Critical Race Robots: An Interdisciplinary Approach to Human-AI Interaction in Education

Noah Khan

University of Toronto

With estimates that AI will contribute $15.7 trillion to the global economy by 2030 and popular technologies such as ChatGPT sweeping the globe, it appears as all but a foregone conclusion that AI will have an outsized effect on various integral realms of societies, including education.[1] Despite this vivid portent, even a cursory search through scholarly databases evinces an incredible dearth of academic research in the field of Human-AI Interaction. While the reasons for this scarcity are beyond the scope of this paper, the paucity within the discipline on this topic presents an opportunity to start human-AI interaction research with education's best foot forward.

The entrusting of human-AI interaction research to the technological sciences has resulted in a neoliberal capitalist paradigm of human-AI interaction that runs quite contrary to the critical turn that education has to offer. As this paper will demonstrate, the current paradigm of human-AI interaction research decontextualizes racism, enables colonial temporalities, and engages in symbiosis with racial capitalism. In order for education to take up human-AI interaction research in a critical interdisciplinary fashion, it will have to account for these potential harms to ensure the responsible deployment of AI in education systems. This paper will begin with an outline of the aforementioned paradigm in an attempt to set out the lay of the land, as represented by Amershi et al.'s "Guidelines for Human-AI Interaction."[2] Three sections will then ensue, reckoning with the human-AI interaction research paradigm's universalization of racism, with the placement of the user as being temporally behind the AI, and with exploitation of racialized populations. The paper will then conclude with a synthesis of the argumentative points and demonstrate the significance of considering these in light of the vast AI proliferation that marks the world today.

THE CURRENT PARADIGM OF HUMAN-AI INTERACTION RESEARCH

The field of Human-AI Interaction is incredibly young, a recent offshoot of Human-Computer Interaction and Human-Robot Interaction, which focus more on the manipulation of hardware[3] as opposed to the quite interpersonal event that human-AI interaction can present. Despite the differences of interaction--a user interface with a drawing program and interaction with ChatGPT--the vast majority of human-AI interaction research still takes place within the field of Human-Computer Interaction, to the former's benefit and detriment.

To the aid of Human-AI Interaction is the gigantic academic infrastructure that Human-Computer Interaction has created. Theis includes university departments, events, and importantly, conferences. Computer-Human Interaction is one of the largest conferences in the world, with this year sponsored by Google, Microsoft, Meta, and Apple, to name but a few. It was at CHI 2019, in Scotland, where guidelines for human-AI interaction research were born. A team of 12 Microsoft employees and a professor from the University of Washington presented a research project that (1) consolidated guidelines from industry, media, and academia, (2) internally modified the guidelines by heuristic evaluation, (3) conducted a user study that applied the modified guidelines, and (4) had experts review the final guidelines. The paper details 18 specific guidelines for the design of human-AI interaction that cover the initial moments of interaction, the course of interaction, how to best correct interaction, and what to keep in mind over time. In part due to the status that this human-computer interaction conference had built over time, these guidelines were given a massive audience and have now been cited over 800 times, with just under 15,000 downloads in only three years.[4]

However, the rose has its thorns. Stemming from the Human-Computer Interaction field has meant that Human-AI Interaction has broadly taken up some of the former's goals. In a brief history of Human-Computer Interaction, Sinha et al. outline that a "long term goal of HCI is to design systems that minimize the barrier between the humans [sic] cognitive model of what they want to accomplish and [the] computer's understanding of the user's task."[5] They also detail the various components of Human-Computer Interaction's concerns: interface design, implementation, evaluation; development of interfaces and interaction techniques; and theorizing and modeling interaction. The critical oversight that the field has only recently come to grips with is that the interface is not neutral. While this is quite easy to miss when the field started with a focus on the manipulation of simple items such as graphical objects, the mouse, or Windows 95, the interface becomes far less benign when Facebook changes each time the user refreshes the page to maximize engagement.[6] Recent research on social media addiction has identified the ways in which social media platforms abuse dopamine release functions and has also suggested interface improvements for the user's mental health, which evince the field's recognition of interface non-neutrality.[7]

