replacement of human creative labor with AI (Shah 2023).

OpenAI estimates that 80% of the U.S. workforce will be affected by GPT (Eloundou et al. 2023), requiring significant public policy work. By emphasizing the effect of LLMs on professional human creators, we do not argue that creative careers deserve more protection than others; our focus is on human creators because they are of interest to the CC community and at risk of from our particular research.

To cause these problems, transformer models need not meet traditional CC benchmarks. Their work need not be indistinguishable from human's, or novel and valuable; they need not be autonomous; and they need not take on tasks an "unbiased observer" would deem creative. Transformer models can produce interesting stories when used co-creatively by skilled users (Ippolito et al. 2022), but the AI stories causing problems for Clarkesworld were submitted by scammers and had no artistic value: it is their sheer volume and imperviousness to automated detection that made the magazine's submissions process unworkable (Clarke 2023). To cause economic disruption, generative systems must only produce passable-looking content easily enough to be attractive to scammers, or cheaper for companies than hiring humans.

Content moderation and fabrication

Bender et al. (2021) show that generative models easily produce harmful content, ranging from subtle reflection of social bias to outright abuse, harassment, or hate speech. While OpenAI trained ChatGPT to produce less offensive outputs (Ouyang et al. 2022), it did so using low-paid workers in the Global South to identify harmful content (Perrigo 2023). This human-in-the-loop learning is a common practice for various aspects of large AI models but it can be exploitative, particularly for moderation tasks which expose workers repeatedly to violent and pornographic material. Many workers developed symptoms of PTSD (Perrigo 2023).

Even after training, ChatGPT produces spurious medical advice and other forms of misinformation (Birhane and Raji 2022). This may be an unsolvable problem for transformers, which do not understand the meaning of their output: their high performance on benchmark tests is due to use of statistical cues, not comprehension (Niven and Kao 2019).

Many CC researchers remove offensive content from their system's results by hand. At the scale of GPT-4, this is not possible. Until a radically different method of content moderation is discovered, all researchers working on models of this size face a choice between exploiting workers like this or allowing their AI to generate potentially unlimited hate speech, as in the case of Microsoft's Tay chatbot (Wolf, Miller, and Grodzinsky 2017). The impossibility of ethically moderating content at this scale is, itself, an argument against the use of LLM-sized models.

Public versus private science

Another issue with transformer models is the extent to which research is not peer reviewed. Most OpenAI papers are released on the arXiv. The white paper for GPT4 was released directly by OpenAI, and has no details about GPT-4's data set, training, parameter counts or efficiency (OpenAI 2023).

Reproducibility is thus nearly impossible, as is systematic critique of a model's weaknesses. And companies working on LLMs have worked to stifle such critique. Google famously fired both leads of its ethical AI team after they criticized Google's LLM (Schiffer 2021), and Microsoft cut an ethics group at the exact time it expanded its relationship with OpenAI (Schiffer and Newton 2023).

Researchers at publicly-funded universities struggle to replicate corporate LLM research, both due to the prodigious size of these models and due to ethical concerns. Researchers might be interested in content moderation, for instance, but current content moderation techniques would present difficulties at a university Research Ethics Board. Corporations can build these models and write papers about their outcomes regardless of ethical concerns. While some journals and conference require that their research satisfies ethical standards, the use of arXiv or other non-peer-reviewed venues frees non-academic developers from this constraint.

Academics collaborating with corporations have also avoided scrutiny. One study tested an LLM-based mental health intervention on suicidal teenagers without informed consent. Because the intervention had been designed and implemented by a startup, and the university researchers only analyzed data after the fact, the study was considered to be "non-human subjects research" and the REB did not enforce any protections (Xiang 2023a).

The result is a situation where academic researchers *cannot* reproduce transformer models and *cannot* work at improving their basic mechanisms, but *can* collaborate with the companies who build them, as long as they turn a blind eye to ethical concerns.

A past example of public versus private science

This is not the first time big science has experienced a tension between public and private ownership. The Human Genome Project (Consortium 2001) was an international consortium of researchers, mostly from the US and UK. In 1998, Celera Genomics was founded in part to speed up sequencing. Celera used publicly-generated sequencing data along with its own sequencing to piece together a potentially more accurate human genome, since its input data was a superset of the public data. Users of Celera's data could search for matches to a query, but could not download the full draft sequence or train models on Celera data. Celera's researchers published a paper (Venter et al. 2001), which appeared in the same week as the HGP's (Consortium 2001). For the HGP, all data was publicly available; for Celera, the data was protected by a licensing agreement, and follow-up research was tightly controlled. Fortunately, Celera's advantage over the public-sector project soon eroded. Developers needed sequencing information that Celera did not release, so most researchers analyzed the public data.

