Applying a principle of explicability to AI research in Africa: should we do it?

Mary Carman
Department of Philosophy, University of the Witwatersrand, Johannesburg, South Africa

Benjamin Rosman
School of Computer Science and Applied Mathematics,
University of the Witwatersrand, Johannesburg, South Africa

Abstract

Developing and implementing artificial intelligence (AI) systems in an ethical manner faces several challenges specific to the kind of technology at hand, including ensuring that decision-making systems making use of machine learning are just, fair, and intelligible, and are aligned with our human values. Given that values vary across cultures, an additional ethical challenge is to ensure that these AI systems are not developed according to some unquestioned but questionable assumption of universal norms but are in fact compatible with the societies in which they operate. This is particularly pertinent for AI research and implementation across Africa, a ground where AI systems are and will be used but also a place with a history of imposition of outside values. In this paper, we thus critically examine one proposal for ensuring that decision-making systems are just, fair, and intelligible—that we adopt a principle of explicability to generate specific recommendations—to assess whether the principle should be adopted in an African research context. We argue that a principle of explicability not only can contribute to responsible and thoughtful development of AI that is sensitive to African interests and values, but can also advance tackling some of the computational challenges in machine learning research. In this way, the motivation for ensuring that a machine learning-based system is just, fair, and intelligible is not only to meet ethical requirements, but also to make effective progress in the field itself.

Keywords Principle of explicability · Machine learning · Intelligibility · Accountability · Africa

---

This is a post-peer-review, pre-copyedit version of an article published in Ethics and Information Technology. The final authenticated version is available online at: https://doi.org/10.1007/s10676-020-09534-2

Please cite the final version.

1. Introduction

Developing and implementing artificial intelligence (AI) systems in an ethical manner faces several challenges specific to this technology. As research and implementation surges forward, it is necessary to develop guidelines for ensuring that we are heading in a desirable direction, ranging from assessing the moral status of AI to ensuring that the process of research and development aligns with the values we hold within our societies. For instance, at the research and development stage, we need to ask the question of how we can ensure that any automated decision-making system is just, fair, and intelligible, to ensure that when we cede decision-making power to artificial agents, they are aligned with our values and lines of accountability can be made clear. The term 'decision-making system' is loaded and quite vague, but by this we mean systems that make use of some form of machine learning. Given that values vary across cultures, an additional ethical challenge is to ensure that these AI systems are not developed according to some unquestioned but questionable assumption of universal norms but are in fact compatible with the societies in which they operate. This is particularly pertinent for AI research and implementation across Africa, a ground where AI systems are and will increasingly be used, but also a place with a history of imposition of outside values.[1] While various frameworks and principles have been developed internationally for guiding 'Good AI', and while discussions about AI in Africa typically draw on these existing frameworks (see, for instance, Microsoft 2019), there is a notable lack of African voices contributing to these discussions. There is thus a need to critically examine whether the frameworks are in fact relevant for and compatible with application in an African context.

In this paper, we take initial steps to address this need by assessing one such proposal for ensuring that decision-making systems are just, fair, and intelligible, to assess whether it should be adopted in an African research context. This is the proposal that we adopt a principle of explicability to generate specific recommendations for guiding the development of ethical AI, a principle that has not yet been assessed for African applicability. It is our contention that a principle of explicability not only can contribute to responsible and thoughtful development of AI that is sensitive to African interests and values but can also advance tackling some of the computational challenges in machine learning research. In this way, the motivation for ensuring that any machine learning-based system is just, fair, and intelligible is not only to meet ethical requirements, but also to make effective progress in the field itself. Our paper, then, firstly begins a critical assessment of the applicability in an African context of a proposal for guiding ethical AI research that has so-far been missing in the literature,

[1] Africa is, of course, a vast continent with many different cultures and peoples. While we talk of 'Africa' in this paper for ease of reference, we do not deny that within Africa there is great complexity.

and secondly builds on the motivation and support for adopting the proposal more generally. It is important to note that, while our particular focus is on a broadly-construed African context, the need for contextual and cultural sensitivity can be echoed more widely, calling attention to the need for care when drawing on generic principles that may or may not be universal in scope.

In section 2, we begin by introducing what we mean by AI and machine learning, and describing some of the AI landscape in Africa. As AI research and implementation is expanding across Africa, we need guidelines to ensure that it is done in an ethical manner. So, in section 3, we turn to existing guidelines for 'Good AI' and, specifically, the European AI4People framework that identifies five guiding principles for Good AI. These are the familiar principles of respect for autonomy, beneficence, non-maleficence and justice from Western bioethics, but also the additional AI-specific principle of explicability. As the first four principles are already well-discussed within Western and African bioethics, our focus will be on the new principle of explicability. In section 4, we therefore engage with the question of whether the principle of explicability should indeed be adopted in an African research context, considering and rejecting two potential reasons for why it should not. Indeed, or so we argue, the relevance and importance of the principle of explicability arises from the kind of research at stake wherever it is conducted, Africa included.

2. Context-setting: AI and AI in Africa

Artificial intelligence is broadly taken to refer to imbuing a system with some form of computational intelligence. Under this broad umbrella, machine learning is the core technology which involves using data to optimise the parameters of a computational model, which are typically used for some form of prediction (typically regression or classification) or decision-making (reinforcement learning, over longer time horizons). The very nature of machine learning, however, as we discuss in more detail in section 4.1, raises the challenge of how we can ensure that systems involving some form of machine learning are intelligible to humans, and that lines of accountability are made clear.

While machine learning and AI research and development in Africa has a long history, this has always happened in small pockets across Africa. Activity across the continent has more generally exploded over the past five years, with strong hubs forming in places such as Johannesburg and Stellenbosch in South Africa, Nairobi in Kenya, and Accra in Ghana.

