Journal of the International Society for the Systems Sciences | 67th Meeting of the International Society for the Systems Sciences
J.M. Wilby, Editor.

DIGITAL PHRENOLOGY: ALGORITHMS AND ETHICS FROM A SYSTEMS PERSPECTIVE

Dimitrius Keykann
dimi.keykaan@nwu.ac.za
Dirk Snyman
Hennie Kruger
Anne Verhoef

Abstract
Decisions made by machine learning algorithms have societal impacts, both good and bad. This study acknowledges the great good that machine learning algorithms contribute to society while also pointing out that there are negative im-pacts such as bias, discrimination and misrepresentation of data, etc. Algorithms often use either irrelevant or incomplete data to make life-changing determinations about certain groups of society. The algorithmic use of data that does not repre-sent the true or actual ability of individuals or groups of people leads to digital phrenology. When algorithms are designed and tested outside the real world, it may not be obvious that digital phrenology may be caused. The ethical conse-quences of this problem can be mitigated if algorithms are not viewed as separate from the world it is eventually deployed into. Therefore, a more holistic systems view of algorithms, including their ethical consequences, should be taken. The purpose of this paper is to show that digital phrenology exists and that, once de-ployed, algorithms become part of a societal system where ethical values cannot be ignored. Various efforts and approaches to mitigate the ethical concerns related to unethical algorithmic behaviour have either done well in raising awareness or produced promising results. However, very little attention has been given to the explicit use of ethical theories as a possible approach to alleviating the problem. Therefore, two major ethical theories are proposed to assess the ethicality of hypothetical machine learning scenarios. A critical analysis of related literature and case studies is central in this study's inquiry method. Results show the existence of digital phrenology. It also reveals how viewing algorithms as a single compo-nent of a larger societal system can give a better perspective on the possible ethi-cal consequences that arise from their use. This research contends that the use of ethical theories can contribute to addressing digital phrenology by viewing algo-rithms as one of many components in a larger system.

Keywords
Digital phrenology, Algorithmic injustice, Machine learning bias, Ethical theories, Systems perspective.

1 | Introduction

Artificial intelligence (AI) and machine learning (ML) have been established as the *de facto* technology which pervades nearly every part of the modern industry and community. Due to the many benefits that come from its use, ML applications are integrated nearly in various areas of modern society. Often people are not even aware that they are making use of them or that their daily lives are influenced by these systems (Fatima & Pasha, 2017). The terms ML and AI are often used interchangeably or understood to be the same thing. In this study, ML is referred to in the context of it being a subset of AI. This implies that ML algorithms enable AI systems; hence ML and AI may be used interchangeably.

An extensive range of ML applications exist, and examples can be found in a diverse number of industries. ML models are applied in government departments such as education (Chen et al., 2020), the judicial system (Das et al., 2019), social welfare (Mladenov et al., 2020), crime prediction (Wheeler & Steenbeek, 2021), government managed medical data and facilities (Uddin et al., 2020), *etc*. A vast number of applications are also found in the private sector, and examples include credit risk assessments, (Bhatore et al., 2020) the hiring of employees (Mahmoud et al., 2019), cyber security (Alhayani et al., 2021), diagnosis of medical cases(Gupta et al., 2021), dentistry (Schwendicke et al., 2020), and economic and financial applications (Akbari et al., 2021).

Despite the widespread use and success of ML applications, there are also several recent studies that raise concerns about the correctness, fairness and ethical aspects of ML (Birhane, 2021; Morley et al., 2020; O'Neil, 2016; Zook et al., 2017). These studies point out that many problems are inherent to the use of ML algorithms. Issues range from results being biased, unfair, gender and racially unjust to problems associated with interpretability, transparency,

responsibility, reliability, *etc*. Researchers that are concerned with these types of issues in ML refer to the problems as *weapons of math destruction* (O'Neil, 2016), *algorithmic injustice* (Birhane, 2021), and *digital phrenology* (Hayes, 2020; O'Neil, 2016). It is, however, important to note that the problems highlighted by these researchers refer to ML models used in the social domain, *i.e*., in welfare, crime prediction, hiring of employees, credit risk assessments, medical care *etc*. For example, ML models used in manufacturing applications with no direct social connection to people or communities usually are not subjected to the same social concerns raised here.

It will be argued that:

- *Algorithmic injustice* (harm caused by unfair ML algorithm decisions that are unable to provide a rationale for decisions taken); and
- *Digital phrenology* (the decision-making process of ML algorithms, that is based on limited or incomplete information) can be addressed.

