Artificial Intelligence in Higher Education in South Africa: Some Ethical Considerations

Tanya de Villiers-Botha
Stellenbosch University
tdev@sun.ac.za

Abstract

There are calls from various sectors, including the popular press, industry, and academia, to incorporate artificial intelligence (AI)-based technologies in general, and large language models (LLMs) (such as ChatGPT and Gemini) in particular, into various spheres of the South African higher education sector. Nonetheless, the implementation of such technologies is not without ethical risks, notably those related to bias, unfairness, privacy violations, misinformation, lack of transparency, and threats to autonomy. This paper gives an overview of the more pertinent ethical concerns that may result from the deployment of various current AI technologies in the South African higher education context. It provides a broad overview of the relevant AI ethics literature and distills a few general AI-ethics principles that can serve as guidelines for the ethical development, adoption, and use of AI systems. Suggestions are made as to how these might be applied to mitigate the ethical concerns in the South African higher education context. Overall, it is argued that AI technologies should only be adopted if they offer demonstrable benefits to stakeholders affected by them and that care should be taken to ensure that any potential harms are adequately addressed.

Keywords: AI ethics; Higher Education; Large Language Models (LLMs); South Africa; AI ethics principles; Ethical AI

Introduction and background

The recent prominence of Large Language Models (LLMs), such as ChatGPT, has given fresh impetus to calls to integrate Artificial Intelligence (AI)-based technologies into various aspects of higher education in South Africa (Govender, 2023; Marwala, 2023). Rightly, it is pointed out that, going forward, AI-based technologies will likely feature even

more prominently in many spheres of our lives. AI technologies are seen as one of the key drivers of the "fourth industrial revolution" (World Economic Forum, 2016; Schwab, 2016) and there is concern that South Africans and Africans as a whole suffer a skills deficit that leaves them ill-prepared for the future job market (Mkansi & Landman, 2021; World Economic Forum, 2023). Strong claims also are made about the benefits that such technologies will hold for higher education along various dimensions, including teaching, learning, and administration (Waghid, Waghid, & Waghid, 2019; Cele, 2021; Evans, 2021; Sedola, Pescino & Greene, 2021; Marwala, 2023). All of these factors play into the call to realign university administration, teaching, learning, and curricula accordingly (Butler-Adam, 2018; Xing, Marwala, & Marwala, 2018; Obi, 2022; Nwosu, 2023). However, while AI technologies are certainly here to stay and while many of the benefits mentioned could very well result from their adoption, it is not a given that any or all of them will *necessarily* do so. Much will depend on how such technologies are developed, deployed, and used. This is a major focus area in the field of AI ethics, which explores the ethical risks posed by AI technologies, something which has not received sufficient attention in the literature on AI in the South African higher education context.

With the advent of LLMs and other generative AI systems it is also essential to take note that there are reasons to be sceptical about the capabilities of these systems, at least in their current iteration (Weidinger, et al., 2022). In addition, even in instances where there is a realistic chance that adopting LLM-based technologies will deliver on their promised benefits, there are, as with other AI-systems, important ethical concerns that need to be addressed in order to ensure i) that these benefits are not undermined by harmful side-effects, and ii) that such benefits are shared equitably among *all* stakeholders impacted by the technology. In the following, I give a brief overview of the most prominent ethical risks relating to the adoption and use of AI-based systems in general, and LLM-based systems in particular, as identified in recent AI-ethics literature. I then discuss five broad ethical principles that can serve as guidelines when assessing the ethical impact of the development, deployment, and use of AI technologies. I also consider how these principles might be applied in the context of incorporating AI-based technologies into the higher education sector in South Africa.

What is AI?

There is no precise, generally agreed upon definition of "artificial intelligence". Roughly, when speaking about AI-systems, people tend to have in mind something along the lines of

"computers/machines that display behaviour that can broadly be characterised as intelligent, with "intelligent" here referring to human-like capabilities.[1] In their seminal textbook, Russell and Norvig (2022) describe AI systems as "agents that do the right thing", where the "right thing" is defined as the objective that we give them. Hence, a standard view of AI systems is that they are machines that can act so as to meet "their" (i.e. human-given) objectives. Thus, AI-systems are computational systems that are able to act in an agent-like way so as to meet a given objective. In addition, "AI" is an umbrella term, used to refer to a set of related technologies that make use of computational systems to perform various tasks, including natural language processing, image recognition, speech recognition, data mining for prediction, and online content recommendation.[2]

In the higher education context, Miao, Holmes, Ronghuai, & Hui (2021) divide possible applications for AI into four categories: "(i) education management and delivery; (ii) learning and assessment; (iii) empowering teachers and enhancing teaching; and (iv) "lifelong learning." Examples of specific applications in this context include the automation of admissions and timetables, using "learning analytics" to identify students at risk of failure, intelligent tutoring systems, automated language teaching, automated writing evaluation, and automated discussion-forum monitoring, among others. When it comes to the student- and lecturer-facing applications listed here in particular, it should be emphasised that there is still a lack of robust evidence for their efficacy (Miao, Holmes, Ronghuai & Hui, 2021). Since much of the current discussion around AI in higher education revolves around the implications of adopting LLMs specifically, we also need a broad understanding of what these systems entail.

LLMs (sometime referred to as "generative AI") are computational models that have been trained on massive amounts of data to model the statistical properties of the language in their training data (Devlin, 2018; Howard & Rude, 2018; Peters, et al., 2018; Radford, 2018; Scao, 2022; Weidinger, et al., 2022; OECD, 2023b). These models can then be used to make probabilistic predictions relating to sequences of tokens (words, subsets of words, and other bits of text or pixels), which allows for text (or other outputs, such as images) to be

[1] Seeing that there is little agreement on what constitutes "human-like" abilities and on whether machines currently truly exhibit such abilities, it is more accurate to refer to "automated" rather than "intelligent" systems.

[2] A distinction is often made between "narrow" or weak AI and "strong" or artificial general intelligence (Searle 1980; Mitchell, 2019). Narrow AI refers to AI that can perform its specific task very well (even surpassing humans' abilities in that task), but which cannot be deployed to perform a different task. All currently existing AI-systems are narrow systems. Artificial general intelligence (AGI) refers to AI-systems than can perform a wide variety or all tasks that humans can perform, on a par with human abilities. We do not yet have AGI.

generated that mirrors the training data. Such outputs can be further refined through, inter alia, fine-tuning and reinforcement learning from human feedback. Currently, depending on their architecture and training, LLMs are able to create text, images, video, and audio. Examples include OpenAI's ChatGPT, Google's Gemini (formerly Bard), Mircosoft's Copilot (formerly Bing chat), and Midjourney. "AI" is now often used as shorthand to refer to generative-AI systems. In what follows, I will use "AI" to refer to artificial intelligence systems in general, and specify when referring to LLMs in particular.

AI ethics

The ethics of artificial intelligence, or AI ethics, involves assessing AI-systems in terms of the ethical risks (and benefits) they may hold. AI ethics is one of the newest branches of the field of applied ethics, a branch of moral philosophy that addresses topics of practical concern, including issues relating to the health sciences, medicine, the environment, business and the public sphere.[3] This might raise the question of why we need a field dedicated to ethical issues relating to AI, specifically. What differentiates AI-related issues from those pertaining to health, medicine, business, and so forth? The answer is twofold: i) AI-systems pose novel problems in as far as they sever *objectives* and the *agency* to carry out those objectives, and ii) for technical and social reasons, AI-systems have unprecedented reach, meaning that any harms that arise from their use can occur at scale (Floridi, 2023).

In as far as AI-systems sever agency and objectives, we now have machines that can make decisions that would previously have been made by humans. Examples include who to admit to university, what information to present to a researcher, or what grade to allocate a student. At least three potential problems arise here: firstly, the systems may, for various reasons, both technical and non-technical, be making these decisions (achieving "its objectives") on the wrong basis, leading to bias, unfairness, or other harms (Friedman & Nissenbaum, 1996; Angwin, Larson, Mattu, & Kirchner, 2016; Barocas & Selbst, 2016; Caliskan, Bryson, & Narayanan, 2017; Buolamwini & Gebru, 2018; Dastin, 2018; Noble, 2018; Barocas, Hardt, & Narayanan, 2019; Sheng, Natarajan, & Peng, 2019; Koenecke, et al., 2020). Secondly, some systems are potentially "black boxes" in that it is often difficult to determine the basis on which any given decision was made. A concern here is the potential lack of recourse if the wrong decision is made (Bryson, 2019; Floridi & Cowls, 2019;

[3] See Singer (1986).