In addition, several recent events and initiatives illustrate the growing interest in developing the capacity to strengthen AI and machine learning research in Africa. These include programmes such as Data Science Africa, Data Science Nigeria, and the Deep Learning Indaba. All of these aim at

explicitly growing the African machine learning community, largely through technical training events and gatherings. The Deep Learning Indaba (2019), for example, is a large Africa-centric summer school which has spawned satellite 'IndabaX' events in 27 different African countries. This kind of programme has led to both growth and better organisation in the African machine learning community, as evidenced in greater participation of Africans in international conferences.

A driver for research and implementation is the vast potential for AI and related technologies to have a positive impact on communities and economies in Africa. Microsoft's 2019 White Paper on the opportunities that AI offers for growth, development and democratisation in Africa, for instance, highlights four core sectors where AI could have a positive impact. These are in agriculture, by improving efficiency and effectivity; in healthcare, by improving quality and increasing access; in public services, by improving efficiency and responsiveness, and enhancing impact; and in financial services, by improving security and expanding reach. The interest in developing and implementing AI in Africa for social good does not just come from outside of the continent but can be found within Africa itself. With the hype around the so-called 'Fourth Industrial Revolution', for instance, various bodies have been set up to explore and promote the use of technologies like AI, machine learning and nanotechnology in Africa, centres such as the South African Affiliate Centre of the World Economic Forum's Centre for the Fourth Industrial Revolution (C4IR).

Given the interest in and likely growth of future research, as well as the trajectory towards implementing more AI and related technologies, there is an urgent need to ensure that any such developments and implementations are done responsibly and thoughtfully. With regard to the powerhouse of machine learning, the need for systems that are just, fair and intelligible is a very real need if we are to guide research in Africa in the direction we want.

3. Guidelines for Good AI and the principle of explicability

A major focus in current AI research, from both the technical and philosophical communities, is on ensuring 'Good AI': that AI is developed and implemented in an ethical and sustainable manner. For instance, at NeurIPS 2018, workshops were held on *Ethical, Social and Governance Issues in AI, Challenges and Opportunities for AI in Financial Services, Machine Learning for the Developing World (ML4D)*, and *AI for Social Good*. In the past few years, several guidelines and frameworks for ensuring Good AI for society and Good AI research have already been drawn up. These include the *Asilomar AI Principles* (2017), the crowd-sourced 'Ethically Aligned Design: A vision for prioritising human well-being with autonomous and intelligent systems' (2017), Microsoft's white paper titled 'Artificial Intelligence for Africa: An opportunity for growth, development and democratisation'

(2019), and the European AI4People's publication, 'AI4People - An Ethical Framework for a Good AI Society: Opportunities, risks, principles, and recommendations' (2018), where this last framework surveys a range of guidelines and frameworks to develop a synthesis of existing principles.[2] With the proliferation of frameworks and guidelines, it makes sense to examine their commonalities and so our focus is on the AI4People framework because of the synthesis it offers of other frameworks. Through their synthesis, the authors identify five recurring ethical principles that are recognised in one way or another by all of the guidelines surveyed by the AI4People project. In this section, we introduce the five principles it identifies, the four Western bioethical principles of respect for autonomy, beneficence, non-maleficence and justice, along with a fifth principle specifically for AI, the principle of explicability, which is our focus.

The principle of respect for autonomy is roughly 'the idea that individuals have a right to make decisions for themselves about the treatment they do or not receive' (Floridi et al. 2018, p. 697). Applied to AI where we might 'willingly cede some of our decision-making power to machines', the principle requires 'striking a balance between the decision-making power we retain for ourselves and that which we delegate to artificial agents' (p. 698). The principle of beneficence, in turn, requires 'promoting well-being, preserving dignity, and sustaining the planet' – basically, developing AI technology that benefits humanity (p. 696). The closely related principle of non-maleficence is one of doing no harm, requiring avoiding certain overuses and misuses of AI technologies. The fourth principle of justice typically requires the fair distribution of goods and services. Applied to AI, justice might require using AI to right previous wrongs, ensuring that the benefits of AI are shared fairly (and, presumably, that the burdens are fairly distributed), and ensuring that any new harms are prevented (p. 699).

The fifth principle is the principle of explicability. In a context where a select few are leading the way in the development and implementation of AI technologies that either directly or indirectly impact the rest of society, the various surveyed guidelines call for 'the need to understand and hold to account the decision-making processes of AI', while recognising that the workings of AI 'are often invisible or unintelligible to all but (at best) the most expert observers' (Floridi et al. 2018, p. 700). As

[2] The guidelines and frameworks surveyed are: The *Asilomar AI Principles* (2017), the *Montreal Declaration for Responsible AI* (2017), the *General Principles* in the IEEE Global Initiative's 'Ethically Aligned Design' (2017), the *Ethical Principles* in the 'Statement on Artificial Intelligence, Robotics and "Autonomous" Systems' of the European Group on Ethics in Science and New Technologies (2018), the principles of the 'AI in the UK: Ready, willing and able?' report of the UK House of Lords Artificial Intelligence Committee (2018), and the *Tenets* of the Partnership on AI (2018).

the authors describe it, the principle of explicability should be understood in both an *epistemological* sense of intelligibility and in an *ethical* sense of accountability.