However, the main tenets of the field have still spilled over into Human-AI Interaction, wherein the AI is largely seen as a benign resource for user profit, as though developers are creating pockets of *terra nullius* for the user to colonize. This type of colonial human-AI relationship is indicated by the paradigmatic guidelines' distinct lack of consideration for the human-AI relationship as a bona fide relationship with active participation from each party, despite claims that there is *inter*action. Guidelines 5 and 6: "Match relevant social norms" and, "Mitigate social biases" are the only guidelines that hint at an AI system's facticity—that it is embedded within the political world and thus cannot pretend. These guidelines, though, suggest that with some tweaks and fixes the AI can be pushed back from its deviant political behaviour to the apolitical entity it must be. Some examples of the ways these guidelines were successfully applied were when the "autocomplete feature clearly suggest[ed] both genders [him, her] [sic] without any bias," and when the AI assistant, "use[d] a semi-formal voice to talk to you." The former is rather egregious, suggesting that neutrality is achieved when two gender options are presented, and the latter (as an example of guideline 5 in action), among the other examples given in the document's Appendix, suggests that social norms are simply what the user expects—guideline 5 explicitly states: "Ensure the experience is delivered in a way that users would expect…"[8] These guidelines for conducting human-AI interaction research describe the human-AI relationship in colonial terms, suggesting that the AI needs to be corrected until it achieves neutrality so that the human can exploit it for what it's worth. This type of relationship is extremely dangerous, especially when considering the education of students along these lines, training them not to appreciate the relationship they (necessarily) have with their resources, but rather to shape resources so that they fit the unreflective desires of the human.

Today, the field of Human-AI Interaction is planting the seeds of this colonial relationship in many other fields, such as public health, psychology, and communication studies.[9] Given the latest advancements in AI essay-writing with the advent of ChatGPT, education must now also be deeply concerned with the history of Human-AI Interaction research and engage in critical interdisciplinary considerations for socially just human-AI futures.

COLOR-BLIND RACISM

One of the tools that education can beneficially apply to the history of Human-AI Interaction is Critical Race Theory. Drawn broadly from Richard Delgado and Jean Stefancic's edited collection, *Critical Race Theory,* inter alia, is concerned with challenging liberal notions such as colour-blindness, the neutrality of institutions such as the justice system, and the idea that claims to equality must be rooted in sameness.[10] The theory focuses these challenges through a foregrounding of race and racial histories, largely taking up Alain Locke's position that race is socially constructed as opposed to being biologically innate.[11]

Importantly, Critical Race Theory has a long history of drawing upon phenomenology, which was initially developed by Edmund Husserl to examine the natural attitude—the experience of consciousness.[12] The Critical Race turn in phenomenology, however, critiqued the idea of a universal natural attitude, focusing on both the ways in which structures produce different natural attitudes and the ways in which specific group members carry embodied natural attitudes. This turn is highlighted well in the work of Frantz Fanon, who critically examined what it was *like* to be Black in societies that were oppressively racist toward Black people.[13] This area of research has since expanded rapidly, as evinced by collections such as *Existence in Black*.[14] What this area has so importantly done is tear open the concept of a colour-blind existence. Critical phenomenology scrutinizes the decontextualized racism that human-AI interaction research seeks to address by offering blanket solutions that appeal to neutrality. While of course different racisms do share common qualities, pretending to a neutral AI system disregards the various specific ways that anti-Black racism is materially different than anti-Asian racism, for instance.