The argument is based on the premise that algorithms seem to be designed from a perspective mainly concerned with algorithmic function and not the environment they will function in. Introducing algorithms into the social domain, without considering other parts of the overarching system may pose fewer ethical concerns if deployed from a systems perspective. According to (Churchman, 1968) a system is made up of many components or subsystems that interact with one another to achieve a combination of greater goals. In its most basic form, a systems perspective is one which acknowledges that large systems rely on the interconnectedness of its smaller subsystems and that the subsystems cannot function in isolation (Arnold & Wade, 2015). In the context of this study, algorithmic systems are conceptualised as a subsystem in a more extensive system, which is modern society. Among many others, another subsystem in society is that of morality or ethics. Taking a systems perspective of ML algorithms involves acknowledging that it is only one subsystem of modern society which needs to interact with other subsystems in an ethical manner. A systems perspective is expected to contribute towards evaluating the ethicality of ML algorithms, minimizing unethical behaviour and harm in modern society. Thus, existing ethical theories will be introduced as possible approaches to minimise the reported problems in ML algorithms.

The aim of this paper is, therefore, to answer the following research question and sub question:

- *To what extent is digital phrenology exposed by a systems perspective?*
- *How can ethical theories be applied to assess algorithmic ethicality?*

The remainder of the paper is structured as follows: In Section 2, the reality and existence of digital phrenology and algorithmic injustice will be demonstrated by presenting an introductory discussion of practical and literature examples. This will begin to expose the effects of ML algorithms that are not approached from a systems perspective. A brief look at possible ethical solution strategies will follow in Section 3, while Section 4 will detail the proposed ethical theories supplemented by an illustration of how they may be applied to evaluate and improve ML ethicality. The paper will then be concluded with a brief reflection on the research questions and closing remarks in Section 5.

2 | Digital Phrenology: A Contributor to Algorithmic Injustice

In this section it will be demonstrated that algorithmic injustice is something that occurs in ML applications and that it is often caused by digital phrenology. It will be shown that various examples can be found in real life applications *i.e.* Makhlouf et al. (2021) and also in many literature discussions (Fuchs, 2018; Green & Hu, 2018). A few examples will be identified and briefly elaborated upon, to show that digital phrenology is a reality in modern society. Firstly, digital phrenology will be briefly introduced, followed by some real-world illustrative case study examples and then a look at a few discussions found in literature. Finally, ethical implications related to digital phrenology will then be considered before moving to the Section 3.

The fallacies perpetuated in digital phrenology will, in theory, contribute to algorithmic injustice. Digital phrenology is a concept which stems from the term phrenology ('*phren*' meaning mind and '*logos*' meaning knowledge). Phrenology is the discipline in which a person's brain function, personality and abilities are believed to be manifested in the cranial features of their head (Bittel, 2019; Schlag, 1997). ML, which is a subfield of AI, enables computers to learn without being explicitly programmed (Gullapalli & Brungi, 2015). Digital phrenology, which is analogue to phrenology, is present in many existing ML applications. The determinations made by many such applications have led to algorithmic injustice.

For the purpose of this study digital phrenology is defined as:

The decision-making process of ML algorithms which consider incomplete and sometimes irrelevant information when making determinations about a subject and causes harm to certain groups of society.

The definition can be explained in terms of the following hypothetical scenario. A responsible unmarried driver applies for car insurance. A ML algorithm is used to determine the driver's risk score and consequently the insurance premium. According to the algorithm, married people have a lower risk score than unmarried people. The outcome is that people who are not married are considered as high risk and consequently pay a higher insurance premium. The fact that the unmarried driver is responsible and may have been claim free, for longer than a married person, is not considered when making a decision. The conclusion "An unmarried driver is a higher risk than a driver who is married" has negative financial implications on unmarried responsible drivers. Additional examples of digital phrenology will be discussed in the remainder of this section

2.1 | Illustrative case studies

Several real-world ML applications have been identified as being either biased or unfair towards certain groups of society. Due to the many possible examples, only three prominent cases will be highlighted in this subsection.

Obermeyer et al. (2019) identified a healthcare risk-prediction algorithm used to determine whether patients will receive medical care of a specialised nature, was biased towards favouring majority groups over minorities. The algorithm was used to make decisions about over 200 million people in the United States. The study found that white American patients were much more favoured than their black counterparts due to an inherent bias in the way training data for the algorithm was constructed. Even though a 'patient's race was not explicitly used as a variable in the decision-making process, another variable (health care cost history) was used. This variable introduced an unintentional bias towards individuals of racial minorities. Historically, members of a lower social class, including black Americans, had limited access to health care due to political discrimination /finances, *etc.* This "non-use" of medical facilities led to lower overall associated costs, and therefore lower histories of expenditure, than people from a higher social class. When these cost histories are included in the training data to determine priority access to specially trained nursing staff for chronically ill patients, the decision is biased due to a perpetuated historical injustice. It is, therefore, argued that the decision is influenced by data that is irrelevant to the issue at hand and amounts to an instance of digital phrenology.