Diakopoulos, 2020). At a more fundamental level lies the question of transparency—how do we determine whether a decision is flawed if we do not know on what basis it was made? Thirdly, even if systems meet their objectives flawlessly (i.e. reach decisions on the right basis) the very efficacy with which they do so, coupled with their extensive reach, could have unforeseen detrimental consequences. I will discuss these concerns in more detail below.

There are other ethical concerns that arise with the adoption and use of AI-systems that have less to do with the technology itself and more with social factors relevant to its adoption and use. Here we can think of issues such as the technological divide in our society and the potential loss of autonomy, for example, as well as the privacy-violating practices of some of the companies who provide these technologies (Floridi & Cowls, 2019). In what follows, I will touch on some of these as they relate to AI technologies in the higher education setting.

On a global level, there are three pitfalls that we should note when assessing the potential ethical (and other) impacts of the adoption of new technology. The first is technological determinism. This is the view that the widespread adoption of a particular technological innovation is inevitable. This assumption creates a sense of inexorableness and urgency that potentially scuppers any due consideration of the relevant ethical concerns (Bearman, Ryan, & Ajjawi, 2022). Often, a determinist narrative is enthusiastically pursued by the companies that develop and market such technologies, for obvious reasons. The second, and related, pitfall is lending too much credence to overhyped claims regarding the potential benefits of any given technology. It goes without saying that no technology is a silver bullet, capable, in and of itself, of solving all manner of social ills. It is also hardly ever the case that a new technology does not also introduce new complexities into our social, economic, and political systems, whatever its benefits. A final pitfall to avoid is conflating the values at stake. For example, a technology that promises "greater efficiency" may or may not be beneficial on the whole, depending on what it is that is being done more efficiently, how this is being achieved, and who gains from it. "Greater efficiency" does not necessarily translate into benefit for all or even any affected stakeholders.

The vastness of the topic means that all potential ethical concerns relating to the adoption of AI-based systems in the higher education environment cannot be addressed in detail here. Such concerns may also differ depending on the kinds of systems adopted and their applications. Hence, in what follows, I will confine myself to the most pertinent concerns relating to current AI-systems in general, with an eye to the kinds of use-cases for AI-based systems in higher education in the South African context, currently and in the near future.

These include automated decision systems (such as student admissions systems or student assessment systems of various kinds), LLMs and other generative systems, personalised tutors (although these may fall under generative systems), and monitoring/classifying technologies, such as facial recognition software, automated proctoring systems, and systems for sentiment analysis.

The aim is to provide an overview of the main ethical risks in adopting AI-systems to serve as a guideline for administrators and policy makers in the higher-education space in South Africa. The overall argument is that ethical risks need to be carefully considered and weighed up against the potential benefits of adopting such systems in various contexts. I want to stress that the argument is *not* that AI technologies should never be adopted in higher education. Instead, my argument is that the benefits of doing so depend on careful consideration of the capabilities and potential pitfalls of these technologies and of their suitability for particular use cases. Moreover, if the foreseen benefits of a given application can plausibly be said to outweigh the foreseen risks, steps should be taken to mitigate the risks at every point during the development, deployment, and use of the application. Care must also be taken to ensure that the benefits and harms that stem from such adoption are equitably shared by all those affected.

Some ethical concerns

It should be noted that the concerns I discuss here do not necessarily all apply equally to each type of AI-system; nevertheless, they all apply to one or more.

Bias

One of the foremost ethical concerns relating to the wide-scale implementation of AI-based technologies is that of bias. "Bias" here refers to the reflection of cultural stereotypes and/or the taking of actions that unfairly prejudice groups.[4] Hence, "biased" AI-systems in this instance are systems that "systematically and unfairly discriminate against certain individuals or groups of individuals in favour of others" (Friedman & Nissenbaum, 1996). "Unfair" discrimination refers to denying an opportunity or assigning an undesirable outcome

[4] This understanding of "bias" should be distinguished from *statistical bias*, a conception of "bias" more prevalent in the context of statistics and machine learning. To avoid confusion, Barocas et al. (2019) characterise bias in AI-systems as "demographic disparities in algorithmic systems that are objectionable for societal reasons". I will retain the concept of "bias" here, but note that it refers to such demographic disparities, not statistical error.

to an individual or group on grounds that are *unreasonable* or *inappropriate* (ibid). Note that this tendency also needs to be systematic for a system to be considered biased.[5]

Bias in AI-systems is a complex issue, with several possible sources and divergent possible remedies with varying rates of success. In their pioneering work on bias in computer-based systems, Friedman and Nissenbaum (1996) identify three broad sources of bias: i) pre-existing bias (i.e. individual and societal biases that become manifest in the system); ii) technical bias (i.e. biases that arises from technical constraints or technical considerations), and iii) emergent bias (i.e. bias that arises when a system is deployed in a context with real users). One of the most important insights to draw from their analysis is that bias need not be the result of intentional design. With the more recent emergence of "big data"[6] and AI-systems trained on this data through machine learning[7], this point becomes even more pertinent. Bias may creep in in the *measurement* phase, where a given state of the world is captured in a dataset; in the *machine-learning phase*, where the data is turned into a model that summarises patterns in the training data and makes generalisations; in the *action stage* where a given action is taken on the basis of the model's prediction; and even in the *feedback* that occurs from implementing this system (Barocas, et al., 2019). In my brief discussion, I will focus on the measurement and machine-learning phases.

As (Barocas, et al., 2019). point out, from the outset, subjective decisions and technical difficulties beset the measurement phase, not only because the world is messy (i.e. complex) but also because machine learning practitioners often need to define new or relevant categories when defining their "target variable", i.e. that which they are trying to predict, such as "likely academic success" in the case of a n automated university admissions system, for example. If "likely academic success" at university is construed as a function of one's academic abilities, and matric results are taken to be an accurate measure of such abilities, then these may be used to quantify "likely academic success". However, matric results in themselves may not be an accurate measure of a given candidate's true academic abilities, especially given the realities of the unequal schooling system in South

[5] Some margin of error in automated systems, as with any other system, is all but inevitable.
[6] "Big data" refers to the massive amounts of digital data available as a result of the internet, social media, and the use of billons of smartphones.
[7] "Machine learning" is one of the means of developing AI-systems, where computation is used to discover useful regularities in data (Bryson, 2019). Systems can then be built that exploit these regularities through categorising them, making predictions, or directly selecting actions. Examples include automated fraud detection, insurance pricing, and recommending online content for a user. Since the "learning" and resultant outputs ("decisions") take place autonomously from direct human intervention, the resultant systems are said to the artificially intelligent.

Africa. There are also other factors beyond academic abilities that determine future academic success. Already at this stage, choices are made about what is and is not relevant that can affect the predictions our AI-system will make.

In the machine-learning phase, when a predictive model is trained on data that reflects the messiness of the world, any disparities, distortions, and biases in the training data will be reflected—and possibly amplified—in the model. An example would be the gender stereotypes embedded in the text scraped from the web (Schiebinger, 2014; Caliskan, et al., 2017; Noble, 2018;), which can then surface in the output of LLMs (Bender, et al., 2021; UNESCO, IRCAI, 2024). Other examples include models that score resumes for programming jobs that discriminate against female applicants due to past biased hiring decisions (Dastin, 2018) and facial recognition software that has much worse accuracy for women of colour than for white men (Buolamwini, et al., 2018). Problematically, even when we are aware of such biases in the data and so try to withhold a demographic category such as gender from the training data, a number of other attributes ("proxies") in the data may still correlate with the withheld category, which the model could pick up on (Barocas & Selbst, 2016; Barocas, et al., 2019). In addition, people of demographic minority groups may be underrepresented in the data, or machine learning may work less well for minority groups, if members of majority and minority groups systematically differ in terms of the prediction task (Kearns & Roth, 2019). In our university admissions example, candidates from historically disadvantaged schools may have systematically lower averages for matric compared to candidates from more advantaged schools, despite being equally capable of academic success at university. If our admissions model is trained on data where candidates from historically disadvantaged schools are underrepresented, the model will likely be calibrated towards the higher averages of historically advantaged groups and will thus have higher errors rates (false negatives) for the already disadvantaged groups.

