Artificial General Intelligence in the Workforce: a Humanist Perspective

OpenAI is a private company that claims to actively develop Artificial General Intelligence (AGI) which they define as "highly autonomous systems that outperform humans at most economically valuable work" (OpenAI, 2025). OpenAI also states that their mission is to "ensure that artificial general intelligence…benefits all of humanity" (OpenAI, 2025). In this essay, we accept their definition of AGI and discuss the benefits of this technology from a digital humanist perspective, employing "Introduction to Digital Humanism" as our foundational text. We begin by introducing digital humanism as the analytical frame through which we examine current AI issues and the implications of OpenAI's conception of the future. We then explore the impacts of AGI on underrepresented groups and AGI's consequences for human systems. Throughout, we ask 2 fundamental questions: (1) Does OpenAI's conception of AGI lead to a better future that "benefits all of humanity"? (2) What do we lose as we increasingly integrate AI tools into our work?

Humanism is a philosophy rooted in the innate values of human agency, freedom, and responsibility. As such, it values human participation in systems and human authorship. This closely aligns with a core value of existentialism, that people form and act on their convictions. Existentialism places a negative value on "bad faith", the self-denial of freedom and acting against one's convictions. Humanism expands this to a value of human agency and responsibility–not only is one in control of their own actions, but they are responsible for the ethical consequences of those actions. Digital humanism extends the ideals of humanism to modern digital systems: since automated systems *can bear no responsibility* for their actions,

humans should be placed at “the center of the digital world” (Werthner et al. 2024). To clarify, digital humanism is not anti-modernist: instead of rejecting the use of digital technology, it advocates for a different style of development with humans at the center. For example, advertising algorithms and social media feeds seek to modify user behavior and their consumption, digital humanism rejects these structures since they remove agency from our lives. With these principles established we can start to address our core questions.

We begin by exploring the impacts of AGI on women and minority cultural groups, which are both historically underrepresented. In the chapter “Digital Transformation Through the Lens of Intersectional Gender Research Challenges and Needs for Action” Draude interrogates how AI algorithms “mirror social bias” (Werthner et al., p88). Since AI algorithms are trained on and look for patterns in human data, they subsequently learn society’s prejudices and reinforce them (Werthner et al., p89). This biased behaviour is demonstrated by the use of word-embeddings. Word embeddings are representations of words in a numerical format that allows computers to parse text and form semantic relationships between words in sentences. Words such as “nurse” are more strongly associated with feminine pronouns, whilst words like “philosopher” are more strongly associated with male pronouns. Consequently, when AI models utilize word embeddings, they employ sexist language, leading to biased outcomes. We are not arguing that sexism does not exist in current decision-making processes– the issue is that AI models are influencing decisions and their inherent sexism is systematically and ubiquitously reducing the social mobility and thus freedom of human stakeholders. Hence, AI models contravene humanism. One example is the use of AI models in hiring decisions which have been found to make misogynistic choices (BBC News, 2018). Such decisions in the workforce

influenced by AI are polluted by prejudice and amplify pre-existing biases. Draude continues to say that there are mathematical methods to debiase AI algorithms in this specific setting, but new partisan problems continue to arise which may not have clear-cut solutions. Recent studies have highlighted a new problem pertaining to minority culture groups: using AI in writing tasks leads to a homogenisation of thought which favors the West. According to Agarwal et al. (2023) "AI caused Indian participants to write more like Americans, thereby homogenizing writing toward Western styles and diminishing nuances that differentiate cultural expression." This is caused by the datasets being primarily – "51.3%" (Dodge et al., 2021) – of American origin and thus induce the model to express mainly American views. Moreover, Agarwal et al. explain that "Everyday products and services offered by big tech companies based in the West capitalize on and profit from the data extracted from people in non-Western settings to enrich the wealthy and powerful."

By considering these impacts of AGI on underrepresented groups, we can begin to answer our core questions. Firstly, when considering whether OpenAI's conception of AGI lead to a better future, there will evidently be losers from AGI, as the inherent bias picked up by the models leads to a loss of opportunity for women, which is a reduction of their freedom and thus is detrimental to them according to the tenets of humanism. Moreover, the beneficiaries will exclusively be in the West, where companies such as OpenAI profit from usage. Therefore, OpenAI's narrative of benefitting <u>all</u> of humanity falls short – the opportunity and monetary gains are limited to a select few. Secondly, regarding what we lose as we increasingly integrate AI tools into our work, it is evident there will be a loss of nuanced cultural communication since the dominance of one culture necessitates the silencing of others. As a result, there is a reduction in freedom of expression, which is humanistically inimical.

We can further analyze OpenAI's conception of AGI by widening our view from individual interactions with AI agents to their effects on human systems. By OpenAI's admission, their end goal is "highly autonomous" AGI agents that are able to replace "most" human labor. But what is their conceived scope of human labor? Current agents operate through responding to a user prompt, performing the work specified, and outputting a result, usually in the form of text. AI agents are prone to occasionally outputting incorrect results, which has led to the concept of prompt engineering: the idea that one can build skill in creating better prompts, thereby guiding AI models to be as effective as possible. Edward Lee coins the term "Digital Creationism" to describe the conception of labor embodied by AI models. He explains it as the view that "technology is the result of top-down intelligent design…[and] every technology is the outcome of a deliberate process, where every aspect of a design is the result of an intentional, human decision" (Werthner et al. 168). In this framework, humans can be abstracted as machines that are assigned tasks and produce labor, making the choice to replace human workers with AI models appear straightforward. However, this ignores significant elements of how we produce technology, which has material consequences for the things we create.