The following case study is that of Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), an ML application used in the United States criminal justice system to predict whether an incarcerated person will be a repeat offender. This application was typically used in bail application hearings or other similar situations. Li (2017) found that COMPAS predicted that black Americans are more likely to be repeat offenders than white Americans with similar profiles. The risk scores of 13 186 offenders assessed in Florida between 2013 and 2014 were used. The results suggested that there was a systematically higher risk predicted for black Americans as compared to white Americans, and also for males compared to females. COMPAS was therefore found to be gender and racially biased. Li (2017) points out that these kinds of biased decisions are based on the data that algorithms are trained with and, therefore, the algorithm can eternize inequalities because the training data contains these informative, but sometimes irrelevant, data. Not only can these inequalities be perpetuated by existing biases but also due to positive data being omitted in the training data sets.

Finally, Amazon.com, Inc. employed a machine learning tool based on natural language processing to automate parts of their hiring process to match candidates with positions. According to Kodiyan (2019), the algorithm was biased against female applicants. The cause of this bias was identified in the training data, collected over ten years prior to the implementation of the ML application. Traditionally it was common for the company to hire more male than female employees, which was not necessarily due to earlier prejudice but could be ascribed to the typically higher uptake of technical fields by men. The training data for their tool was then based on these hiring patterns for which the data was readily available. Due to statistical inferences made by the ML algorithm, a bias towards recommending male candidates was introduced simply because more men than women were hired in the past. Candidates were, therefore, not just screened based on their skillset but their ability was determined by irrelevant attributes, such as their gender.

The above examples indicate how the decision-making process of these ML algorithms are similar to that of phrenology. A common contributor in the examples is that of pre-existing bias being introduced into the training data sets of the algorithms as well as the omission of diversity in the data.

Real-world case studies present a good indication of algorithmic injustice, caused by digital phrenology. To further demonstrate its relevance and emphasise this problem, a few examples from literature will be highlighted in the following subsection.

2.2 | Examples from literature

The reality of digital phrenology found in real-world ML applications can be supplemented with descriptions from literature which expose and corroborate the fact that the problem exists and needs to be addressed. Birhane (2021) contends that ML algorithms that are used to predict social outcomes often allow historical, unjust and discriminatory patterns to be continued and therefore considers them inadequate for this purpose. The perpetuation of these patterns, and the harm they cause to already marginalized communities, is what she calls *algorithmic injustice*. This has come to the attention of ML designers, leading to the development of various techniques to address its potential harm. These techniques are, however, based on technical solutions instead of focussing on the adversely impacted communities. The answer to algorithmic injustice is, therefore, not necessarily just a technical one. More human-centric avenues, *e.g.*, ethics, should be considered.

The algorithmic injustice described by Birhane (2021) can be closely related to what O'Neil (2016) calls *weapons of math destruction* (WMD). She corroborates how algorithms may have elements of digital phrenology which leads to algorithmic injustice. Early in the 20th century, bankers would approve loans based on irrelevant aspects of an applicant, including reference to aspects of a non-financial nature such as church attendance, family connections, race and ethnicity, and the geographical origin of an applicant. This form of decision-making was based mainly on how *applicants of a certain archetype* are known for repaying loans instead of transparently considering more representative factors such as an 'individual's repayment history, income, expenses, *etc*. O'Neil (2016) further states that semi-automated lending systems, approving loans based on ML algorithms, are very similar to these 'bankers' approach. The only difference is that human deliberation has been removed from the equation, and decisions are now made based on mathematics and statistics. As a result, these algorithms are WMDs when decisions go awry. The algorithms, for instance, consider information such as an 'applicant's residential address or zip code as a deciding factor in whether the applicant will be able to make repayments or not. This is not different from considering how *applicants of a certain archetype* (residential information forming part of the archetype) have historically repaid their loans. The result of such a decision-making process enables the continuation of past injustices. Similar WMDs are found in models for employment and insurance. Applicants are often unaware that they have a right to refuse or argue the use of these WMDs in decision-making relating to them.

Ploug and Holm (2020) argue that patients have a right to refuse the use of ML algorithms for medical diagnosis and treatment planning due to bias and transparency problems related to such systems. They assert that the bias can be due to human prejudice present in the training data sets and/or biased objective functions at the core of AI systems. For instance, if the objective function is to maximize *health care outcomes* with *patient compliance* as a decision variable, then the patient age group with the highest compliance parameter will be given the best possible treatment. Thus, if high compliance is associated with a 60-70 'years' age group, then this group will receive preferential access to treatment. In effect, the system will discriminate against less compliant age groups and as a result, will not get the treatment they require.

Literature also shows that digital phrenology is indeed a reality and that the algorithmic injustice that is caused by it leads to harm to certain groups of society based on factors such as age (Stypińska, 2021), socioeconomic status (Juhn et al., 2022) and gender (Leavy, 2018). Literature on this problem is increasing due to its relevance in modern society and academia. See the following books, for example: *Artificial unintelligence* (Broussard, 2018) and *Algorithms of oppression* (Noble, 2018).