In the context of applications built on LLMs, such as ChatGPT, many of these problems of bias also surface (UNESCO, IRCAI, 2024). Any biases in the training data can be reflected in the model output, including sexualised depictions of women, gender biases in output and the generation of other pernicious stereotypes (Sheng, et al., 2019; Hutchinson, et al., 2020; Abid, et al., 2021; Nozza, et al., 2021). In addition, underrepresentation in the data of some and overrepresentation of others ultimately leads to the homogenisation of outputs, where dominant views in the training data are reflected and less common, or less *datafied* worldviews and opinions, are lost (Weidinger, et al., 2022). Many LLMs reflect an Anglophone worldview (Bommasani, et al., 2021) as these models are largely trained on

English text (Scao, 2022). This can also lead to disparities in performance, such as less facility in underrepresented languages and dialects (Koenecke, et al., 2020). In addition, models can *amplify* training data biases, thereby exacerbating what is already a pernicious problem (Zhao, et al., 2017; Wang, et al., 2019).

The takeaway here should be, firstly, that, far from being value-neutral computation-based systems, AI-systems can reflect and amplify biases in the societies in which they exist in complex ways. Secondly, biases can also arise as a result of the technicalities involved in building and training AI systems. It is thus unlikely that bias can be completely avoided or corrected for at model level. Hence, while developers have an obligation to attempt to mitigate biases in the systems they develop, users should be made aware of the potential for biased output. If researchers, lecturers, and students were to use LLMs, for example, they would need to know that the output likely represents only the most common or dominant view or group represented in the training data and that some of it may be problematic.

Fairness

As may be expected, the problem of bias in AI-systems can give rise to the problem of fairness in the use of AI-systems. One aspect of fairness has to do with whether individuals or groups affected by the decisions of AI-systems are affected in ways that are *justified*, i.e. whether they have not somehow benefitted or been harmed through decisions based on irrelevant factors. Fairness, at least in part, requires that people *merit* the prediction of a given AI-system. This seemingly straightforward requirement is subject to several complications, however, not least of which is determining what 'merit' entails in particular cases. And even where we have a clear sense of what merit entails, it is not always clear how fairness will be ensured. There is also the question of different conceptions of "fairness". In addition to merit-based outcomes, "fairness" also connotes something like "equity"—this may involve equity in performance but also equity in terms of the distribution of the benefits and burdens of AI-based technologies.

In terms of merit-based conceptions of fairness, we may characterise "fairness" as being violated when a classifier model gives different scores to otherwise-identical members of different demographic groups (Obermeyer, 2019). This assumption seems justified in an instance where a hiring algorithm systematically rejects the applications of women that are otherwise similar to those of male applicants (Dastin, 2018). Here, equally deserving candidates from different groups are not treated equally. However, as is evident from our university admissions algorithm example, treating otherwise identical candidates from

advantaged versus disadvantaged schools equally will potentially bar deserving candidates from tertiary education, which also seems unfair (Kearns & Roth, 2019). In both cases, the problem is that an AI-system is potentially reinforcing and even amplifying underlying societal inequities. Moreover, there is no obvious technical remedy, given the problem of proxies. When "fairness" is understood as involving "equity", we might require that different groups be treated differently in a given context, such as in our university admissions example above. Yet, these definitions can be mutually exclusive, depending on our conceptions of "merit" and "equitable treatment", and optimising a model for one of these may lead to unfair solutions on the other understanding of fairness (Kearns & Roth, 2019). Inevitably, optimising for fairness, on whichever conception, also tends to negatively impact accuracy (Kearns & Roth, 2019).

The above problems notwithstanding, proponents of machine learning argue that data-driven decision-making has the potential to be more transparent than human decision-making in that it forces us to clearly articulate our decision-making objectives and thus enables us to be explicit about the trade-offs between desiderata such as accuracy and fairness (Barocas, et al., 2019; Fry, 2019; Kearns & Roth, 2019). If anything, algorithms force us to be explicit about what we want to achieve with decision-making. Neither transparency nor fairness come for free, however. Deliberate and careful design is needed to ensure fair and transparent decisions. Moreover, human judgement and norms should still play an important role in outcomes, even if (part of) the process is automated.

A further problem relating to fairness is that prefect prediction is impossible, which means that mistakes are inevitable (Kearns & Roth, 2019). It is thus important to appropriately temper trust in the outcomes of automated processes. In lieu of doing away with automated decision making, we may need to take steps to ensure that such mistakes are equitably shared, meaning that a particular group should not bear the brunt of erroneous outcomes. Infamously, the COMPAS recidivism-prediction model in the United States of America was shown to have a higher false positive error rate for black defendants and higher false negative error rate for white defendants (Angwin, et al., 2016). While the workings of the algorithm remains proprietary, the reason for this discrepancy does not seem to be due to bias in the algorithm, but is an artifact of statistical analysis due to the differences in the base rates of the two groups involved (which in itself can be drawn back to various societal factors). Hence, to ameliorate the patently unfair outcomes of the model (i.e. it systematically harms black defendants by overestimating their recidivism rates, other things being equal), deliberate decisions may need to be taken on how to adjust the model in

order to reduce this discrepancy in outcomes and equalise its error rate. Moreover, such an adjustment would need to come at the cost of some accuracy (Kearns & Roth, 2019). Mechanisms also need to be put in place to detect and ameliorate mistakes.

When it comes to using LLMs in an educational context, an important fairness consideration is the potential disparity in performance for different social groups. Of particular concern is the possibility that groups who are already marginalised will be subject to harmful stereotyping and other social biases and/or exclusion (Buolamwini, et al., 2018; Bender, et al., 2021; Weidinger, et al., 2022; UNESCO IRCAI, 2024). This means that some students will potentially carry a heavier burden with regard to possible harms resulting from LLM use, while reaping less benefit than students who are not form the affected marginalised groups. Hence, we run the risk of unfairly distributing the risks and benefits of the technology and further perpetuating and even compounding existing social inequalities. Such inequalities may also be exacerbated by the fact that students may not have equal access to the technology, either due to a lack of hardware (e.g. laptops, internet connectivity, etc.) or the computational skills required to utilise the technology effectively. Moreover, some LLMs are currently behind a paywall, meaning that students with greater economic means will be able to benefit more from the technology than others.[8] As discussed above, LLMs also risk "locking in" values and views dominant in their training data as well as in their downstream tweaking and applications, potentially leading to the homogenisation of perspectives in those using them (Bommasani, et al., 2021).

In order to address some of these biases and fairness concerns raised here, higher education authorities and institutions should take steps to minimise the discrimination and exclusion that may come from adopting these systems. These include (UNESCO, 2023):

i) Identifying and assisting those who do not have access to the necessary hardware, internet connectivity, data, or digital skills needed to use the systems;
ii) making provision for students with disabilities or special needs who may face unique challenges in accessing these systems;
iii) developing criteria and systems for validating AI systems for biases and discrimination;

[8] For example, the ChatGPT version based on GPT-3.5 is free while the better-performing GPT-4 version is behind a paywall (Open AI, 2023a)

iv) building validation mechanisms to test whether systems are trained on data representative of diversity (including in terms of gender, disability, social and economic status, ethnic and cultural background, and geographic location); and

v) requiring from developers that systems are trained in multiple languages and that support multilingual use; requiring that developers put measures in place to mitigate against promoting dominant cultural norms

It should be kept in mind that such measures are not guaranteed to fully address the problem of disparate performance for different groups. If they are to be employed in the educational context, students, researchers, and other stakeholders need to be informed that such systems do not necessarily reflect the plurality of worldviews and values that exist in the world. Care will have to be taken to mitigate this limitation, possibly by making use of other sources of information and by educating users to maintain a critical perspective towards outputs. Moreover, if larger LLMs are fine-tuned with specific data for particular uses in the context of teaching and research, care must be taken that the data used is as representative of diverse views and values as possible (at least in as far as is appropriate within the confines of the particular subject matter).