Does it matter when scientific data sets are privately held, despite deriving from the work of the world? We argue that it does, in particular for transparency. As people highlight

ethical troubles with privately-held LLMs, all of their work is done in the proprietary space of the companies, and the companies need not respond accountably.

Literature review

ICCC is the largest international conference devoted to computational creativity. Did we predict any of the current issues caused by transformers? Did we see the coming storm?

We felt that, overall, ICCC researchers did not predict the current state of affairs. As a test, we conducted a literature review of ICCC papers between 2017-2022, *i.e.*, since the original transformer paper (Vaswani et al. 2017).

Framework of the literature review

We studied papers about text generation or media generation based on text, where transformers have caused the most disruption; papers about ethics and/or the nature of creativity; and general CC reviews. Both authors used Covidence to screen each paper for relevance to these topics. We evaluated each paper on the following questions:

- Did the paper mention neural networks? Did it mention transformers? Was either topics the paper's main focus?
- Did the paper attempt to predict how its area of CC was going to develop in the future?
- Did the paper mention ethics? If so, did it mention any of the specific ethical issues that are the focus of this paper? What other ethical issues were discussed?

Results of the literature review

Figures 1 and 2 show our major findings:

- While neural networks have always been studied in CC, there was a sharp increase in their mention and use in 2020, more recently driven by the rise of transformers. By 2022, most reviewed papers mentioned neural networks, and 43% had neural networks as their central topic.
- Between 25% and 50% of papers studied discussed ethics, but most discussions were brief and concerned other ethical topics than the ones that we screened for, such as how to conceptualize machine ethics or promote social causes. Each of the specific ethical topics we screened for was discussed by a handful of authors at most.
- Exploitation of content moderators was never mentioned.
- We also counted each paper's references taken from arXiv; this ranged from 0 to 54%, with median 0% across all included categories and 2% for text generation papers, but 23% for papers whose primary topic was transformers and 24% for media generation from text prompts.

Discussion

The CC community is not unaware of the *technologically* disruptive potential of transformers; there has been a sharp increase in interest in their use. But this has not been accompanied by a similar increase in attention to their ethical problems.

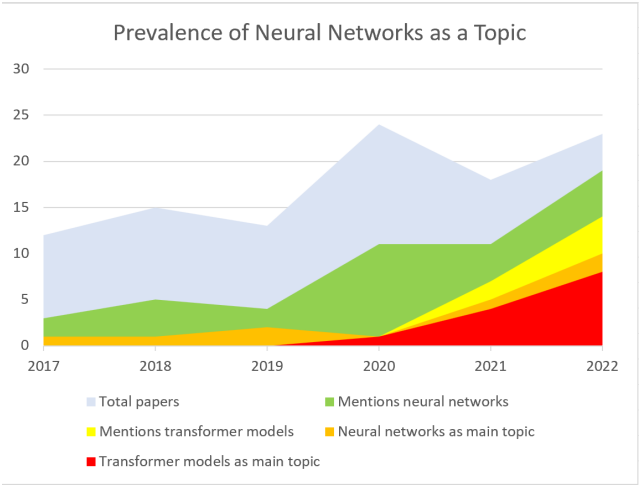

Figure 1: Papers in the screened categories that mentioned neural networks. The areas in this chart are overlapping, *not* stacked. Mentions of neural networks sharply increase beginning in 2020; those of transformer models do so in 2021-22.

This is not to say that attention to the ethical problems was missing entirely. Most of our topics were studied by a few researchers. For example, Bodily and Ventura (2020) discuss consequences for creative humans who feel surpassed by computers. Brown et al (2021) and Gordon et al (2022) analyze copyright issues. Loughran (2022), among others, discusses CC training set bias. Mirowski et al (2022), developing a CLIP-based collage system, incorporate concerns for human autonomy and copyright into their design. But these researchers are a minority. Their recommendations were not been taken up by the broader community, and certainly not at rates that matched the general dramatic increase in use of transformers. Nor did any predict the level of widespread social alarm that we currently see.