As argued in the introduction, algorithmic injustice causes harm and avoiding harm is often related to discussions of ethics (Dixon & Quirke, 2018). The ethical implications around algorithmic injustice and digital phrenology will be considered in the following subsection.

2.3 | Ethical implications that stem from digital phrenology

Given the aforementioned examples of the prevalence of digital phrenology, it can be argued that digital phrenology involves and can be mapped to ethical implications that need to be addressed or at least taken note of. This notion is corroborated by Mittelstadt et al. (2016), who contend that at least six ethical concerns can be conceptually mapped to the use of algorithms, how they process data to produce evidence and motivate actions. The six types of ethical concerns raised by algorithms have also been referred to and briefly elaborated on by Morley et al. (2020), as seen in the first two columns of Exhibit 1 below.

The objective of Exhibit 1 is to demonstrate the digital phrenology in ML algorithms, as well as the injustices caused by it, can be considered ethical concerns. In every second row of Exhibit 1, the ethical concerns (Morley et al., 2020) are highlighted, and the first column shows a brief rationale relating to the concern. The second column introduces practical remedial considerations to the ethical concerns related to ML algorithms. These practical considerations,

based on O'Neil (2016), intend to guide the design process of algorithms so that they are "less guilty" of digital phrenology. Each ethical concern is directly linked to a list of practical remedies, and as a result, each practical remedy is an ethical solution to an ethical concern related to ML algorithms. This research proposes that the contents of Exhibit 1 be extended and further investigated to allow the explicit use of ethical theories to assess, and possibly address digital phrenology.

Exhibit 1. Mapping ethical concerns to algorithms.

Rationale (adapted from (Morley et al., 2020))	**Practical remedial considerations** (based on (O'Neil, 2016))
Ethical concern: Inconclusive evidence	
Certain algorithmic and model conclusions are based on probabilities, and evidence produced by these techniques may be fallible.	People should refrain from *blindly accepting* algorithm results, treating *algorithms as a god* that cannot make mistakes. Admit that *algorithmic systems cannot do everything* and not believe in the myth that *algorithms do not have to provide evidence for their decisions*. In addition to the above-mentioned paradigm shifts, the implementation of *appeal processes* and *algorithm audits can also contribute to minimalizing* Inclusive evidence.
Ethical concern: Inscrutable evidence	
An algorithm or model needs to be understood to control and monitor the output produced. A lack of interpretability and transparency would lead to inscrutable evidence and irresponsible applications and decisions.	The solution to inscrutable evidence may lie in addressing the modellers of algorithms first. Similar to what medical doctors have in the form of a *Hippocratic oath*, data scientists and modellers of algorithms must pledge to something alike. It is important to remember that algorithms results cannot be blindly accepted as if the algorithms are gods who do not have to provide any evidence. It is also essential to have *diversity in modelling teams* – not just data scientists so that different perspectives may be considered.
Ethical concern: Misguided evidence	
Models and ML algorithms are trained on data. This implies that an algorithm's reliability depends on the reliability of the data used during the training process. Algorithms trained on irrelevant data will produce results that are misguided.	The *evaluation of inputs* – and considering whether they are ethical or not- when addressing this ethical concern. Another problem with the input is not checking that the *data collection is approved by the user/client*. Modellers should also not use data from one application for a different purpose. So, it is important to interrogate *the data collection process – not just the algorithm*. Therefore, algorithms should be *open to audits*.
Ethical concern: Unfair outcomes	
The outcome of an algorithm can be considered unfair or discriminatory based on how it treats a particular group or subgroups of people. For example, bank loan approval systems may favour certain clients based on historical and unjust patterns (O'Neil, 2016)	There must exist *an appeals proces*s for everyone. Appeals will make sense if models are *open and available for public scrutiny. O*therwise, outcomes may still be considered unfair – even after an appeals process. *Fairness must be put ahead of profit*, and the use of algorithms *should benefit society as a whole*. Because algorithms do not create themselves, modellers need to be addressed because *they (modellers) are key* and can *impose human values on systems.* Modellers must *measure the impact of an algorithmic decision,* and *government should play a powerful regulatory role*. Regulatory roles can ensure that choices or *outcomes are fundamentally moral, considering context and harm*.

Exhibit 1. Continued.