Misinformation and other information harms

LLMs are prone to outputting false, misleading, made-up information or poor-quality information (Bender, et al., 2021; Bommasani, et al., 2021; Metzler, et al., 2021; Weidinger, et al., 2022; Floridi, 2023; Mialon, et al., 2023; UNESCO, 2023). According to one school of thought, this technology, in itself, will never overcome the problem of generating false information, since LLMs are unable to access or identify information in terms of which they can ground truth (Bender & Koller, 2020; Weidinger, et al., 2022; Lenat & Marcus, 2023). As Weidinger et al. (2022) assert LLMs "are trained to predict the likelihood of utterances. Yet, whether or not a sentence is likely does not reliably indicate whether the sentence is also correct." There is broad agreement that current LLM-based systems are prone to "hallucination" (i.e. making factual errors) and correcting for this with current methods is time-consuming, costly, difficult to scale, and by no means foolproof (Metzler, et al., 2021; Mialon, et al., 2023; UNESCO, 2023).

In particularly high-risk domains, such as medicine or law, there is, of course, the danger of material harm that may arise from users' acting on misinformation. But outputting misinformation on the basis of statistical prominence can also lead to the further marginalisation of minority opinion (where this happens to be true) in that majority opinion

(as represented in the data) is represented as "fact" (Weidinger, et al., 2022). An added risk in the context of factual inaccuracy is "algorithm bias" where the users of AI-systems are prone to overestimating the capabilities of the system and hence to not be critical of the output of these systems.

It goes without saying that, in the context of education, information resources that deliver factual inaccuracies are highly problematic. If used, it is imperative that students, lecturers, researchers and other stakeholders are made aware of the limitations of these systems and of the potential for algorithmic bias in order to remain appropriately critical towards their output. This shortcoming should also be given due consideration before LLM-based technologies are deployed in the educational setting. In most teaching and research contexts, there seems little value to be had from technologies that are not always factually accurate. It is also important that users develop and retain the skills to effectively search for information outside of these systems in order to assess the veracity of their outputs.

The accessibility and fluency of LLM-based systems makes it extremely easy for students to generate passable text and pass it off as their own. At the same time, it is next to impossible to accurately determine whether text has been generated by an AI (UNESCO, 2023). Universities will have to carefully rethink their assessment practices and take measures to accurately determine whether students are indeed the authors of written work. The most effective methods will probably entail some form of oral assessment to compliment any work not done in a controlled setting (e.g., in-person examinations). However, South African educators are already time constrained, and our current large classes will make this option infeasible, especially at undergraduate level. Measures will need to be put in place to support lecturers in this regard. It is clear that assessment cannot simply continue as in the past without running the risk of such assessments not reflecting students' knowledge, understanding, or abilities—be it in writing, research or cognition—in any meaningful way.

A related risk is the generation of content that is subject to copyright or consists in intellectual property, given that many large language models contain such material in their training data. Users will need to exercise care and ensure that content generated does not constitute plagiarism or copyright and intellectual property violations. This can prove difficult, however, given that current systems are unable to accurately identify the sources that form the basis of the content that they generate (UNESCO, 2023).

Transparency/accountability

As discussed above, even though, when implemented carefully, automation can help ensure some measure of consistency in decision-making, mistakes are all but inevitable.[9] For one thing, there are natural limits to prediction. For another, Barocas et al. (2019) note that the "typical" model deployed in practice may have accuracy of between 0.7 and 0.8. This is better than a random guess, but still leaves room for a substantial number of false negative and false positives. In addition, machine-learning-derived algorithms may develop a decision-making scheme which during training may *seem* to be equivalent that of the human scheme it is being trained on, but the system may, in fact, be reaching its decisions differently and may produce different error patterns. Such learned decision schemes may end up relying on criteria that we would find objectionable; yet, due to the nature of machine learning, this would be opaque to us (Burrell, 2016; Barocas, et al., 2019).[10] Finally, the system may simply be buggy. Examples of automated decisions that are patently erroneous or unfair are not hard to find (O'Neil, 2016; Obermeyer, 2019; De la Garza, 2020).

All of this means that some measure of human oversight is necessary to ensure fairness. Of course, this also means that people need to have insight into the system and the basis on which it makes decisions, which is complicated by the opacity that accompanies AI systems that rely on machine-learning ,where even an expert may not know what pattern was identified by the system and where the programme keeps on evolving (Coeckelbergh, 2020). Hence, especially when it comes to uses of prediction models that have consequential impacts on people's lives, to establish legitimacy, developers and deployers should be able to *justify* a given decision scheme in that they are able to explain how the chosen targets relate to goals, they need to validate the accuracy of the deployed system, and allow methods for recourse in case of mistakes (Barocas, et al., 2019). Yet, the extent to which this is possible and practicable remains in dispute (Coeckelbergh, 2020; Diakopoulos, 2020; Müller, 2023).

In the context of higher education, the issues of transparency and accountability become especially relevant where AI systems are used to assess student work and where researchers use AI-technologies to do research. In both instances, it is imperative that there remains a "human-in-the-loop", i.e. some mind of human oversight. It needs to be possible to determine on what basis assessment of student work was done and to ensure that such

[9] See Barocas et al. (2019) for a detailed discussion.
[10] An example here is the risk-scoring model used by the city of Rotterdam to rank people according to their risk for committing welfare fraud which produced biased outputs. It was found that the system seemed to be taking into account categories such as being a parent, a woman, young, not fluent in Dutch, or struggling to find work (Burgess, et al., 2023).

assessment is fair. In addition, there needs to be avenues for students to query specific assessments and the possibility of recourse, if it becomes apparent that a system has erred. In terms of research, it is imperative that researchers remain accountable for their research. This means that researchers need to ultimately take responsibility for the veracity and integrity of their research. They also need to disclose and document their use of LLMs, something which many journals and publishers already require (e.g. Cambridge University Press, 2023).

Privacy

Loss of privacy looms as a major ethical concern within the context of AI-based systems both because of the scale at which such systems are able to access, collect, and process private data and because of the increasing ease with which even ostensibly anonymized or non-personally identifiable data can be "de-anonymised" (Narayanan & Shmatikov, 2008; Kearns & Roth, 2019; Zuboff, 2019; Véliz, 2020). Consequently, granular surveillance of individuals and blanket surveillance of whole populations by various entities other than legally mandated law enforcement is now possible, including surveillance by other state agencies, businesses, and even individuals (Thompson & Warzel, 2019). In addition, the massive amounts of digital data that we generate when we use the internet and internet-connected devices (smartphones, laptops, smart watches, etc.) and services is traded between various entities (e.g. data brokers, companies, governments and the like) (Müller, 2023). In fact, the business model of most of the internet-based services that we make use of can be described as consisting of "surveillance capitalism" where our data is harvested in order to granularly target content at us, be it advertising, or user content such as videos and social media feeds, designed to keep us engaged on a particular platform in order to collect more of our data and target more advertising at us (Vold & Whittlestone, 2019; Zuboff, 2019). This data can also be sold on to anyone, including insurance companies, potential employers, and governments (Véliz, 2020). As the European Commission's High-Level Expert Group on AI put it (2019), AI "enables the ever more efficient identification of individual persons by both public and private entities."

Of particular ethical concern are facial recognition systems and other involuntary methods of identification using biometric data (e.g. sentiment analysis through assessing micro expressions, automatic voice detection, automated proctoring, etc.). Such technologies cannot be introduced into the educational setting without consent and, in some instances, the meaningful option to opt out without being prejudiced. Clarity is also needed as to who retains access to the data collected, the purposes it is used for, and the security of

the data. Finally, the option sometimes needs to exist for such data to be deleted upon request (sometimes referred to as the "right to be forgotten"), always keeping in mind that simply removing personal identifying information from a given data set does not guarantee its anonymity (Kearns & Roth, 2019).