In the epistemological sense, the principle asks for an answer to the question of *'how does it work?'*. This epistemological sense can be found in the Asilomar AI Principles (2017), for instance, as a requirement for 'failure transparency': 'if an AI system causes harm, it should be possible to ascertain why'. The General Principles of 'Ethically Aligned Design' (IEEE 2017) call for a need for the basis of a decision to be discoverable, as does the Partnership on AI (2018), calling for 'the operation of AI systems to be understandable and interpretable by people, for purposes of explaining the technology'. The need for transparency and explainability is identified in the European Group on Ethics in Science and New Technologies (2018) and in the UK House of Lords Artificial Intelligence Committee report (2018). All of these call for an answer to the epistemological sense of '*how does it work?'*

In the ethical sense, the principle of explicability tackles issues of accountability by asking for an answer to the question of *'who is responsible for the way it works?*[3] For instance, the Asilomar Principles include a requirement that the designers and builders of AI systems have a duty to shape the moral implications of the use, misuse and abuse of those AI systems. Both the Partnership on AI and the European Group recognise that AI research and development needs to be accountable to a range of stakeholders, with the House of Lords report calling for clear lines of accountability.

The principle of explicability is valuable on a number of fronts. Firstly, it addresses the uneven power structure already apparent in the development of AI, between those who are developing the technologies (typically large corporations) and those who will be affected by them (the consumers and the rest of society). Secondly, it both complements and enables the other four principles. For AI to be both beneficent and non-maleficent, we need to understand what kinds of benefits and harms it can actually do within a society. Similarly, if we are to respect human autonomy, we need to know how an AI system would choose and act. Additionally, for the principle of justice to be respected, we need to ensure that there is accountability. Thirdly, the principle recognises the role that intelligibility and accountability can play in engendering public trust and understanding, necessary for ensuring that AI is accepted within society. Without public trust and understanding, the potential economic and societal benefits of AI and related technologies could fail to materialise (Winfield and Jirotka 2018). For instance, a lack of trust can be seen to underlie fears about automation negatively

[3] This question is an ethical question. There are, of course, questions about legal accountability and responsibility but we do not attend to them in this paper.

impacting human employment. With powerful bodies like South Africa's Congress of South African Trade Unions (COSATU) not wholeheartedly behind such technologies – 'You can't be talking about the future of work when you describe displacement and unemployment' (Steyn 2017) – the benefits will be difficult, if not impossible, to achieve. The motivation that the AI4People framework draws on for the principle of explicability, which includes requirements for intelligibility and accountability, is in large part based on societal benefits such as public trust and understanding.

4. Towards ethical AI in and for Africa

The AI4People framework uses the five identified principles to generate a set of recommendations for Good AI within a European context, acknowledging that recommendations based on the principles may differ in different cultural contexts – at least to the extent that 'different cultural frameworks inform attitudes to new technology' (Floridi et al. 2018, p. 701). So, if we are to apply the principle in an African research context and to the design of systems to be implemented in Africa, we need to develop our own recommendations based on principles that are culturally and socially sensitive. However, truly acknowledging the impact of different cultural contexts is not limited to appreciating that different cultural frameworks inform attitudes to new technology. It more fundamentally requires ensuring that the principles themselves are applicable.

Take, for instance, the communitarian nature of many African cultures and worldviews. While not all African cultures and worldviews are communitarian, while those that are need not be identical to one another, and while communitarian cultures and worldviews exist outside of Africa, the centrality of community is widely agreed to be a salient and dominant feature that can be found in various forms across the continent below the Sahara, and a feature that has been drawn on by those working within sub-Saharan African philosophy and ethics.[4] In many communitarian societies, people often engage in joint decision-making or refer to authority figures for guidance as part of their decision-making, thereby legitimately including others in a normal process. This is in stark contrast to a typical Western worldview that centralises the individual, and which is reflected in bioethical principles like the principle of respect for autonomy, frequently understood as respecting the decisional autonomy of an individual who makes decisions without undue coercion (see Beauchamp and Childress 2012). Similarly, the AI4People report expresses the principle of respect

[4] For a sample of seminal philosophical work highlighting community within different cultural contexts, see Mbiti 1990 (Kenya, although with a systematic review of other cultures), Gykeye 1987 (Akan, Ghana), Gbadegesin 1991 (Yoruba, Nigeria) and Ramose 2005 (South Africa).

for autonomy in a way that highlights the focus on the individual: 'individuals have a right to make decisions for themselves' (Floridi et al. 2018, p. 697).

The salience of community versus a strong individualism illustrates why we require, on one hand, sensitivity in how we adopt and adapt the principles in different contexts, if we are to apply them. For instance, the World Health Organisation (WHO) has made provisions to allow partner agreement in reproductive research in certain countries with a cultural tradition of involving partners or family in decision-making, despite usually taking the involvement of a partner as a violation of participant autonomy (Moodley 2007, see also WHO 2020). The principle of respect for autonomy is still applied, but it is adapted to reflect the real communitarian-infused decision-making processes that people engage in.

On the other hand, we might reject that any of these principles should be applied at all. Over the past few decades, an increasing amount of work has been done developing the field of African bioethics in response to exactly this kind of challenge (as samples, see Murove 2005; Behrens 2013, Chukwuneke et al., 2014; Rakotsoane and Van Niekerk 2017; Barugahare 2018).

Reasons given for rejecting – or at least seriously critiquing the applicability of – Western principles include both the pragmatic and the theoretic. Pragmatically, simply adopting foreign principles that are divorced from the ethical worldviews that govern ordinary peoples' lives can result in practices that are inefficient in achieving their aims. In general, people are more inclined to accept ethical ideas or interventions if they are consistent with their own worldviews (Behrens 2013). The former director of Médecins Sans Frontiers, Roy Brauman, gives the example of emergency food supply in famine-stricken Uganda. Medical workers prioritised giving food to the most vulnerable, women and children, only to discover that the food was being taken away and given to local elders in lines with local customs that prioritise respecting social orders (Hellsten 2006, p. 73). Such an example illustrates how applying a principle like justice without sensitivity to local context can be ineffective in achieving its aims. This could be addressed by applying the principle in a culturally sensitive manner; however, and more crucially, the example also illustrates how a reliance on predetermined principles can exclude other principles that in fact govern people's behaviour.