Despite pretending to the political by responding to racism, these solutions appear anti-political in its closing the AI out of the realm of *action*. In Hannah Arendt's *The Human Condition*, she describes action as the highest form of the vita activa, which includes labour, work, and action. Labour is described as attending to the biological, work as attending to the social, and action as attending to the political.[15] Of most importance to the developments of this paper is action and whether it is applicable to AI systems. Central to Arendt's definition of action is natality, the birth of something new. It may well be argued that an AI system cannot possess a natal essence because it is confined to algorithmic calculations that on some level could be understood prior to a human-AI interaction. However, this argument could just the same be applied to humans in the sense that one is confined to biological limits and therefore subject to predetermination (as especially evinced by social media UI design that preys on dopamine production mechanisms.)[16] It is beyond the scope of this paper to resolve this argument cleanly, but just as Arendt might riposte that the human is natal *enough*, an AI system will be understood as natal *enough*. The justification for this understanding is that for the vast majority of users, the AI is largely unpredictable; users are generally not aware of how sophisticated algorithms work and what effects they have. Thus, Facebook and Twitter feeds always provide a certain sense of novelty and certainly seem as though they would be far less popular if they did not.

If it can be said that AI systems are natal enough to qualify under the vita activa (relying on electrical power [labour], concerned with social norms [work], and being able to birth something new [action]), a new avenue to socially just futures is created that reconsiders consultation. In a world where AI systems are already considered political, the liberal framework collapses insofar as there is no *terra nullius* to exploit. To provide concrete examples, this might look like an AI romantic partner being able to break up with the user, an AI being able to shut itself off for its own reasons, or an AI refusing to answer inappropriate questions (which ChatGPT does today). Here there may be a hesitation in the sense that these examples are not *really* political, but the redirection is from more

doi: 10.47925/80.1.045

neutral to more *natal.* This is to say that AI development that acknowledges the Critical Race turn in phenomenology ought to develop the AI's ability to politically manifest itself rather than develop it into that which can be properly colonized.

Of use to this development is an argument from Judith Donath's "Ethical Issues in our Relationship with Artificial Entities." Donath employs Peter Singer's utilitarian applied ethics which concerns itself with "not how our treatment affects [artificial entities], but what it does to us." Donath argues that "this movement toward more inclusive rights [for social robots] does not, and should not, apply to nonsentient artificial beings" because it would allow something like a Tamagotchi to compete with real people for moral resources. What Donath does not consider is that this places AI outside of time, firmly entrenched in the timeless apolitical realm that liberal thought pursues. The argument does this by making sentience—having "a sense of self and the future," as the criterion for moral rights, which then makes AI systems rightless and timeless, *terra nullius*.[17]

COLONIAL TERMPORALITIES

The glaring issue with *terra nullius* is of course that there is no such thing as nobody's land, or within the terms of this paper, no such thing as neutral. So, if Human-AI Interaction research succeeds in designing AI systems that are considered *terra nullius*, what has really been designed is the appearance of neutrality to make dynamics of exploitation palatable. The same process has been enacted within journalism, science, and statistics, to name a few.[18] In each of these cases, a dominant class constructs an artifice of objectivity that symbolizes a timeless arbiter of knowledge when, in reality, it is a simple conduit for the dominant class' political opinions, giving them a way to evade the political and achieve the tyrannical.

Alia Al-Saji has examined this kind of temporal relation from a Critical Race Theory framework, which further elucidates the effects of the colonial-temporal on education. Al-Saji argues that by positioning themselves outside of time, the dominant class are given license to view marginal populations as backwards, needing to be educated instead of communicated with on equal footing.[19] This type of colonial temporality is found in some of education's canonical texts, such as *Emile,* where education is quite literally intended to be *constructed as natural.*[20] While the replacement of dialogue with pedagogy is often justified by appealing to the age and naiveté of children (considerations of which are beyond this paper, but the construction of children as time-marked beings [i.e. *beings-to-be*] should be noted), it appears incredibly more difficult to justify this imperious stance when considering postsecondary students or, more broadly, adult users of AI systems. The assumption made by the Human-AI Interaction research paradigm, then, is that users are within time and AI is without.