LLM-based technologies also pose privacy hazards in that they can potentially leak private information present in their training data or entered into them via prompts (Weidinger, et al., 2022). Such models may also be used to infer sensitive information about individuals. Potentially, malicious actors could leverage LLM-integrated applications to attempt to surreptitiously extract users' data, among other security risks (Greshake, 2023). These risks may be exacerbated if the LLM-technology used in an educational context is provided by an external entity, such as a large corporation, and where users are required to create an account in order to use the products (e.g. ChatGPT). This leaves users open to being profiled for purposes such as content curation and online advertising, potentially leaving them vulnerable to manipulation and exploitation (UNESCO, 2023). Hence, students should not be compelled to make use of a commercial AI chatbot such as ChatGPT to complete a research assignment, for example.[11] Care also needs to be taken to ensure that all users of such technologies are made aware of privacy risks that may result. In addition, students, lecturers, and researchers need to have the option of opting out of using such systems if they have legitimate privacy concerns. Institutions should also vet potential AI service providers on their use and handling of any user data that they collect.

Autonomy

The collection of our personal data is so widespread because it ultimately gives companies and governments the power to "forecast and influence" our behaviour (Véliz, 2020). This is one area where autonomy becomes a central ethical concern. Autonomy refers to our right to make informed decisions regarding our own lives—i.e. to have power over our own lives— and is taken to be a central, fundamental human right (United Nations, 1948). The concern with digital technologies is that that they, along with insights gained from the analysis of our personal data, can give companies, governments, and others unprecedented power to influence our behaviour and manipulate us. This need not only take place in the form of targeted advertising but also through targeting online content at us that is specifically tailored to our individual susceptibilities and hence more likely to persuade us (Susser, 2019). As already mentioned, this can take the form of videos and other online

[11] An example might be an assignment where students are required to generate output on a topic with a commercial AI application and then critique it.

content that is likely to keep us engaged for longer, but as became clear with the Cambridge Analytica scandal, can also result in attempts to manipulate our voting and other political behaviour (Cadwalladr, et al., 2018; Pham, et al., 2022).

An aspect of autonomy that has become even more pertinent with the advent of ubiquitous platforms and, more recently, LLMs is what we can call "informational autonomy". At issue here is the fact that the information that is consumed online is mediated by recommendation algorithms, which help determine and/or rank which users see what content and thus carry the risks of amplification and distortion (Alfano, et al., 2020; Narayanan, 2023). Recommendation algorithms for online content on platforms such as YouTube "predict" which posts a user are most likely to engage with, given their behavioural data. The result is that the online content that we consume becomes highly personalised to ensure optimum engagement. Hence, people have less control over what they encounter online than they might suppose. This applies to social media, streaming platforms, and even search (to a greater or lesser extent) and includes educational content. The concern here is that various platforms and the handful of commercial companies that own them have extensive control over what people are able to access online. This poses a threat to autonomy in that users cannot make informed decisions if they have no knowledge of the content that they are *not* served with, and they have very little information about and control over the curation processes that mediate their access to the online world (de Villiers-Botha, 2022; Kiri Gunn, 2022; Nys & Engelen, 2022). This dynamic is likely to continue with LLM-based systems deployed by commercial entities and that may be deployed in educational settings (Bommasani, et al., 2021). Without some insight into how such LLM-based systems fit into the above dynamic and are influenced by commercial considerations, we cannot adequately assess the impact that such systems have on our ability to access relevant information and reach our educational goals.

The risk to autonomy that comes from LLM-based systems is not limited to such systems being deployed by companies whose commercial interests might not coincide with our own interests. There is potential risk from the systems themselves. For example, research suggests that using LLM-based writing assistants can affect users' views (Jakesch, et al., 2023). In this study, participants making use of an "opinionated" language model, i.e. one that preferentially produces a particular point of view, shifted their opinion on the topic under discussion to align more closely with "its view". A further threat is the homogenisation that such models bring, as discussed above. Because of their reach, the homogenous worldviews encoded in the dominant LLMs can potentially become default views, robbing

users of the possibility of exploring alternative viewpoints and hence meaningfully reflecting on their own and developing their analytic skills.

A final potential risk to autonomy to consider is the possibility that students may fail to acquire important skills if they become too reliant on LLM-based and/or other AI systems. A recent Organisation for Economic Cooperation and Development Report (OECD) (2023a) argues that recent developments in AI technology means that AI may soon outperform many humans in many literacy and numeracy skills. This does not necessarily mean that such systems will replace humans in jobs where such skills are required, as jobs generally require complex skillsets. Nevertheless, the OECD forecasts that this will almost certainly affect employment, especially for workers with lower proficiency in literacy and numeracy than machines, which make up a "considerable share of the [OECD member countries'] workforce" (OECD, 2023a, p. 100). They argue that the workers best placed to deal with automation will be those with "solid skills" in three key areas: literacy, numeracy, and problem-solving skills in technology-rich environments (i.e. digital skills). As we adapt education to the emergence of AI technologies, we need to make sure that we do not produce students who use such technologies as crutches and who do not develop the literacy, numeracy, and problem-solving skills that may allow them to compete with such technologies. Accordingly, educational institutions should not use AI technologies in ways that will deprive students of opportunities to develop their cognitive abilities and social skills. In essence, the use of technology should clearly contribute to learning and research in a way that would not be possible in the absence of the technology.

Legitimacy

Closely related to the phenomenon of technological determinism (i.e. the view that widespread adoption of a given technology is inevitable and that we should simply find a way to accommodate it) is a form of "ethics washing" where companies claim to be developing ethical or responsible AI through explicitly addressing particular ethical concerns such as "safety" and "security" when developing their products. This can serve to detract from a more general ethical question relating to legitimacy—i.e. the question of whether a given AI system *should* be implemented in the first place. It is not always obvious that implementing AI systems rather than not is ethically justifiable. An example here is surveillance technology (High-Level Expert Group on AI, 2019). In addressing the bias in facial recognition and related systems, we should not lose sight of the overarching question of the ethical legitimacy of such systems—should anyone have such unprecedented, unbridled ability to surveille the public?

An example in the context of AI-powered applications for the higher education sector might be the adoption of LLM-based tutors. While the prospect of the "personalised teaching" is attractive for many reasons, careful consideration needs to be given to whether the technology available is capable of delivering on this promise (UNESCO, 2023). As (Bommasani, et al., 2021) point out, at a minimum, such a model would need to i) have an "understanding" of the subject matter at hand, ii) need some capacity for determining what misconceptions a student might have about subject matter in order to address these effectively, and iii) would need to have some kind of appreciation of pedagogy—i.e. effective and appropriate teaching techniques for a given subject matter and cohort of students. These are highly complex capabilities that current automation technologies are unable to replicate.[12]

To the extent that currently available AI-based systems are able to automate aspects of teaching, scholars highlight the risks that need to be mitigated before their implementation, most notably risks to "professional authority, institutional accountability, and public policymaking in education" (Zeid, 2020). Zeid points out that using technology that allows for "outsourcing instruction, assessment, and credentialing functions to [edtech] companies, leads to outsourcing more fundamental decisions to them as well." Lecturers do not only draw on their own skills, expertise, and experience to make decisions regarding instruction in classroom settings, but they also create curricula, course content, lesson plans, textbooks, syllabi, assessments, and education standards. All of this needs to be in line with specific institutional policies and decision-making structures and with broader national policies and guidelines. International, subject-specific norms, standards, and best practice guidelines also inform all of the above functions. In the incorporation of educational technologies from external service providers, the designers of the technologies take on part or all of these functions. In addition, the way in which the technologies are designed and trained will have a significant impact on how and what students are taught and how they are evaluated. There is a tremendous amount of individual and institutional expertise that goes into all of these functions, and not all of them are obviously automatable, especially in a way that leaves room for professional judgement. Moreover, education takes place in highly specific contexts, each with unique circumstances and localised needs and values. The South African higher educational context, for example, is highly democratic, with educators and students having a say in institutional policy and practices. This is not easily reconcilable

[12] Arguably, computer scientists underestimate the expertise required to master all of these capacities (both task specialisation and domain specialisation).

with the unilateral decisions that determine how educational technology is designed and trained in corporate settings. In outsourcing teaching to technology, some pedagogical and policy decisions are also outsourced to the creators and vendors of the technology, without the necessary oversight and transparency. As Zeid (2020) states, "[w]ithout tools for greater transparency, decisions embedded in code shut students, parents, and educators out of this loop. Instead of more readily available classroom teachers or on-site administrators, corporate entities handle these important decisions." Hence, before educational technology is adopted, careful consideration should be given not only to its capability to perform particular functions, but also to how the responsibilities and duties that accompany those functions will be met.