While pragmatic issues could potentially be addressed by adapting the principles in a context-sensitive manner while also being open to the existence of other principles, a deeper theoretical set of issues remain. These are particularly pertinent in the postcolonial African context where there is a tradition of postcolonial critique and an expressed need for the reclamation of human dignity, authenticity, and a positive assertion of African identity following centuries of subjugation by

Western powers.[5] As Andoh writes, drawing on this tradition, an attitude of 'assimilating Western values and ideologies into Africa can give rise to a situation of self-dehumanisation and outright self-subversion both in terms of dignity and self-esteem' (Andoh 2011, p. 69). For instance, given the salience of community in many sub-Saharan African cultures, a prioritising of individualism can sever the person from her 'relational spheres of existence' (Murove 2005, p. 27). Rather than simply adopting an individualistic principle like respect for autonomy, we might thus instead adopt a broader principle of respect for persons, which captures the essence of the principle but allows more cultural nuance (Behrens, 2013). Alternatively, we could introduce new principles like 'human life invaluableness' (Rakotsoane and Van Niekerk 2017), or even ones that capture the respect for social order brought out in the Ugandan example.

These critiques of the traditional bioethical principles put pressure on the applicability of the very same principles found in the AI4People framework for use within an African context, or indeed any other cultural context that does not share similar features to the central European context in which the AI4People framework was derived. We cannot simply, and without critical engagement, adopt the principles and use them to generate context-specific recommendations, at risk of resulting in ineffective processes or causing genuine moral harm. We first have to assess if the principles themselves can be recruited within the context at hand. Applying the principles of respect for autonomy, beneficence, non-maleficence and justice in an African AI context requires more examination, but our current focus is on the *new* principle of explicability. This is because there is already a rich literature within African bioethics, as well as the field of global bioethics more widely, examining the supposedly universal applicability of the standard principles of Western bioethics, whereas the principle of explicability is AI-specific and has yet to receive critical attention. So, while we could question what *form* such a principle would take in a particular cultural context – if explicability is closely related to communication and public trust as described at the end of the previous section, for instance, then different cultural norms of communication may inform what needs to be made explicable, to whom, when, and to what degree – we must first question whether the principle of explicability is relevant and applicable at all. This second level of critique, which is our focus, is crucial given the theoretical issues that warn against the uncritical assimilation of Western values into African contexts.

In the rest of this section, we critically assess whether the principle of explicability should be applied in an African context, rather than just what form it should take if it were applied. In section 4.1 we

[5] This is a tradition that draws on a diverse range of theorists from across the continent, such as Senghor (1988), Mbembe (2001), wa'Thionga (1986) and Biko (2002).

argue that the importance of the principle arises in part from the very nature of the technical research at hand, while in 4.2 we consider two possible reasons not to adopt it in an African research context, arguing that neither takes hold.

4.1. The importance of the principle of explicability

The principle of explicability, *prima facie* at least, is not obviously based on strongly Western values like the individualism that underlies the principle of respect for autonomy. In fact, it is a principle that could allow us to be sensitive to cultural nuances as a matter of necessity and, as we will suggest, a principle that arises out of the nature of much of the research in question. The epistemological sense of the principle, at least, is required to address some of the computational challenges within machine learning. In this subsection, we thus illustrate how the principle of explicability ties in with some real issues and risks that computational and technical researchers are addressing, including those working in Africa, as a way of motivating for the adoption of the principle within an African research context.

The epistemological sense underpinning the principle of explicability seeks an answer to the question of 'how does it work?' There are two primary issues in which the epistemological sense of explicability is worth considering.

The first issue is that the fundamental modality of machine learning comes down to a human specifying a 'goal', or more technically an objective function, and the learning procedure is required to optimise a model, or technically the parameters of the model, for achieving this goal. This stands in stark contrast to more traditional programming paradigms where a human specifies the full set of steps (algorithm) that is to be executed by some software. Machine learning may instead, for example, require that a set of positive and negative outputs are provided to the learning system, which then infers the procedure for distinguishing between these categories itself. While this shift in attitude to problem solving has been transformative, there are at least two broad risks that arise.

The first risk is that of the human misspecifying the desired objective function. This may happen for example in reinforcement learning (RL), where an artificial agent is required to learn to take a sequence of decisions to achieve some long-term goal. In RL, desired states of the world are typically annotated with some positive reward, and undesirable states with some negative reward. In general, these attributions are arbitrary (both in location and magnitude), and it is trivially easy for an incorrect scaling or attribution to lead to undesired behaviours. A human, however, may not be fully cognisant of her own objectives, and as such may incorrectly imbue them into a learning system. An

oft-cited example of this is the thought experiment of the paperclip maximiser: an intelligent agent tasked with running a facility to produce a maximal number of paperclips could self-improve to the extent that the result is 'a superintelligence whose top goal is the manufacturing of paperclips, with the consequence that it starts transforming first all of earth and then increasing portions of space into paperclip manufacturing facilities' (Bostrom, 2013). The scenario shows how even this simple seemingly benign goal could be sufficient to generate behaviours antithetical to human life and flourishing.