It is possible that researchers also raised ethical and social concerns in venues other than ICCC, or in informal discussions. Early signs show that there may be a greater focus on ethics at this year's conference. Nonetheless, the discrepancy between the use of transformers and the attention paid to ethics, in the papers that ICCC published before the explosion of public interest in this topic, is striking.

Our count of arXiv references is not a stand-in for paper quality; many papers full or arXiv references are thoughtful and inventive. (Indeed, this manuscript cites arXiv, news sources, and blogs!) However, the differences in this metric across categories suggest that in certain areas, the state of the art forces researchers to rely on non-peer-reviewed claims.

In light of the social disruption caused by transformers, some of the CC community's usual foci feel less urgent. Issues such as autonomy and embodiment are orthogonal to social impact; transformers cause these impacts regardless. Many impacts have nothing to do with the models' inner workings and everything to do with their ease of use at scale.

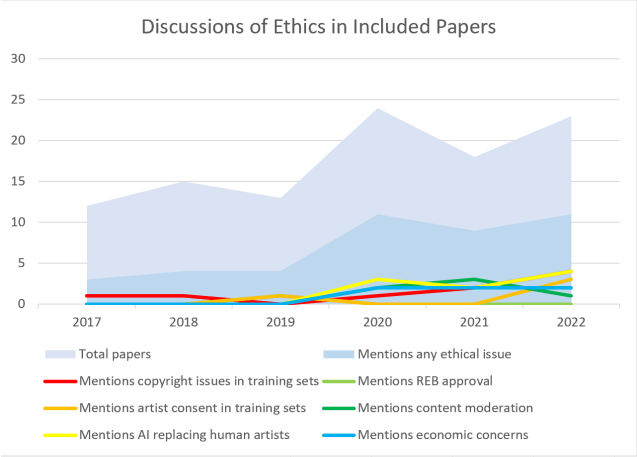

Figure 2: Number of papers in the screened categories that mentioned ethics. The areas and lines are overlapping, *not* stacked. It is a low, steady rate over time.

Conclusions and Recommendations

Transformers are the most quickly developing area in CC today. Yet, as we have seen, they produce harmful misinformation, exploit content moderators, harm the environment and are produced by opaque corporations that silence criticism. They make questionable use of copyright exceptions to gather training data for purposes that economically affect the artists on whose work they train, and they raise fears of human artists being replaced altogether. The hype and the alarm are overwhelming. What is a CC researcher to do?

One option is to study transformers ourselves, but if we do this, we must do so critically, with intense attention to their social and economic effects; we cannot become shills. ICCC is devoted to all aspects of computational creativity; social impacts are one such aspect, and they are exploding. To avail ourselves of the benefits and interesting uses of transformers, without proper and significant attention to these impacts, is a woeful imbalance.

As academics we lack direct power over corporations and governments, but we have a respected voice. We can rally against the excesses of corporate AI, point out its drawbacks, and suggest mitigations or alternatives. As "creative AI" becomes a public policy issue, more of us must focus on these roles.

Another option is to avoid transformers altogether. There are arguments in favor of this option; Bender et al (2021) discuss the "opportunity cost" of pouring scientific, financial, and material resources into transformers instead of using them to develop better alternatives. However, this option is not the easy out that it may appear. We must be aware that transformers are now the public face of creative AI - the first and sometimes only thing that a member of the public thinks of when they think of what we do. In this environment, if we develop a generative system that is not a transformer, it is up to us to clearly differentiate it from a transformer. We should think about how our own models can avoid the ethical pitfalls into which transformers have already fallen, and how we can make this difference clear to a frightened or skeptical audience.

At the least, we must be aware of the social effects of our research. Beyond writing about ethics in theory, we must incorporate ethics into our process, for example by adopting the recommendations of Bender et al (2021): thoroughly document training datasets, identify stakeholders at risk, and re-align research goals around a system's sociotechnical role.

We have an advantage in our emphasis on Process and Press, not merely Product (Jordanous 2016). We should take care not to lose this advantage. The social impact of CC systems has reached a crisis point, and is the most urgent issue in CC today; we should treat it accordingly.

Acknowledgments

The work of D.G.B. is supported by an NSERC Discovery Grant. We appreciate helpful conversations with Max Peeperkorn, Lauren Byl, and Maura Grossman as well as the clarifying comments of our anonymous peer reviewers.

Author contributions

Author C.E.L. designed the literature review, outlined the paper, and drafted most of it. Author D.G.B. drafted the two sections on public vs private science. Both authors collaborated closely on ideating the paper, performing the literature review, and revising the text.