It should also be emphasized, however, that the technology is not yet at the point where it can be usefully deployed in a generalised manner in an educational setting. Referring specifically to the context of educational applications of LLMs, the authors of the UNESCO (2023) guideline caution that

> before significant progress is possible, it is essential that efforts are put into refining foundation models [LLMs and other forms of generative AI] not only through adding subject knowledge and de-biasing, but also through adding knowledge about relevant learning methods, and how this can be reflected in the design of algorithms and models. The challenge is to determine the extent to which EdGPT models can go beyond subject knowledge to also target student-centred pedagogy and positive teacher-student interactions. The further challenge is to determine the extent to which learner and teacher data may ethically be collected and used in order to inform an EdGPT. Finally, there is also a need for robust research to ensure that EdGPT does not undermine student human rights nor disempowers teachers.

If the technology does develop to the point that significant parts of the teaching process can be automated, questions still need to be asked, such as how the efficacy of such technology will be established, where the data to train these systems will come from, the role of the vendors of these systems in the educational process, whether or not something important is lost in replacing the human element in teaching and also whether or not the adoption of such systems would not mainly be adopted as a cost-cutting measure, resulting in the well-off being educated by human teachers and the less well-off being taught by machine (i.e. humans for the rich and machines for the poor).

Five Ethical Principles for AI

The above does not necessarily address all current ethical concerns relating to adopting AI technologies in the higher educational context, let alone possible future concerns as the technology develops. Consequently, it may be useful to keep a set of general ethical principles in mind to guide future discussions. A plethora of AI-related principles and guidelines have recently emerged, which is much needed but can be overwhelming. Strikingly, AI ethicists (Floridi et al. 2018; Floridi, 2019) have pointed out the convergence of many of these guidelines with one another and with the well-established ethical principles proposed by Beauchamp (1979) in the fields of bioethics and the ethics of healthcare namely beneficence, non-maleficence, justice, and autonomy. Floridi et al. add the principle of *explicability*, to capture an additional ethical dimension posed by AI-systems that stems from the potential opacity of their workings. Hence, important ethical considerations relating to the development, adoption, and use of AI-systems can be summarized into the following set of guiding principles (adapted from Floridi, et al., 2018; High-Level Expert Group on AI, 2019; IEEE, 2019; OECD, 2019, 2023b; UNESCO, 2022).

Beneficence

A central consideration in the development, adoption, and use of AI technologies is beneficence. The technology should bring some kind of benefit. Before developing, adopting, and using given technologies stakeholders should ask, "What *good* will it bring about?", or "What benefits follow from this technology?" At the risk of sounding glib, if there is no discernable benefit, this may be a good indication that it should not be developed, adopted, etc. The "good" that accrues could apply to individuals, society at large, or even the environment. In the context of higher education, incorporating technology such as large language models or other generative-AI systems for teaching or research, for example, should ultimately benefit students, lecturers, and researchers, such as by helping them develop important skills or conduct effective and relevant research.

Non-maleficence

Bringing about an apparent good is not yet *sufficient* reason to create and adopt given technology. We also need to consider possible harms that may be brought about in the course of the development, adoption, and use of given technologies. The risks discussed above highlight the kinds of harms that need to be considered. For example, we need to consider whether reliance on large language models potentially undermines important cognitive skills or whether students from marginalised communities will be even more

marginalised through underrepresentation or lack of access. Ultimately, possible harms need to be outweighed by the foreseen benefits. Steps also need to be taken to mitigate foreseen harms at every point in the lifecycle of technologies. Included here would be things like technical robustness and safety, including ensuring that the system is secure, resilient to attack, accurate, representative, reliable, accessible, auditable, assures privacy, and avoids unfair bias.

Autonomy and other rights

When assessing the foreseen benefits and harms of a given technology, it is not enough to argue that the benefits somehow outweigh the harms on aggregate. We also need to determine whether some foreseeable harms are categorically unacceptable. In other words, we need to determine whether some harms cannot be "outweighed" by the benefits. For example, basic human rights, such as the right to life, liberty, and security of person, are taken to be inviolable—they may not be infringed on the basis that this will bring about a "greater good" for others or society at large. An example we touched on in the AI-context is human autonomy. It is generally thought that it is a fundamental human right to be free to make important decisions about one's own life—to act as an agent (High-Level Expert Group on AI, 2019). This right is often also seen as encompassing a variety of other human rights, including the right to dignity, freedom of expression, the right to a private life and privacy, and freedom of movement (United Nations, 1948). Such rights are, of course, also enshrined in the South African constitution (Constitution of South Africa, 1996).

With AI, as discussed in the introduction, we have *artificial agents* to which we potentially cede some of our decision-making powers. As autonomous agents, we are entitled to *meaningfully choose* to cede some of these powers if we feel that this would be to our advantage; however, such a choice should be free and informed (i.e. not a prerequisite to accessing an essential service or buried in a lengthy and obscure list of "Terms and Conditions"). Other requirements in to ensure that stakeholders' autonomy is respected in the AI context include: transparency about when they are dealing with an AI-system (European Comission, 2023), securing human oversight over work processes in AI systems, restrictions on surveillance (High-Level Expert Group on AI, 2019), enhanced data protection when it comes to sensitive domains, such as health, family planning and care, employment, education, criminal justice, and personal finance, and having the option to opt out of using systems without being excessively adversely affected (Office of Science and Technology Policy, 2022). For example, students would need to be made aware if they were being

tutored by an AI-tutor. In such a scenario, their data would need the requisite level of protection and they may even need to be given the option of opting out of using the system, if they have legitimate privacy or other concerns.

An influential formulation according to which we can test whether or not we are respecting others' right to autonomy comes from Kant (1785) when he states that we should always treat people as ends in themselves and never as means to an end. In essence, this means making sure that we are allowing people to pursue their own, freely chosen ends and not simply using them as objects for our own benefit. Clearly, deceiving, manipulating, and exploiting people or covertly exposing them to risk falls foul of this principle.

Justice and fairness

Once it has been determined that a particular technology is beneficial and does not violate fundamental rights, it needs to be determined whether the benefits and harms are equitably distributed among stakeholders. Hence, unfair discrimination needs to be addressed and care must be taken not to perpetuate and amplify past injustices. Care should also be taken to ensure that some social groups are not disproportionately adversely affected by AI-systems.[13] Moreover, AI-systems should be as inclusive as possible, both in terms of representation and being accessible to different groups and populations. Asymmetries of information about such systems that could lead to asymmetries of power should be taken into consideration, e.g. between employers and workers, governments and citizens, or between businesses and consumers ((High-Level Expert Group on AI, 2019). Accountability should also be maintained so that redress against decisions made by AI systems is possible.