This first risk speaks directly to what is known as the value alignment problem. The value alignment problem refers to the challenge of ensuring that the goals of an artificially intelligent system do not contradict (typically inadvertently) the values of humans, or society in general. In an epistemological sense, to ensure that the values are aligned requires seeking an answer to the question: 'how does it work?' This in itself is a nontrivial question to answer, as discussed in more detail below. In a case where the goals do contradict and go against human values, we are faced with the ethical question: 'who is responsible for how it works?' For instance, who is best able to explicitly identify all the relevant values that are often implicit within ourselves and our societies, and can we hold someone accountable for failing to ensure that the goals of an AI system cohere with some set of implicit values of humans? Would explicitly identifying relevant values using quantitative methods even be sufficient for capturing the complexity and flux of human values, something that Sloane & Moss (2019) question?

This problem is exacerbated by the fact that human values differ across different societies and contexts. As a simple example, consider the 'rules of the road' which change between countries, from the side of the road on which people drive, to the rules for entering roundabouts. This is even more notable in how societal preferences change between countries in versions of the 'trolley problem', where people around the world have been surveyed on which of two random sets of people should be spared in a vehicle collision (Awad, et al, 2018). One can note, for example, that there appears to be a preference for sparing the lawful and pedestrians in Japan, versus a preference for preserving humans over animals as well as high status individuals in Nigeria.[6]

The second risk is that the general form of a machine learning model may be difficult to interpret. The quintessential example of this is artificial neural networks, which are widely acknowledged to operate as 'black boxes'. This arises from the basic modelling assumptions, that the relationships between various inputs and the desired outputs are typically learned to be complex nonlinear

[6] http://moralmachineresults.scalablecoop.org/

functions, which are recursively embedded in other nonlinear functions. As the numbers of parameters that are learned in these models can easily be in the millions, interpreting these functions quickly loses feasibility. The question of 'how does it work?' thus becomes of paramount importance, with limitations for just how much detail we can give in answer. The question of 'who is responsible for how it works?' also highlights how very few humans, if any, are able to understand the processes that are followed, yet that select few still have considerably deeper understanding than the many who might be impacted by the technology.

It is interesting to note that other classes of machine learning models do not necessarily possess this same characteristic, although they very seldom achieve state-of-the-art performance that neural networks do. An example here is that of decision trees, which involve learning an ordering of features to treat as conditional rules for classifying data points. By following this sequence which is learned, one can easily trace out and validate decisions made by these models. Different models may thus generate different requirements in terms of answering the question 'how does it work?'

Another issue concerns biases that may be presented to the learning system in its training data. Examples of this have been widely seen in supervised learning, with numerous cases of racial biases being reported globally (Buolamwini and Gebru, 2018). This is particularly troublesome when confounded with the previous challenge of difficult-to-interpret models, in that it could easily become unclear that these biases have been introduced to the system. Establishing how a system is making decisions, especially when it is fed with data that we may not realise to be ethically suspect, is therefore important if we are not simply to recreate our own human inefficiencies as decision-makers and agents. When these biases reflect real-world human biases and have the potential for profound and detrimental real-world impact, we again face the ethical question of who is responsible and who should be held accountable.

The above examples illustrate some general computational issues within current research the world over, and they are issues wherever research is taking place and systems implemented, including in Africa. What becomes apparent from these examples is that we need something like the epistemological sense of the principle of explicability not just for engendering public trust and understanding and for ensuring alignment with societal values, but for actually informing research to tackle some of the computational challenges that are being faced. This is because those computational challenges themselves require alignment with societal values and needs, in turn requiring that certain values and objectives are made explicit. We thus need to be sensitive to the values and needs of the societies where AI technologies are developed and implemented. In the

AI4People report, the authors describe the 'dual advantages' of an ethical approach to AI, which allows identifying and leveraging 'new opportunities that are socially acceptable or preferable' and 'enables organisations to anticipate and avoid or at least minimise costly mistakes' arising from 'courses of action that turn out to be socially unacceptable and hence rejected' (Floridi et al., 2018, p. 694). As we have argued, the advantages of an ethical approach go even a degree deeper than the authors originally discuss. Various challenges facing technical research are in fact 'socio-technical' in nature (Crawford 2017). As such, applying the principle of explicability, especially in its epistemological 'how does it work?' sense, does not only have the dual advantages identified in the AI4People report, but an additional advantage in that it can help solve some of the computational problems facing AI researchers, avoiding courses of action that are ineffective in addition to but distinct from prioritising their social acceptability.

4.2. Are there reasons not to apply the principle in Africa?

While the principle of explicability is an important principle for guiding research both to achieve computational ends and to strive for societal benefits, there may nevertheless be reasons not to adopt and apply it in African research contexts. Here, we consider two such reasons and argue that they do not show that the principle itself is problematic or irrelevant. Rather, the potential problems highlight at least two lessons: the principle of explicability absolutely requires contextual sensitivity in its application, and it must be balanced with other relevant principles.

A first potential problem, the trade-off problem, relates to the epistemological sense of the principle of explicability. This is the problem that we might face undesirable trade-offs in demanding explicability. One of the recommendations put forward by AI4People is that, in a European context at least, a framework that enhances the explicability of AI systems that make socially significant decisions is developed, where 'central to this framework is the ability for individuals to obtain factual, direct, and clear explanation of the decision-making process' (Floridi et al. 2018, p. 702). One way to meet this demand is by requiring that systems produce explanations of their own behaviour (see, for instance, Selbst and Powells 2017; Doshi-Velez and Kim 2017; Winfield and Jirotka 2018). Yet, requiring explanations in this manner for a system to meet explicability requirements could hypothetically mean that the capabilities of that system are severely handicapped (Wachter, Mittelstadt and Floridi 2018). In a similar fashion, London (2019) has recently argued that a demand for explicability or for making something interpretable is typically a demand for an explanation of causal relations. Domains such as AI and even medical decision-making, however, typically involve associations that are not necessarily causal. As London argues, in a domain where we lack causal

knowledge but where predictive and diagnostic accuracy are nevertheless important, a demand for explicability can needlessly detract from accuracy and reliability.