Rationale (adapted from (Morley et al., 2020))	**Practical remedial considerations** (based on (O'Neil, 2016))
Ethical concern: Transformative effects	
ML applications are integrated nearly everywhere in modern society, and often people are not even aware of it (Fatima & Pasha, 2017). The application of these algorithms has the potential to transform the world in new but unexpected ways, leading to actions that are directly informed by insights generated by the outcomes of the algorithms (Mittelstadt et al., 2016).	Since algorithms have a direct influence on society, the ability to *measure the impact of an algorithmic decision* is essential. *Choices or decisions should be fundamentally moral, taking context and harm into account*.
Ethical concern: Traceability	
Models and algorithms cannot be held accountable for their outcomes. Furthermore, no laws exist to hold their creators accountable. This makes it hard to trace algorithmic injustice to a point where someone can be held responsible for undesired outcomes. According to (Morley et al., 2020), the lack of traceability may also lead to issues with moral responsibility.	*Modellers are key* in moral responsibility and traceability of algorithms outcomes. Therefore, something like a *Hippocratic oath for modellers* can only be helpful. Modellers can *impose human values on systems*, and society can hold governments responsible if it does not regulate and establish legislation for algorithmic outcomes.

Exhibit 1 has highlighted how certain ML problems relate to ethical concerns. This creates an opportunity to consider the explicit use of ethical theories to address these problems. Digital phrenology and the algorithmic injustice to which it leads have been demonstrated to directly impact society. Considering the application of ML algorithms, the contexts in which they are employed, and the outcomes their use engenders, as an interrelated system highlights the different perspectives that must be considered, especially when it results in harm. This harm has successfully been mapped to ethical concerns relating to algorithms. Considering these mappings, an ethics-based solution to the identified problem(s) related to ML algorithms can be conceptualised. In the next section, a few possible ethical solution strategies will be considered to make an ethics-based contribution to the mitigation of digital phrenology and algorithmic injustice.

3 | Ethics-based Strategy to Address Digital Phrenology

An ethics-based solution to phenomena such as digital phrenology has been considered in various studies before. Researchers of these studies refer to terms such as *Ethical AI* (Mittelstadt, 2019), *Responsible AI* (Eitel-Porter, 2021), and even *AI Ethics* (Coeckelbergh, 2020). For this reason, ethical AI can be considered an instantiation of its 'designers' good intentions (Winfield et al., 2019). Organisations can direct their good intentions into a principles-based approach to achieving ethical outcomes. Organisations often share five common "pillars" that each of them considers responsible AI principles, namely *fairness*, *accountability*, *transparency*, *explicability*, and *privacy* (Eitel-Porter, 2021).

These common principles for ethical AI as identified by (Eitel-Porter, 2021) are not exhaustive. Floridi et al. (2018) argue that the four core principles commonly used in bioethics – which is the application of ethics to the field of medicine and healthcare – adapt well to the ethical challenges relating to ML algorithms. They contextualise the four bioethical principles (*beneficence*, *non-maleficence*, *autonomy*, and *justice*) and argue that adding a fifth principle (explicability) to complement the first four and contribute to a more complete ethics-based approach for ethical AI. The rationale is to address questions of accountability and transparency in algorithmic decision-making and to have a chain of command if its outcomes cause injustices.

The principles presented by Eitel-Porter (2021) and Floridi et al. (2018) can serve as sound guiding principles for the designers of algorithms and also serve as a reminder of what is expected from an ethical ML model. It is, however, not enough to simply have these principles listed for designers. Principles alone cannot guarantee ethical AI, partly because a principled approach may have a limited impact on design and governance. Lauer (2021) asserts that ethical principles must be implemented in ethical organizations and that AI ethics cannot be considered separate from its requirement to be adopted into a space where ethics is already part of organizational culture. Because it may be difficult to ensure that every organization has adopted an ethical culture before it adopts principles, it may not be the right time to use the principle-based approach just yet.
In an attempt to ensure that ML algorithms and AI systems behave ethically, Anderson and Anderson (2021) propose the consideration of six general approaches to effect ethics into algorithms without simply implementing ethical principles:

- Algorithms should be hardcoded to prevent actions that are generally considered to be unethical;
- Put the burden on the user to ensure that ML systems will only be used ethically, simply providing the user with warnings;
- Learn from earlier decision-'makers' judgements or current polls of what people generally think is ethically acceptable behaviour and let this guide an ML system;
- Use an existing ethical theory to guide the behaviour of ML systems*;*
- Impose a hierarchy of ethical principles on programs or machines to guide their behaviour; and
- Attempt to learn what is ethically acceptable from those with expertise in ethics, deriving from their input not only ethical principles appropriate for such systems that function in particular domains but also a way to represent the building blocks of ethics.