Explicability and transparency

In order to facilitate the principles set out above, measures need to be taken to mitigate the blackbox nature of automated systems. Those who develop and deploy AI-based systems need to document the design and development of the systems and provide accessible explanations of their functionalities, including highlighting their proper use cases and their limitations. In addition, the outcomes of such systems need to be interpretable, i.e. the meaning of a given output in relation to their designed purpose needs to be clear, or at

[13] In our large language model example, an assignment that requires using commercial AI-models such as ChatGPT, for example, may disadvantage students who are unable to pay for the premium version of the product or who do not have the hardware or skills to use the system. Moreover, since most commercial LLMs are trained in Anglophone data (see discussion above), their output would disproportionately reflect worldviews that do not necessarily apply to our context.

least explicable (High-Level Expert Group on AI, 2019; National Institute of Standards and Technology, 2023)

It should be noted that, due to their architecture and training, LLMs pose a challenge to the criterion of explicability. Currently, it is a challenge to understand what some of these models can do, why they output certain behaviours, and how they do so (Bommasani, et al., 2021). Even in task-specific neural network models, it is difficult to know what a model will do with particular input. Large language models turn out to be capable of a variety of tasks, some unforeseen and some likely as-yet undiscovered. Moreover, they tend to be trained on vast amounts of data—some scraped from the web—and developers are not always forthcoming regarding the provenance and scope of the training data (OpenAI, 2023b). This is at odds with the requirement that the provenance of training data should be maintained and the AI system's decisions should be attributable to subsets of its training data (National Institute of Standards and Technology, 2023) and strengthens the arguments that LLM-based technologies should not be deployed in high-risk settings, where it is necessary that outcomes or decisions be verifiable and/or contestable.[14]

In the educational setting in general, it is vital that possible use cases be tested, and deployment should occur on an evidence-based basis (UNESO, 2023). Applications should be validated by interdisciplinary experts before being adopted. After adoption, the characteristics, limitations, and potential shortcomings of the AI system need to be clearly communicated to users. Users should also always be made aware when they are interacting with an AI system. To prevent algorithmic bias and anthropomorphising, it should be clearly communicated that the system has no capacities of feeling or understanding.

[14] This obviously applies to a research setting where factual errors are unacceptable. If researchers were to make use of LLMs in their research in some capacity, it is imperative that they verify all generated output against trusted and legitimate sources.

Conclusion and recommendations

It is undeniable that there is currently a lot of hype around AI-technologies. Whereas the prospect of artificial intelligence holds exciting possibilities, it is important to keep in mind that these tools, as with other tools, can be both beneficial and harmful. Because of the complex nature of the technology, the many ways it could potentially be deployed and used, and its potential reach, care needs to be taken that possible harms are identified and mitigated. Using an ethical framework like the one above to do such an assessment is a start. Ultimately, expertise from many fields will be needed to ensure that any AI-system that is adopted is deployed in a way that benefits all who will be affected by it. At a policy level, it is recommended that the following suggestions are taken into account in the development of policies relating to the incorporation of AI into higher education:

- Adopt a human-centred, ethics-first approach to incorporating AI into management, teaching, learning, and assessment. The primary purpose of applying AI should be to enhance the capacities of students, lecturers, support staff, and other stakeholders in the field.

- Develop a master plan for using AI for higher education, drawing on a broad range of interdisciplinary expertise, including educators, researchers, computer scientists, engineers, and ethicists, to ensure that the adoption of such systems serves educational priorities and meets genuine educational needs in an ethical manner.

- Adopt an evidence-based approach. Make use of pilot testing, monitoring, and evaluation, and building up an evidence base to test for efficacy and to identify and mitigate any ethical harms before wide-scale adoption.

- Prioritise and enhance AI literacy for all stakeholders in the higher education sector, focusing not only on how to use AI-based technologies, but also on how they work and ethical concerns and responsibilities around their use. Foster the development of AI and general digital skills in all stakeholders in the higher education space.

- Foster local AI talent and local AI innovations for higher education to ensure the development of AI-systems appropriate to the local educational context.

References

Abid, A., Farooqi, M. & Zou, J. (2021). Persistent Anti-Muslim Bias in Large Language Models. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society,* 298-306

Alfano, M.; Fard, A. E.; Carter, J. A.; Clutton, P. & Klein, C. (2020). Technologically scaffolded atypical cognition: The case of YouTube's recommender system. *Synthese* (1-2): 1-24

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016). Machine Bias. *ProPublica*, 23 May 2016

Barocas, S. and Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, 103(4): 671-732

Barocas, S., Hardt, M. & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. [Online] Available at: https://fairmlbook.org [Accessed June 2023]

Bearman, M., Ryan, J., & Ajjawi, R. (2022). Discourses of artificial intelligence in higher education: a critical literature review. *Higher Education,* 86(2): 369-385

Beauchamp, T. L. (1979). *Principles of Biomedical Ethics*. New York: Oxford University Press

Bender, E. M. & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 5-10, 2020: 5185–5198

Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 3 March 2021, 610-623

Bizcommunity (2023). AI and higher education: What must tertiary institutions consider? *bizcommunity,* 22 February 2023
Avaliable at: https://www.bizcommunity.com/Article/196/499/236028.html. [Retrieved 10 March 2023]

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E. & Brynjolfsson, E. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*

Bradshaw, S., Neudert, L.M. & Howard, P. (2019). Government Responses to Malicious Use of Social Media, NATO StratCom Centre of Excellence, Riga, Working Paper

Bryson, J. J. (2019). The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation. In: M. Dubber, F. Pasquale & S. Das, eds. *The Oxford Handbook of Ethics of Artificial Intelligence*. Oxford: Oxford University Press, 3-25

Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: *Conference on Fairness, Accountability, and Transparency. Proceedings of Machine Learning Research*, 81: 1–15

Burgess, M., Schot, E. & Geiger, G. (2023). This Algorithm Could Ruin Your Life. *Wired*, 6 March, 2023

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society*, 3(1)

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *Guardian*, 17 March 2018

Caliskan, A., Bryson, J. J. & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183-186

Cambridge University Press, 2023. Authorship and contributorship. Available at: https://www.cambridge.org/core/services/authors/publishing-ethics/research-publishing-ethics-guidelines-for-journals/authorship-and-contributorship#ai-contributions-to-research-content [Accessed 30 September 2023]

Christie, J. (2020). The Post Office Horizon IT scandal and the presumption of the dependability of computer evidence. *Digital Evidence and Electronic Signature Law Review*, 17: 49-70

Coeckelbergh, M. (2020). *AI Ethics*. Cambridge, Massachusetts: MIT Press

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 11 October 2018

De la Garza, A. (2020). States' Automated Systems Are Trapping Citizens in Bureaucratic Nightmares With Their Lives on the Line. *Time*, 28 May 2020

De Villiers-Botha, T. (2022). Re-assessing Google as Epistemic Tool in the Age of Personalisation. In: *Proceedings of SACAIR2022 Conference, the 3rd Southern African Conference for Artificial Intelligence Research*, Online: SACAIR2022 Organising Committee, 323-337

Devlin, J., Chang, M., Lee, K, & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186

Diakopoulos, N. (2020). Transparency. In: M. D. Dubber, F. Pasquale & S. Das, eds. *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press, 197-213

European Commission (2023). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) 2021/0106 (COD)

Evans, N. (2021). Nurturing the African higher education teaching and learning ecology using data, information and knowledge from our physical, digital and biological learning spaces. In D. Ocholla, N. Evans, & J. Britz (Eds.), *Information knowledge and technology for development in Africa*, 85-102. Cape Town: AOSIS

Fayyad, U., 2001. The digital physics of data mining. *Communications of the ACM*, 44(3): 62–65

Floridi, L. (2023). AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models. *Philosophy and Technology*, 36(1): 1-7

Floridi, L. & Cowls, J. (2019). A united framework of five principles for AI in society. *Harvard Data Science Review*, 1(1)

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4): 689-707

Friedman, B. & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, 14(3): 330–347

Fry, H. (2019). *Hello World: Being Human in the Age of Algorithms*. New York: W. W. Norton and Company

Govender, R. (2023). The impact of artificial intelligence and the future of ChatGPT for mathematics teaching and learning in schools and higher education. *Pythagoras, 41*(1): a787

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T. & Fritz, M. (2023). More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models. *arXiv preprint arXiv:*2302.12173

High-Level Expert Group on AI (2019). *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:*1801.06146

Hutchinson, B. et al. (2020). Social Biases in NLP Models as Barriers for Persons with Disabilities. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5491-5501

IEEE (2019). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition., Online: IEEE.
Available at: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html [Accessed 6 October 2023]

Jakesch, M. et al. (2023). Co-Writing with Opinionated Language Models Affects Users' Views. *Association for Computing Machinery*, 1–15

Kant, I. (1785). *Groundwork for the metaphysics of morals*. New York: Oxford University Press. Thomas E. Hill & Arnulf Zweig, eds. (2002)