So, if a system would lose accuracy or reliability in diagnosing some life-threatening disease by some percentage, how should we view this trade-off? This is a challenge that all societies need to address, but it is particularly pertinent in many African contexts where other solutions, such as medical professionals and state-of-the-art laboratories, are not easily at hand. As a representative example, consider Tanzania: as high as 71% of the population lives in rural, difficult-to-access areas with poor infrastructure, a fact that informs the Tanzanian government's current five-year health sector strategic plan for increasing access to healthcare services (United Republic of Tanzania 2015) and a fact that explains the welcoming of the use of drones to provide basic medical supplies (Landhuis 2017). This challenge is significant in the health sector, as a result of a number of factors such as different disease profiles around the world. Malaria, for example, poses a much greater risk in Africa than it does in Europe, and this coupled with a shortage of experts necessitates automated solutions (Brown et al. 2019).

Challenges also exist in the social sphere. Africa is home to an estimated 2,000 languages, and addressing communication barriers is an important step towards advancing these societies. The sheer scale of this problem again calls for AI-based solutions in automated translation (Abbott and Martinus 2019).

Part of the attraction of the development and implementation of AI solutions in Africa is that doing so can address challenges like these and others faced by African societies that arise from social, historical and geographical inequities that make solutions available elsewhere in the world untenable. This is something that the Microsoft White Paper discussed earlier does indeed highlight, by focusing on the way AI could be used to improve various sectors, such as agriculture, healthcare, public services and finance (Microsoft 2019). In this context, the stakes for demanding explicability at the expense of accuracy or reliability can be particularly high.

This problem, however, does not show that the principle is itself problematic or irrelevant in an African context. For one thing, there are two senses of explicability contained within the principle, the epistemological and the ethical. In the epistemological sense, we might seek alternative and less demanding ways to account for how a system works. We have argued, for instance, that solving some of the computational problems facing machine learning requires making objectives and goals explicit. As such, we could plausibly achieve explicability in the epistemological sense by specifying a system's design goals more carefully. Indeed, Kroll (2018) has proposed such an approach for

intelligibility, one that shifts from a focus on understanding technical tools to understanding the overall system, which includes people. Such a tactic need not require an explanation in terms of causal relations, as per London's (2019) worry. In fact, making an overall system intelligible, including the people that are part of it, may require non-causal explanations, such as functional or hermeneutical explanations and those found more widely in the social sciences. In this way, the machine learning-based decision-making system can be made intelligible, in terms of its goals and objectives, and accountability can be demanded of the entire system, which includes the people specifying those goals and objectives.

Alternatively, perhaps in a situation like the trade-off described above, we should step back from the epistemological sense and focus instead on the ethical sense of explicability, establishing a clear line of accountability, such as holding those who are specifying the goals and objectives accountable. But more generally, the principle is not intended as a standalone principle. It would still need to be balanced with other acceptable principles, such as the principle of beneficence – how can we best capitalise on AI technologies to ensure the well-being of people? – or, even, the principle of justice, concerning the fair distribution of goods or what constitutes fair compromises to ensure that, say, access to health services is available to all. This allows variability in what is demanded of a particular system with particular goals and within a particular context.

A second problem, a problem of compromise, focuses more on the ethical sense of the principle of explicability, and the concomitant demand for accountability. This is the problem that a demand for accountability could plausibly limit progress in a field where African nations and research institutions could be firmly entrenched among world leaders. The development of AI and related technologies promises to tackle African-specific problems that can aid in social and economic development, can create jobs, and is an arena where African researchers are already increasingly active. Indeed, the growth and appetite for events such as the Deep Learning Indaba, Data Science Africa and Data Science Nigeria shows that there is interest in upskilling in this direction. Imposing lines of accountability could result in onerous regulatory constraints on an industry we want to encourage, with parties becoming less willing to pursue potentially fruitful but risky research or implementation.

This problem is speculative and overlooks that the principle of explicability does not state what the accountability requirements are, just that we establish lines of accountability. Like the previous problem, this allows contextual sensitivity in devising recommendations from the principle, where that contextual sensitivity may consider different cultural norms regarding what needs to be explained, to whom, when and to what degree, but also must consider the cost-benefit ratio of

potential regulations. Further, and as discussed with the previous problem, devising recommendations from the principle of explicability would work in tandem with other principles. For instance, justice might require differential treatment for how research is conducted in Africa, to target economic and historical imbalances between African nations and centres in the developed world. The principle itself is not obviously problematic but we have to take care with how it is applied.

Further, if Africa is to capitalise on the progress it is making with growing AI research across the continent, we do still need to ensure that it does so in a way compatible with the values and needs of those living there. In order to actually do the computational research, we will often have to address requirements that are closely related to those put forward by the principle of explicability, such as by tackling the value alignment problem. Tackling the value alignment problem requires that we identify our own objectives; that is, we must go some way towards answering the question 'how does it work?' in order to make it, the system, work for what it is designed for. If this is the case, the research and technology that we want to flourish and advance will often depend on addressing questions that the principle of explicability raises, such as demanding intelligibility. In also requiring accountability for the systems, we would be requiring accountability for something that would have to be done, at least to some extent, in order to advance the research itself.

These two problems, as we have seen, can be dealt with by highlighting that the principle of explicability would, ideally, be applied in tandem with other principles and applying it requires contextual sensitivity, not just to ensure that values are aligned with a society's values, but also to ensure that computational challenges are themselves addressed. These problems do not show that the principle of explicability itself is problematic or irrelevant.