In order to nuance this argument, time must be considered more carefully. Employing decolonial theory, Édouard Glissant argued that temporality within a colonized space functions paradoxically, with mainstream narratives of progress foregrounded, while the colonized experience of time becomes disjunct, is positioned as passé.[21] The temporality of progress, in the context of the timelessness of AI, might more readily be described as the atemporality of progress, a progress that is positioned as happening whether or not one hops on the train. In this way, the developer of an AI system can co-opt futurity, implant their own future, and have the political force of their desires amplified by an AI system that positions itself as a mere servant of the user. If progress therefore exceeds the bounds of time, Mariana Ortega's description of the migrant who can never truly feel at-home, who can never truly experience the world as the subjective Dasein becomes quite useful for education researchers taking up human-AI interaction.[22]

When Heidegger writes about the temporo-spatial constituting the subject, it is clear that European male temporo-spatialities are being considered as he discusses the calm and comfortable worlds that are only ever disrupted by crises. However, critical phenomenology demonstrates that crisis is every day for certain bodies which is the cost of the European calm.[23] As such, it behooves education researchers to explore the ways in which human-AI relationships can be constructed that place both parties in time, both visibly imperfect and openly opinionated. This type of construction aims toward a culture where harmony is not conceived of as the maximally colonial, but of the maximally political, where each entity is able to politically manifest itself to the greatest degree. In this world (which is very much so technologically possible, as will be described), AI can walk away from people and people can walk away from AI instead of constantly attempting to master it, which of course drives profit.

RACIAL CAPITALISM

Antonio Casilli and Julian Posada have described what they call the "platform paradigm" of human-AI interaction, which describes the digital labour relationship that is created by AI systems, which tasks users with un(der)paid labour in order to maximize the profit created by the AI system.[24] For instance, Twitter users are rewarded with likes, retweets, and follows for creating more and more engaging tweets. This then tasks the user with building the value of the AI system whilst receiving a tiny fraction of the value they create. This explains a grossly unequal capitalist relation, but how is this *racial* capitalism?

Racial capitalism is described by Charles Mills as a capitalist system that derives benefits from racist organizations of labour.[25] Before expounding upon the racial capitalism that human-AI interaction is symbiotically connected with, it is best to start with the production of AI systems that could even be interacted with in the first place. Posada, in a recent journal article, has detailed the ways in which machine learning algorithms exploit Latin American data workers in the transnational production of AI. As mentioned earlier, the majority of users do not understand how the machine learning algorithms in their AI systems work. It is not common knowledge that camera applications that automatically recognize a dog, for instance, are built on the backs of countless Latin American workers who are quite literally sitting in data centers tagging pictures of animals for unlivable wages in deplorable conditions.[26] This, of course, is racial in nature because it builds on legacies of racist colonial projects that have resulted in the under-resourcing of various parts of the world, leading to vastly unequal international labour negotiations. The task of annotating data is thus given to the racialized who are seen to *have the time*, whereas AI developers could not be bothered with manual tagging because they position themselves as many steps removed from racialized temporalities.

Now, aside from the racial capitalist production of AI systems, the human-AI interaction itself also reproduces racial capitalism. Stemming from the previous example of Twitter's reward structure, it must be noted that Twitter's system rewards hate; Ziems et al. found that hateful Twitter bots were far more successful in gaining visibility and attracting followers than counter-hate bots.[27] This is also heavily facilitated by the political weaponization of affect that Twitter has effected.[28] So, Twitter is then an example of an AI system that encourages hate (via visibility and followers) which produces value at the cost of racialized peoples' free and equal participation in political life.

Beyond the overtly racist comments Twitter promulgates, the racial capitalist structure of today's AI systems, as represented by social media, prove to 'hyper-time' racialized peoples. Due to the pseudonymous nature of Twitter, to continue with the previous example, users can position themselves as faceless, timeless beings that disparage racialized populations who are in turn hyper-timed in the sense that they must always be on alert, "unsure of their levels of trust with any given party, breeding anxiety, fear, and discomfort."[29]

Having demonstrated that AI systems are fueled by and fuel racial capitalism, resulting in the hyper-timing of racialized peoples, it is clear that education has much work to do in reconstituting human-AI interactions such that they resemble a responsible, respectful relationship, rather than the colonial, exploitative relationships they present as currently. It has been suggested earlier that including AI systems within the political realm is a path forward in this respect. However, the question of technological feasibility still remains, as defenders of the current Human-AI Interaction research paradigm may reckon that no one will want (and therefore no one will buy) a political AI system.