Kearns, M. & Roth, A. (2019). *The Ethical Algorithm The Science of Socially Aware Algorithm Design*. Oxford: Oxford University Press

Kiri Gunn, H. (2022). Is There a Duty to Disclose Epistemic Risk?. In: F. Jongepier & M. Klenk, eds. *The Philosophy of Online Manipulation*. London and New York: Routledge, 275-291

Koenecke, A., Nam, A.J., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J.R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America,* 117: 7684–7689

Lenat, D. & Marcus, G. (2023). Getting from Generative AI to Trustworthy AI: What LLMs might learn from Cyc. *arXiv preprint arXiv:*2308.04445

Marwala, T. (2023) Resistance is futile – South Africa must urgently adapt to the new age of artificial intelligence. *Daily Maverick*, 7 December 2023

Metzler, D., Tay, Y., Bahri, D. & Najork, M. (2021). Rethinking Search: Making Domain Experts out of Dilettantes. *ACM SIGIR Forum*, 55(1): 1–27

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A. & Grave, E. (2023). Augmented language models: a survey. *arXiv preprint arXiv:*2302.07842

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Penguin UK

Müller, Vincent C. (2023). Ethics of Artificial Intelligence and Robotics. *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition). Edward N. Zalta & Uri Nodelman (eds.)

Available at: https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/ [Accessed 14 March 2024]

Narayanan, A. & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets, *IEEE Symposium on Security and Privacy*, 111-125

Narayanan, A. (2023). *Understanding Social Media Recommendation Algorithms*. New York: Knight First Amendment Institute

National Institute of Standards and Technology. (2023). *AI Risk Management Framework*. Online: U.S. Department of Commerce.
Available at: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf [Accessed 30 September 2023]

Noble, S. U. (2018). *Algorithms of Oppression*. New York: New York University Press

Nozza, D., Bianchi, F. & Hovy, D. (2021). HONEST: Measuring Hurtful Sentence Completion in Language Models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2398-2406

Nwosu, L. I. (2023). Fourth Industrial Revolution tools to enhance the growth and development of teaching and learning in higher education institutions: A Systematic Literature Review in South Africa. *Research in Social Sciences and Technology, 8*(1): 51-62

Nys, T. & Engelen, B. (2022). Commercial Online Choice Architecture: When Roads Are Paved With Bad Intentions. In: F. Jongepier and M. Klenk, eds. *The Philosophy of Online Manipulation*. London: Routledge, 135-155

Obermeyer Z, Powers B, Vogeli C, & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447-453

OECD (2019). *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449

OECD (2023a). *Is Education Losing the Race with Technology?: AI's Progress in Maths and Reading.* Paris: Educational Research and Innovation, OECD Publishing

OECD, (2023b). AI language models: Technological, socio-economic and policy considerations. In: *OECD Digital Economy Papers*, 352. Paris: OECD Publishing.

Office of Science and Technology Policy (2022). *Blueprint for an AI Bill of Rights*. The White House https://www.whitehouse.gov/ostp/ai-bill-of-rights/

OpenAI (2023a). ChatGPT. [Online] Available at: https://chat.openai.com/ [Accessed 5 October 2023]

OpenAI (2023b). GPT-4 Technical Report. [Online] Available at https://doi.org/10.48550/arXiv.2303.08774 [Accessed 5 October 2023]

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Volume 1.* New Orleans, Louisiana: Association for Computational Linguistics, 2227–2237

Pham, A., Rubel, A. & Castro, C. (2022). Social media, emergent manipulation, and political legitimacy. In: F. Jongepier & M. Klenk, eds. *The Philosophy of Online Manipulation*. New York and London: Routledge, 353-369

Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI
Available at https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [Accessed 6 October 2023]

Russell, S. J. & Norvig, P. (2022). *Artificial Intelligence A Modern Approach*. Fourth ed. Harlow: Pearson Education Limited

Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M. & Tow, J., 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv*:2211.05100

Schiebinger, L., 2014. Scientific research must take gender into account. *Nature*, 507: 9

Schwab, K. (2016). *The fourth industrial revolution.* Cologny: World Economic Forum

Searle, J. R. 1980. *Minds, brains, and programs. Behavioral and Brain Sciences,* Vol. 3, No. 3, pp. 417–424

Sedola, S., Pescino, A. J., & Greene, T. (2021). *Blueprint: Artificial Intelligence for Africa.* Kigali: Smart Africa, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) and GFA Consulting

Sheng, E., Chang, K.W., Natarajan, P. & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv*:1909.01326

Singer, P., 1986. Introduction. In: P. Singer, ed. *Applied Ethics*. Oxford: Oxford University Press, 1-8

Susser, D. & Roessler, B. & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2)

Thompson, S. A. & Warzel, C. (2019). Twelve Million Phones, One Dataset, Zero Privacy. *New York Times*, 19 December 2019

UNESCO (2022). *Recommendation on the Ethics of Artificial Intelligence*. Paris: United Nations Educational, Scientific and Cultural Organization

UNESCO (2023). *Guidance for generative AI in education and research*. Paris: United Nations Educational, Scientific and Cultural Organization

UNESCO, IRCAI (2024) *Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Models*. Paris: United Nations Educational, Scientific and Cultural Organization

United Nations (1948). Universal Declaration of Human Rights 217 A (III). Paris

Véliz, C. (2020). *Privacy Is Power: Why and How You Should Take Back Control of Your Data*. London: Penguin Random House

Vold, K. & Whittlestone, J. (2019). Privacy, Autonomy, and Personalised targeting: Rethinking How Personal Data is Used. In C. Véliz (Ed.), *Data, Privacy and the Individual*, Madrid: Centre for the Governance of Change

Waghid, Y., Waghid, Z., & Waghid, F. (2019). The fourth industrial revolution reconsidered : on advancing cosmopolitan education. *South African Journal of Higher Education, 36*(3): 1-9

Wang, T., Zhao, J., Yatskar, M., Chang, K.W. & Ordonez, V., 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5310-5319

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P., Mellor, J.F., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S.M., Kenton, Z., Hawkins, W.T., Stepleton, T., Birhane, A., Hendricks, L.A., Rimell, L., Isaac, W.S., Haas, J., Legassick, S., Irving, G., and Gabriel, I. (2022). Taxonomy of Risks posed by Language Models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* New York, NY: Association for Computing Machinery, 214–229

World Economic Forum. (2016). *The Future of Jobs: Employment, skills and workforce strategy for the fourth industrial revolution.*
Available at: https://www.weforum.org/publications/the-future-of-jobs/: World Economic Forum [Accessed 2 March 2024]

World Economic Forum. (2023). *Future of Jobs Report.* Cologny: World Economic Forum

World of Work (2023). Artificial intelligence & Higher Ed in SA: What Universities Must Consider Now. *World of Work*, 13 February 2023
Available at: https://iieworldofwork.iie.ac.za/artifical-intelligence-higher-ed-in-sa-what-universities-must-consider-now [Accessed 1 March 2024]

Xing, B., Marwala, L., & Marwala, T. (2018). Adopt Fast, Adapt Quick: Adaptive Approaches in the South African Context. In N. W. Gleason (Ed.), *Higher Education in the Era of the Fourth Industrial Revolution.* Singapore: Palgrave Macmillan

Zeid, E., (2020). Robot teaching, Pedagogy, and Policy. In: Markus D. Dubber, Frank Pasquale, & Sunit Das (Eds.), *Oxford Handbook of Ethics of AI.* Oxford: Oxford University Press, 788-803

Zhao, J., Wang, T., Yatskar, M., Ordonez, V. & Chang, K.W., (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, 2979-2989

Zuboff, S., (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs

Acknowledgements

Competing interests

The author declares that she has no financial or personal relationships that may have inappropriately influenced her in writing this article.

Ethical considerations

This study followed all ethical standards for research without direct contact with human or animal subjects.

Funding information

This research work received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author profile

Dr Tanya de Villiers-Botha is a senior lecturer in the Department of Philosophy at Stellenbosch University. She is also head of the Unit for the Ethics of Technology in the Centre for Applied Ethics. Her research interests include: philosophy of mind and cognitive science, meta-ethics and the epistemology and ethics of AI.