The fourth approach: "*Use an existing ethical theory to guide the behaviour of ML systems*, will be used in this paper. It involves the consideration of ethics during its design and implementation to govern a machines behaviour, rather than having designers try to block unethical behaviour or simply warn users after a biased implementation. However, a much more comprehensive ethical approach may be needed, and the question will then be on what ethical theory such an approach will be based on.
The evaluation of ML algorithmic ethicality will be done using existing ethical theories. The two most discussed ethical theories that can be employed in this context, are Utilitarianism and Deontology (Anderson & Anderson, 2021). The former is a theory that determines whether a particular act is morally correct by assessing the consequence(s) of an action. As long as the consequence(s) brings about the most happiness for the most people (society), then according to Utilitarianism, that act is morally correct or ethical. Deontology determines whether an action is morally correct by assessing the act itself, irrespective of its consequences or any emotions that it may cause. For example, according to Deontology, the act of stealing is wrong and remains wrong even if the theft of bread saved someone from dying of hunger (Jiang *et al*., 2021). The most popular Deontology approach is that of Immanuel Kant, known as the categorical imperative. This approach depends on the application of reason and considers an act moral if the agent agrees that it may become a universal law (Kant, 1785). Simply put, the categorical imperative only considers an act ethical if the 'agent's reasoning allows him to believe the act would be correct if it were universally done by or against everyone (including the agent) all the time, with absolutely no exceptions.

4 | Illustrative Application Examples

The use of existing ethical theories as an approach to assessing the ethicality of algorithms will be demonstrated in this section. Two short hypothetical examples of digital phrenology, motivated by the case studies in Section 2.1, will be evaluated using Utilitarianism and Deontology. The purpose is to illustrate that those ethical concerns relating to ML, as identified in Exhibit 1, can be assessed by referring to existing ethical theories. Even though a variety of ethical theories exist (Granitz & Loewy, 2007), using them to assess the ethicality of algorithms, along with more detailed case studies, fall beyond the scope of this paper and will form part of future research.
Before assessing the ethicality of hypothetical digital phrenology examples, the stage in the ML lifecycle where ethics would make the most significant difference, will be identified. According to Ashmore et al. (2021), the ML cycle consists of four separate yet interdependent stages, namely: *data management*, *model learning*, *model verification* and model *deployment*. The primary activity in the data management stage is *data collection*, and for these illustrative examples, this activity will be emphasised.

The ethical concern, *misguided evidence*, is mainly concerned with the data collection and management activities of ML. and the two hypothetical examples of digital phrenology also deal specifically with this concern. The hypothetical examples will be structured as follows: A machine learning scenario will be presented, followed by an ethical question that pertains to the scenario. The two ethical theories will then be used to evaluate the scenario and attempt to offer an answer to the ethical question. This form of an ethical assessment was employed by other application studies (Mouton et al., 2015).

Hypothetical example 1: motivated by a Health care ML application (Obermeyer et al., 2019).

Scenario: A hospital system includes a patient's health care cost history. An ML algorithm determines whether patients require specialized healthcare or not. The algorithm determines that patients with a low-cost history would not require specialized health care (and therefore denies it based on this reasoning), while patients with high-cost histories were determined to require specialized care. This application assumes that every person ill enough to require specialized health care would have a long history of illness which would relate to a high health care cost history of each person.
Ethical question: Is it ethical to recommend specialized healthcare based on data, such as health cost history, that does not indicate the seriousness of the 'person's current health status?
Deontology: Data that does not indicate the seriousness of a 'person's health status is a misrepresentation of the actual health need. There is also an element of lying to the patient when not telling them that the outcome is based on their cost history instead of seriousness of their current illness. Cost history cannot reasonably represent an 'individual's current need. Deontology would consider misrepresentation of data and lying to patients to be unethical. Misrepresentation and lying do not adhere to the requirement of being considered a universal law or becoming the same.
Utilitarianism: Cost history does not indicate and may be a misrepresentation of a 'person's actual health needs, but Utilitarianism may consider it ethical to use cost care history to determine health care needs. The misrepresentation would be ethical if its consequences lead to the good of society. In this case the good of society would be realized if the majority of patients receive specialized health care.

Hypothetical example 2: motivated by a Correctional services ML application (Li, 2017).

Scenario: A correctional services system includes offenders' race as input to an ML algorithm. The algorithm determines whether a parole applicant is likely to re-offend (and not receive parole) or not likely to re-offend (and receive parole). Offenders of a particular race are determined more likely to re-offend in future, and subsequently, offenders of that race would be denied parole. This application focuses more on race (which is not in the control of an offender) while neglecting other factors such as good behaviour (which can be controlled by an offender).
Ethical question: Is it ethical to recommend parole based on data such as race that the offender has no control over while neglecting information that the offender has control over, such as good or bad behaviour?
Deontology: Using racial data of a person in determining an 'applicant's parole constitutes racial discrimination. Deontology would consider racial discrimination unethical, because racial discrimination does not adhere to the requirement of being considered a universal law or becoming the same.
Utilitarianism: To use racial data to determine an 'applicant's parole, *i.e.*, racially discriminate, might be ethical if the algorithmic outcomes yield consequences that lead to the good of affected applicants. In this case, the good of affected applicants can be realized if most parolees receive parole. However, in countries such as South Africa, where most of the population is black, it is assumed that most of the jail population will also be black. If the ML algorithm determines people from the black race are more likely to re-offend, then Utilitarianism would consider racial discrimination unethical.