In her analysis, Sharp uses the software industry as a particular example of how people "sit within a community of designers, developers, users, and other stakeholders who contribute to creation in one way or another" rather than being siloed, individual elements (Werthner et al. 360). These informal communities can form across organizations as a result of shared software and tools, but also as a result of shared cultural backgrounds and identities. Developers are in turn influenced by the practices embodied by these groups (Werthner et al. 361). This isn't merely conjecture about company culture; the effect of human communities on production is directly measurable. As found by Lopez et al. (2022), decisions critical to the security of a

codebase were “not always made by developers and their teams but instead reflect the attitudes and priorities of companies and their clients” (Werthner et. al 362). This extends beyond the software world: Sharp cites an example of nurses in an infusion clinic improvising a repair to medication pumps, outside of any formal organizational process (Werthner et al. 361). Clearly, these interpersonal networks act in meaningful ways to affect the results of our labor through sharing common practices, information, methodologies, and solutions to problems. These important yet often intangible properties of human agency in systems are part of why digital humanism values their presence so highly. Notably, OpenAI’s goals for AI in the workforce lack any consideration of this fact.

AI agents don’t participate in networks like these: they are overall siloed and context-free, don’t make independent discoveries or developments, and don’t share information or practices between each other. Overall, the only way to steer AI agents is through their prompts. This limits them to a very digital creationist loop of receiving top-down instructions and executing them. Returning to the software development example, we can see this falls apart in real world scenarios. Take a team of autonomous agents tasked with building and maintaining a large codebase. Over time, security issues will arise for a number of reasons which often follow similar patterns, for example changing requirements and new features causing stress to old code. Addressing these issues long-term often involves changes to everyday processes that can’t be easily expressed by modifying an agent’s prompt, such as working to address untenable product requirements or subtly shifting the ongoing code style. We can conclude that there is a fundamental gap between the agentic future OpenAI is selling and how effective development works in reality. Not only should we deny the notion that workers are akin to (and replaceable by) machines that take in instructions and output labor, but we should also avoid seeing

ourselves as masters of these agents, capable of ordering them to execute our will and expecting our own agency to not be affected in the process.

As a counterpoint to the framework of digital creationism, Lee introduces the notion of technological development as co-evolution: “Facebook changes its users, who then change Facebook. The LLMs will change us. For software engineers, the tools we use, themselves earlier outcomes of software engineering, shape our thinking. These tools have more effect on the outcome than all of our deliberate decisions” (Werthner et al. 169). As such, Lee asserts that as we increasingly integrate AI agents into our workflows, we start to lose control over the direction of our organizations. We must question if this is worth it or not. The tech industry especially prizes the idea of the individual innovator, someone whose immaculate vision paired with the ability to implement it brings about a brighter future. How can this fantasy stand up to a future where we will supposedly give up the agency to implement our ideas, in favor of deferring to agents who fulfill our whims? And on an industry-wide scale, if the direction of technological development begins to be driven by factors out of our control, the concept of innovation begins to fail as a whole, and any good outcomes become hard to attribute to human effort. We’ve always valued human agency in organizations, but while previously most of this value has been focused at the top, digital humanism asserts that agency at *all* levels is crucial.

We have introduced digital humanism as a lens to scrutinise the implications of AGI and explored its potential impact on underrepresented groups and the development process in the workforce. When exploring the impacts on underrepresented groups, we showed that OpenAI’s conception of AGI will be harmful for women and minority culture groups as it infringes on their

freedoms. Furthermore, we've demonstrated that OpenAI's idea of the future likens humans to the AI agents designed to replace them–machines that take in assigned tasks and produce labor–and thus ignores significant parts of the labor humans do in the process of creation. Ultimately, we have found that the implementation of AGI will be deleterious and diametrically opposed to the beliefs of digital humanism. If OpenAI wants to accomplish its mission of building AGI that "benefits all humanity" it must pay greater attention to what is being lost with current AI systems and build an implementation that puts humans at the centre of their design.

Works Cited

"Our Structure | Openai." *OpenAI*, 5 May 2025, openai.com/our-structure/.

Werthner, H., et al., editors. *Introduction to Digital Humanism : A Textbook*. First edition 2024., Springer Nature Switzerland, 2024, https://doi.org/10.1007/978-3-031-45304-5.

Agarwal, Dhruv, et al. *AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances*. 2024, https://doi.org/10.48550/arxiv.2409.11360.

Turkle, Sherry. *Alone Together: Why We Expect More from Technology and Less from Each Other*. 1st ed., Basic Books, 2011, pp. xvii–xvii.

Jakesch, Maurice, et al. *Co-Writing with Opinionated Language Models Affect*, 2023, dl.acm.org/doi/pdf/10.1145/3544548.3581196.

Nass, Clifford, et al. “Are People Polite to Computers? Responses to Computer‐based Interviewing Systems1 - Nass - 1999 - Journal of Applied Social Psychology - Wiley Online Library.” *Are People Polite to Computers? Responses to Computer-Based Interviewing Systems*, onlinelibrary.wiley.com/doi/abs/10.1111/j.1559-1816.1999.tb00142.x

BBC News *Amazon scrapped ‘sexist AI’ tool*, 2018 https://www.bbc.com/news/technology-45809919

Jesse Dodge, , Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, Matt Gardner. ,2021 *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus*. https://arxiv.org/abs/2104.08758