5. Closing thoughts on who is accountable for how a decision-making system works

We have proposed that the principle of explicability, when applied in the epistemological sense to typical areas of machine learning research, requires identifying objectives and goals for a system that cohere with those of a given society in which the system will operate. But what implications does this have for the ethical sense of the principle and the question of who, exactly, should be held accountable for how such a decision-making system works? A third potential problem could arise here, to do with demandingness: the principle of explicability as we have developed it may be too demanding on researchers in a developing field in Africa who are frequently dependent on international input. Holding those researchers to account would be unfair. To address this problem,

we tentatively propose that the demands of explicability require a division of labour, and as a result accountability could in fact be diffuse.

Suppose that the machine learning researchers based in Africa and developing a system to be implemented in an African context are to be held accountable for how the system works. As part of the demand for explicability, in the epistemological sense, we have argued that objectives and goals of the system need to be identified. But who should identify these objectives and goals?

Identifying the goals, objectives and underlying values of a society is not a straightforward matter and not something one can simply consult a rulebook for. In South Africa, for instance, it is officially required that vehicles yield right of way to pedestrians crossing at a pedestrian crossing (Department of Transport 2012). In practice, however, this seldom happens and stopping at a pedestrian crossing can surprise other vehicles on the road. Simply consulting the rulebook would not prepare anyone, person or machine, for actual driving.

The machine learning researchers, however, are technical experts, not necessarily experts in identifying the goals, objectives and underlying values of a society with which their system's goals and objectives need to be aligned. Further, they may be contributing to global work or be part of an international research team, such as by working at one of IBM's research labs in South Africa or Kenya, or Google's research lab in Ghana, or be funded by international bodies like Google and Facebook, who fund students pursuing the African Masters in Machine Intelligence in Kigali, Rwanda. Yet, demanding that they make goals and values explicit and then holding the African-based researchers accountable for a system that is not entirely in their hands would be placing an onerous and unfair burden on them.

Identifying the goals, objectives and underlying values of a society, as those working on the ethical design of technology already emphasise (see, for instance, Crawford & Calo 2016; Crawford 2017; Friedman, Henry & Borning 2017; Sloane & Moss 2019), needs to draw on a wider body of stakeholders, which includes those who are experts on the values and goals of a given society, such as researchers in the social sciences and humanities, and even members of society themselves. This would obviously have to be within reason, as not just any lay person will be knowledgeable, nor should they be held accountable for something over which they have no knowledge or control. However, even with bringing in the expertise of a range of people, technical researchers may still shoulder a higher degree of burden because of the fact that these researchers must acknowledge that other players need to be involved and consulted, not as a matter of courtesy or annoyance, but as central to advancing the actual research. Alternatively, organisations driving research should be

required to engage a diversity of relevant experts to ensure that the epistemological sense of explicability is met, and be held accountable if they fail to do so. It is those involved in or driving the actual research who are in a position to ensure that a range of interests and values are acknowledged and, ideally, addressed in both local and international research, or that international research is not uncritically implemented across a range of differing contexts.

Ensuring that relevant experts from a range of backgrounds are engaged speaks in favour of promoting interdisciplinary research and societal engagement as a matter of necessity, not simply for ethical considerations but for advancing the research itself. Luckily, the value of interdisciplinary and multi-stakeholder engagement is already recognised in the various centres and initiatives being set up in Africa, such as the Centre for Artificial Intelligence Research (CAIR) and the South African Affiliate Centre of the C4IR. Applying a principle of explicability in an African context that recognises the necessary involvement of a range of actors, a kind of division of labour for addressing the epistemological sense of explicability, could thus generate diffuse patterns of accountability when addressing the ethical sense of explicability. What exactly this would entail in terms of recommendations, and whether accommodating a diffuse notion of accountability is feasible on the ground, however, is beyond the scope of this paper. Nevertheless, adopting and applying a principle of explicability in an African research context should aim to address these complexities.

In closing, then, we have argued that existing principles and frameworks for the development of Good AI should not be adopted uncritically into an African research context. We thus took initial steps for critically assessing one such framework, that of the AI4People report, by addressing whether the AI-specific principle of explicability should be applied in an African context. We argued that, when designing a decision-making system making use of some form of machine learning, an approach that requires adhering to a principle of explicability in both an epistemological sense (of 'how does it work?') and an ethical sense (of 'who is responsible for how it works?') not only contributes to the responsible and thoughtful development of AI that is sensitive to African interests and needs, but can also advance tackling some of the computational challenges in machine learning research. The principle thus should be adopted in an African context. Adopting the principle, however, requires that African researchers and societies, as well as organisations driving research, ensure that values are aligned, and doing so requires the involvement of a range of knowledgeable stakeholders.

Acknowledgements

We would like to thank participants at the Third CAIR Symposium on AI Research and Society, held at the University of Johannesburg in March 2019, for feedback and discussion on an earlier version of the paper. We would also like to thank the two reviewers and editors for this journal for their comments.

Conflict of interests

Benjamin Rosman is one of the founders and organisers of the Deep Learning Indaba that we mention as an example of the growth of machine learning across Africa.

References

Abbott, J. & Martinus, L. (2019). Benchmarking neural machine translation for southern African languages. *Proceedings of the 2019 Workshop on Widening NLP.*

Andoh, C. (2011). Bioethics and the challenges to its growth in Africa. *Open Journal of Philosophy*, 1(2), 67–75.

Asilomar AI Principles. (2017). https://futureoflife.org/ai-principles. Accessed 29 May 2019.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The moral machine experiment. *Nature,* 563(7729), 59.

Barugahare, J. (2018). African bioethics: Methodological doubts and insights. *BMC Medical Ethics*, 19(98), https://doi.org/10.1186/s12910-018-0338-6.