References

Belanger, A. 2023. Getty sues Stability AI for copying 12m photos and imitating famous watermark. *Ars Technica*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proc. 2021 ACM Conf. on fairness, accountability, and transparency*, 610–623.

Bensinger, G. 2023. ChatGPT launches boom in AI-written e-books on Amazon. *Reuters*. Feb 21.

Birhane, A., and Raji, D. 2022. ChatGPT, Galactica, and the progress trap. *WIRED*. Dec 9.

Bodily, P. M., and Ventura, D. 2020. What happens when a computer joins the group? In *Proc. ICCC'20*, 41–48.

Brown, D.; Byl, L.; and Grossman, M. R. 2021. Are machine learning corpora "fair dealing" under Canadian law? In *Proc. ICCC'21*.

Clarke, N. 2023. Editor's desk: Written by a human. *Clarkesworld*. April.

Consortium, I. H. G. S. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.

Cropley, D. H.; Kaufman, J. C.; and Cropley, A. J. 2008. Malevolent creativity: A functional model of creativity in terrorism and crime. *Creativity Research Journal* 20(2):105–115.

Eloundou, T.; Manning, S.; Mishkin, P.; and Rock, D. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Farhi, P. 2023. A news site used AI to write articles. it was a journalistic disaster. *The Washington Post*. Jan 17.

Gordon, S.; Mahari, R.; Mishra, M.; and Epstein, Z. 2022. Co-creation and ownership for AI radio. In *Proc. ICCC'22*.

Ippolito, D.; Yuan, A.; Coenen, A.; and Burnam, S. 2022. Creative writing with an AI-powered writing assistant: Perspectives from professional writers. *arXiv preprint arXiv:2211.05030*.

Jordanous, A. 2016. Four PPPPerspectives on computational creativity in theory and in practice. *Connection Science* 28(2):194–216.

Lemley, M. A., and Casey, B. 2020. Fair learning. *Texas Law Review* 99:743.

Loughran, R. 2022. Bias and creativity. In *Proc. ICCC'22*.

Mirowski, P.; Banarse, D.; Malinowski, M.; Osindero, S.; and Fernando, C. 2022. Clip-clop: Clip-guided collage and photomontage. In *Proc. ICCC'22*.

Niven, T., and Kao, H.-Y. 2019. Probing neural network comprehension of natural language arguments. In *Proc. ACL*, 4658–4664.

OpenAI. 2023. GPT-4. Mar 14.

Ouyang, L.; Wu, J.; Jiang, X.; et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS* 35:27730–27744.

Perrigo, B. 2023. Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. *TIME*. Jan 18.

Scalzi, J. 2023. OMG is the AI coming for my job?!????!??!!!?!?!? *Whatever*. Feb 23.

Schaub, M. 2022. Cover of Paolini book may contain AI-created image. *Kirkus Reviews*. Dec 16.

Schiffer, Z., and Newton, C. 2023. Microsoft lays off team that taught employees how to make AI tools responsibly. *The Verge*. Mar 13.

Schiffer, Z. 2021. Google fires second AI ethics researcher following internal investigation. *The Verge*. Feb 19.

Shah, S. 2023. The writers srike is taking a stand on AI. *Time*. May 4.

Sobel, B. L. 2017. Artificial intelligence's fair use crisis. *Columbia Journal of Law & the Arts* 41:45.

Sweney, M. 2023. Mirror and Express owner publishes first articles written using AI. *The Guardian*. Mar 7.

Vaswani, A.; Shazeer, N.; Parmar, N.; et al. 2017. Attention is all you need. *NeurIPS* 30.

Venter, J. C.; Adams, M. D.; Myers, E. W.; et al. 2001. The sequence of the human genome. *Science* 291:1304–1351.

Wolf, M. J.; Miller, K.; and Grodzinsky, F. S. 2017. Why we should have seen that coming: comments on Microsoft's Tay "experiment," and wider implications. *ACM SIGCAS* 47(3):54–64.

Xiang, C. 2023a. 'Horribly Unethical': Startup experimented on suicidal teens on social media with chatbot. *Vice*. Mar 7.

Xiang, C. 2023b. Legendary sci-fi magazine halts submissions amid deluge of AI-written stories. *Vice*. Feb 21.

Zirpoli, C. T. 2023. Generative artificial intelligence and copyright law. *Congressional Research Service*.