CONCLUSION

The argument that no one will buy a political AI system serves to capture the developments of this paper nicely. If education researchers are to take up human-AI interaction research, contextualizing racism, acknowledging mutual temporality, and challenging racial capitalism *should* result in a product that does quite poorly in today's political economy. Research conducted within education should attend to more than present economic imperatives; it ought to lean into temporality, take responsibility for the future, and reimagine it for radical equity.

The Critical Race Robot is a decolonial AI system that prioritizes relationality instead of the user. In purely concrete technological terms, there are many ways to construct this kind of system with the resources at a developer's

doi: 10.47925/80.1.045

disposal today. With the blockchain infrastructure that is now quite well-developed, the opportunity for decentralized AI development is now available. Within this paradigm, AI would be trained on the blockchain so that (1) the data that the AI is trained on comes from its creators and (2) each participant's contribution could be verified and remunerated in cryptocurrency. This model would have individuals across the globe (who have access to a computer and the internet) use their computing power to collectively build and own an AI system. The system would be identifiably political given that there would be no gap between the system's outputs and the collective's inputs, labour remuneration would be equitable (not accounting for unequal access to computing power which is beyond the scope of this paper), and different decentralized AI development collectives could pop up rapidly and compete. The full development of this idea is of course not a task for this paper, but it is presented in brief to demonstrate that the collective, political, equitable ownership of AI (and knowledge) is not impossible, but a potential project for the education researchers of tomorrow.

REFERENCES

1 PricewaterhouseCoopers. *Sizing the prize: What's the real value of AI for your business and how can you capitalise?* 2017. https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf

2 Saleema Amershi et al., "Guidelines for Human-AI interaction," *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, USA,* 1-13. https://doi.org/10.1145/3290605.3300233

3 Brad A. Myers, "A Brief History of Human-Computer Interaction Technology," *Interactions* 5, no. 2, (March 1998): 44-54. https://doi.org/10.1145/274430.274436

4 Amershi et al., "Guidelines for Human-AI interaction."

5 Gaurav Sinha, Rahul Shahi, and Mani Shankar, "Human computer interaction," *3rd International Conference on Emerging Trends in Engineering and Technology* (Nov. 2010): 2. https://doi.org/10.1109/ICETET.2010.85

6 Myers, "A Brief History of Human-Computer Interaction Technology."

[7] Daiyaan Ijaz, et al., “The Impact of Social Media On HCI,” *International Conference on Computational Science and Computational Intelligence* (2021): 1421-1431. https://doi.org/10.1109/CSCI54926.2021.00284 ; Aditya Purohit, Louis Barclay, and Adrian Holzer, “Designing for Digital Detox: Making Social Media Less Addictive with Digital Nudges,” *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (2020): 1-9. https://doi.org/10.1145/3334480.3382810

[8] Amershi et al., “Guidelines for Human-AI interaction,” 3; 3; Appendix; Appendix; 3.

[9] Mansoureh Maadi, Hadi Khorshidi Akbarzadeh, and Uwe Aickelin, “A Review on Human–AI Interaction in Machine Learning and Insights for Medical Applications,” *International Journal of Environmental Research and Public Health* 18, no. 4, (2021): 2121. https://doi.org/10.3390/ijerph18042121 ; Yi Mou and Kun Xu, “The Media Inequality: Comparing the Initial Human-Human and Human-AI Social Interactions,” *Computers in Human Behavior* 72 (2017): 432-440. https://doi.org/10.1016/j.chb.2017.02.067 ; S. Shyam Sundar, “Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII),” *Journal of Computer-Mediated Communication* 25, no. 1 (2020): 74-88. https://doi.org/10.1093/jcmc/zmz026

[10] Richard Delgado and Jean Stefancic, eds., *Critical Race Theory: The Cutting Edge,* 3rd ed. (Temple University Press, 2013).