Based on these two examples, using ethical theories to evaluate ethical concerns in ML can be helpful. It is important to note that the application of Utilitarianism would require context, the consideration of consequences, and consideration of who the actual "society" or majority, in a given scenario is. Nonetheless, Utilitarianism can arrive at an explicable conclusion on the ethicality of every ML ethical question. Even though Deontology does not consider the consequences, or even context of a scenario, it is still possible for the two ethical theories to reach the same conclusion. The concept of ethicality from a deontological perspective and that from Utilitarianism should, therefore, not be assumed to always conflict with one another. The two may reach the same conclusion depending on the consequences observed from the utilitarian perspective.
The two hypothetical examples were chosen to illustrate the ethical concerns related to misguided evidence (evaluation of inputs and data collection) as part of the data management stage of ML. The resulting evaluation contributes to ethical data management by highlighting moral issues that arise from including discriminatory variables in the input data. For example, in the first scenario, it was clear that the input variable "cost" may not be used in isolation to

determine health care, while the second scenario identified "race" as a problematic input variable. Therefore, the misguided evidence ethical concern can be addressed by not using irrelevant input data in algorithms or at least not focusing too much on these kinds of data.
Using existing ethical theories provides more insight into the digital phrenology problem and can contribute to the development of ethical ML algorithms. This leads to the harmonious functioning and interacting of algorithms, with other sub-systems of society.

5 | Conclusion

Digital phrenology may have dire consequences, and if the risks associated with it are not mitigated, social ills become inevitable. To this effect, in Section 1, a research question and sub-question were identified and are subsequently revisited here:

To what extent is digital phrenology exposed by a systems perspective? – Section 2 presented examples of how digital phrenology is a reality in modern society. The examples presented are that of ML systems, used to make decisions with an impact on society as on overarching system. The examples were presented in a twofold manner. Firstly, digital phrenology was identified in three real-world case studies, which were then supplemented with discussions from literature. It was mainly attributed to biased input data used to train the underlying ML algorithms. The harmful impact found in ML algorithms' decisions were pointed out and then mapped to ethical concerns. Digital phrenology and the algorithmic injustice it causes, show that algorithms do not always present with ethical behaviour. ML algorithms are not inherently unethical if one simply looks at the function it is expected to perform. From a systems perspective, however, it revealed that digital phrenology is present in certain ML algorithms.

How can ethical theories be applied to assess algorithmic ethicality? – The use of ethical theories to evaluate the ethicality of ML scenarios was illustrated in Section 4. This was done by identifying a strategy (Section 3) for applying such theories and then applying the chosen approach to hypothetical ML scenarios. The process and resulting outcomes illustrate how the ethical theories can be used to assess algorithmic ethicality. These illustrations also revealed that the use of ethical theories to identify whether an ML algorithm is ethical (or not) can be a suitable approach.
Given the ethical challenges arising from the use of ML systems, it is nonetheless important to recognize and acknowledge the amount of good these systems have contributed to society. This study does not suggest the abolishment of algorithms or that the use of algorithms is inherently bad. However, the ethical coexistence of technology and humans needs to be prioritized over the "good" that can come from using technology in society.
As this is part of an ongoing research project, future work will include an in-depth study of other existing ethical theories. In this paper, only two of the most commonly used ethical theories were considered and briefly illustrated. The investigation of others may contribute toward the combatting of digital phrenology.

6 | References

Akbari, A., Ng, L., & Solnik, B. (2021). Drivers of economic and financial integration: A machine learning approach. *Journal of Empirical Finance*, *61*, 82-102.

Alhayani, B., Mohammed, H. J., Chaloob, I. Z., & Ahmed, J. S. (2021). Effectiveness of artificial intelligence techniques against cyber security risks apply of IT industry. *Materials Today: Proceedings*, 1-6.

Anderson, S. L., & Anderson, M. (2021). AI and ethics. *AI and Ethics*, *1*(1), 27-31.

Arnold, R. D., & Wade, J. P. (2015). A definition of systems thinking: A systems approach. *Procedia computer science*, *44*, 669-678.

Ashmore, R., Calinescu, R., & Paterson, C. (2021). Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)*, *54*(5), 1-39.

Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, *4*(1), 111-138.

Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns*, *2*(2), 1-9.

Bittel, C. (2019). Testing the truth of phrenology: Knowledge experiments in antebellum American cultures of science and health. *Medical history*, *63*(3), 352-374.

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. MIT Press.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access*, *8*, 75264-75278.

Churchman, C. W. (1968). *The systems approach*. Delacorte Press.