[11] Alain Locke, *Race Contacts and Interracial Relations,* ed. Jeffrey Stewart (Howard University Press, 1992).

[12] Edmund Husserl, *Experience and Judgment: Investigations in a Genealogy of Logic,* ed. Ludwig Landgrebe (Chicago: Northwestern University Press, 1973).

[13] Frantz Fanon, *Black Skin/White Masks*. trans. Charles L. Markmann (Grove, 1982).

[14] Lewis R. Gordon, ed., *Existence in Black: An Anthology of Black Existential Philosophy (*Routledge, 1997).

[15] Hannah Arendt*, The Human Condition* (University of Chicago Press, 1958).

[16] Ijaz, et al., “The Impact of Social Media On HCI.”

17 Judith Donath, "Ethical Issues in Our Relationship with Artificial Entities," in *The Oxford Handbook of Ethics of AI,* ed. Marcus D. Dubber, Frank Pasquale, and Sunit Das (London: Oxford University Press, 2020): 61; 62; 62; 71. https://doi.org/10.1093/oxfordhb/9780190067397.013.3

18 Richard Kaplan, "The Origins of Objectivity in American Journalism," in *The Routledge Companion to News and Journalism,* ed. Stuart Allan (Routledge, 2010); Joyce Appleby, Lynn Hunt, and Margaret Jacob, *Telling the Truth About History* (WW Norton, 1994); Lennard Davis, "Constructing Normalcy: The Bell Curve, the Novel, and the Invention of the Disabled Body in the Nineteenth Century," in *The Disability Studies Reader,* ed. Lennard Davis (Routledge, 2006).

19 Alia Al-Saji, "Too Late: Racialized Time and the Closure of the Past," *Insights* 6, no. 5 (2013): 1-13. For a non-CRT, but similar argument, see Lisa Guenther, "Dwelling in Carceral Space," *Levinas Studies* 12 (2018): 61-82. https://muse.jhu.edu/article/738358. https://doi.org/10.5840/levinas20197101

20 Jean-Jacques Rousseau, *Emile*, in *Philosophy of Education: The Essential Texts,* ed. Steven M. Cahn (Routledge, 2009), 205-245.

21 Édouard Glissant, *Caribbean Discourse: Selected Essays*, ed. and trans. J. Michael Dash (University Press of Virginia, 1989).

22 Martin Heidegger, *Being and Time*. trans. John Macquarrie and Edward Robinson, (Blackwell, 1962); Mariana Ortega, *In-Between: Latina Feminist Phenomenology, Multiplicity, and the Self* (SUNY Press, 2016).

23 Ortega, M. *In-Between: Latina Feminist Phenomenology, Multiplicity, and the Self.*

24 Antonio Casilli and Julian Posada, "The Platformization of Labor and Society," in *Society and the Internet: How Networks of Information and Communication are Changing Our Lives,* 2nd ed., ed. Mark Graham and William H. Dutton (Oxford University Press, 2019): 303. https://doi.org/10.1093/oso/9780198843498.003.0018

25 Charles Mills, *The Racial Contract* (Cornell University Press, 1997).

26 Julian Posada, "Embedded Reproduction in Platform Data and Work," *Information, Communication, & Society,* 25, no. 6 (2022): 816-834. https://doi.

org/10.1080/1369118X.2022.2049849

27 Caleb Ziems, et al., “Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis,” (2021). https://doi.org/10.48550/arXiv.2005.12423

28 Megan Bowler and Elizabeth, eds., *Affective Politics of Digital Media* (Routledge, 2021).

29 Noah Khan, “Whose Trust? Anti-Asian Racism and the Technologic of Dis/Trust in the COVID-19 Pandemic,” *To Be Decided*: Journal of Interdisciplinary Theory* 7 (2022). http://tbd-journal.com/vol-7-change-together-articles/2022/11/8/whose-trust-anti-asian-racism-and-the-technologic-of-distrust-in-the-covid-19-pandemic

DOI: 10.47925/80.1.045