Coeckelbergh, M. (2020). *AI ethics*. MIT Press.

Das, A. K., Ashrafi, A., & Ahmmad, M. (2019). *Joint cognition of both human and machine for predicting criminal punishment in judicial system* 4th International Conference on Computer and Communication Systems (ICCCS),

Dixon, S., & Quirke, L. (2018). What's the harm? The coverage of ethics and harm avoidance in research methods textbooks. *Teaching Sociology*, *46*(1), 12-24.

Eitel-Porter, R. (2021). Beyond the promise: implementing ethical AI. *AI and Ethics*, *1*(1), 73-80.

Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, *9*(01), 1-16.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., & Rossi, F. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689-707.

Fuchs, D. J. (2018). The dangers of human-like bias in machine-learning algorithms. *Missouri S&T's Peer to Peer*, *2*(1), 1.

Granitz, N., & Loewy, D. (2007). Applying ethical theories: Interpreting and responding to student plagiarism. *Journal of business ethics*, *72*(3), 293-306.

Green, B., & Hu, L. (2018). *The myth in the methodology: Towards a recontextualization of fairness in machine learning* Proceedings of the machine learning: the debates workshop,

Gullapalli, V. K., & Brungi, R. (2015). A novel methodology to implement optimization algorithms in machine learning. *International Journal of Computer Applications*, *112*(4), 33-36.

Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., & Kumar, P. (2021). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity*, *25*(3), 1-46.

Hayes, S. (2020). Cautionary Ethics Tales: Phrenology, Eugenics...and Data Science? In B. Franks (Ed.), *97 Things About Ethics Everyone in Data Science Should Know* (pp. 9-12). O'Reilly Media.

Juhn, Y. J., Ryu, E., Wi, C.-I., King, K. S., Malik, M., Romero-Brufau, S., Weng, C., Sohn, S., Sharp, R. R., & Halamka, J. D. (2022). Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index. *Journal of the American Medical Informatics Association*, *29*(7), 1142-1151.

Kant, I. (1785). The categorical imperative.

Kodiyan, A. A. (2019). An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool. *Researchgate Preprint*.

Lauer, D. (2021). You cannot have AI ethics without ethics. *AI and Ethics*, *1*(1), 21-25.

Leavy, S. (2018). *Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning* Proceedings of the 1st international workshop on gender equality in software engineering,

Li, Y. (2017). *Algorithmic discrimination in the US justice system: A quantitative assessment of racial and gender bias encoded in the data analytics model of the correctional offender management profiling for alternative sanctions (COMPAS)* Johns Hopkins University].

Mahmoud, A. A., Shawabkeh, T. A., Salameh, W. A., & Al Amro, I. (2019). *Performance predicting in hiring process and performance appraisals using machine learning* 10th International Conference on Information and Communication Systems (ICICS),

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021). Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, *58*(5), 1-32.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501-507.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 1-21.

Mladenov, M., Creager, E., Ben-Porat, O., Swersky, K., Zemel, R., & Boutilier, C. (2020). *Optimizing long-term social welfare in recommender systems: A constrained matching approach* 37th International Conference on Machine Learning,

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, *26*(4), 2141-2168.

Mouton, F., Malan, M. M., Kimppa, K. K., & Venter, H. S. (2015). Necessity for ethics in social engineering research. *Computers & Security*, *55*, 114-127.

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

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (Vol. 1). Crown.

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.

Ploug, T., & Holm, S. (2020). The right to refuse diagnostics and treatment planning by artificial intelligence. *Medicine, Health Care and Philosophy*, *23*(1), 107-114.

Schlag, P. (1997). Law and phrenology. *Harvard Law Review*, *10*(4), 877-921.

Schwendicke, F. A., Samek, W., & Krois, J. (2020). Artificial intelligence in dentistry: chances and challenges. *Journal of dental research*, *99*(7), 769-774.

Stypińska, J. (2021). *Ageism in AI: new forms of age discrimination in the era of algorithms and artificial intelligence* 1st Conference on AI for People, Bolgona, Italy.

Uddin, M. A., Stranieri, A., Gondal, I., & Balasubramanian, V. (2020). *Dynamically Recommending Repositories for Health Data: a Machine Learning Model* Proceedings of the Australasian Computer Science Week Multiconference (ACSW 2020),

Wheeler, A. P., & Steenbeek, W. (2021). Mapping the risk terrain for crime using machine learning. *Journal of Quantitative Criminology*, *37*(2), 445-480.

Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). Machine ethics: The design and governance of ethical AI and autonomous systems [scanning the issue]. *Proceedings of the IEEE*, *107*(3), 509-517.

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., & Metcalf, J. (2017). Ten simple rules for responsible big data research. *PLOS COMPUTATIONAL BIOLOGY*, *13*(3), 1-10.