Beauchamp, T. L. & Childress, J. F. (2012). *Principles of Biomedical Ethics*. 7th edn. Oxford: Oxford University Press.

Behrens, K. (2013). Towards an indigenous African bioethics. *The South African Journal of Bioethics and Law*, 6(1), 32–35.

Biko, S. (2002). *I Write What I Like*: *Selected writings*. Chicago: University of Chicago Press.

Bostrom, N. (2003). Ethical issues in advanced Artificial Intelligence. In I. Smit et al. (eds), *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, volume 2 (pp. 12-17). International Institute of Advanced Studies in Systems Research and Cybernetics.

Brown, B. J., Przybylski, A. A., Manescu, P., Caccioli, F., Oyinloye, G., Elmi, M., Shaw, M. J., et al. (2019). Data-driven malaria prevalence prediction in large densely-populated urban holoendemic sub-Saharan West Africa: Harnessing machine learning approaches and 22-years of prospectively collected data." *arXiv preprint arXiv:1906.07502*.

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR*, 81, 77-91.

Chukwuneke, F. N., Umeora, O. U. J., Maduabuchi, J. U. & Egbunike, N. (2014). Global bioethics and culture in a pluralistic world: How does culture influence bioethics in Africa? *Annals of Medical and Health Sciences Research*, 4(5), 672-675.

Crawford, K. & Calo, R. (2016). There is a blind spot in AI research. *Nature*, 538, 311-313.

Crawford, K. (2017). The trouble with bias. *NIPS 2017 Keynote Address*. https://www.youtube.com/watch?v=fMym_BKWQzk. Accessed 19 August 2019.

Deep Learning Indaba. (2019). *Together We Build African AI: Outcomes of the 2nd Annual Deep Learning Indaba*. http://www.deeplearningindaba.com/uploads/1/0/2/6/102657286/annualindaba2018report-v1.pdf. Accessed 1 March 2020.

Department of Transport. (2012). *SA Learner Driver Manual: Rules of the road.* Pretoria: SA Department of Transport.

Doshi-Velez, F. & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *ar*Xiv.*org.* https://arxiv.org/abs/1702.08608. Accessed 29 May 2019.

European Group on Ethics in Science and New Technologies. (2018). *Statement on Artificial Intelligence, Robotics and 'Autonomous Systems'*. European Commission. http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf. Accessed 29 May 2019.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People - An ethical framework for a Good AI Society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28, 689–707.

Friedman, B., Hendry, D. & Borning, A. (2017). A survey of value sensitive design methods. *Foundations and Trends in Human-Computer Interaction*, 11(23), 63-125.

Gbadegesin, S. (1991). *African Philosophy: Traditional Yoruba philosophy and contemporary African realities.* New York: Peter Lang.

Gyekye, K. (1987). *An Essay on African Philosophical Thought: The Akan conceptual scheme*. Cambridge: Cambridge University Press.

Hellsten, S. (2006). Global bioethics: Utopia or reality? *Developing World Bioethics*, 8(2), 70–81.

House of Lords Artificial Intelligence Committee. (2018). *AI in the UK: Ready, willing and able?* UK Parliament. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf. Accessed 29 May 2019.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically Aligned Design: A vision for prioritising human well-being with autonomous and intelligent systems* (first edition) https://ethicsinaction.ieee.org. Accessed 29 May 2019.

Kroll, J. A. (2018). The fallacy of inscrutability, *Philosophical Transactions A*, 276(20180084).

Landhuis, E. (2017). Tanzania gears up to become a nation of medical drones, *npr.org*. https://www.npr.org/sections/goatsandsoda/2017/08/24/545589328/tanzania-gears-up-to-become-a-nation-of-medical-drones. Accessed 28 February 2020.

London, A. (2019). Artificial Intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, 49(1), 15–21.

Mbembe, A. *On The Postcolony*. Berkeley: University of California Press.

Mbiti, J. (1990). *African Religions and Philosophy (2nd Edition)*. Heinemann.

Microsoft. (2019). *Artificial Intelligence for Africa: An opportunity for growth, development and democratisation*. https://info.microsoft.com/ME-DIGTRNS-WBNR-FY19-11Nov-02-AIinAfrica-MGC0003244_01Registration-ForminBody.html. Accessed 29 May 2019.

Montreal Declaration for a Responsible Development of Artificial Intelligence. (2017). https://www.montrealdeclaration-responsibleai.com/the-declaration. Accessed 29 May 2019.

Moodley, K. (2007). Microbicide research in developing countries: Have we given the ethical concerns due consideration? *BMC Medical Ethics*, 8(1), doi:10.1186/1472-6939-8-10.

Murove, M. F. (2005). African Bioethics: An explanatory discourse. *Journal for the Study of Religion*, 18(1), 16–36.

Partnership on AI. (2018). *Tenets*. https://www.partnershiponai.org/tenets. Accessed 29 May 2019.

Rakotsoane, F. & Van Niekerk, A. (2017). Human life invaluableness: An emerging African bioethical principle. *South African Journal of Philosophy*, 36(2), 252–262.

Ramose, M. *African Philosophy Through Ubuntu (Revised Edition)*. Harare: Mond Books Publishers.

Selbst, A. D. & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4), 233–242.

Senghor, L. (1988). *Ce Que Je Crois*. Paris: Grasset.

Sloane, M. & Moss, E. (2019). AI's social sciences deficit. *Nature Machine Intelligence*, 1, 330–331.

Steyn, L. (2018). Watch your job, the bots are coming. *Mail & Guardian*, 8 December. https://mg.co.za/article/2017-12-08-00-watch-your-job-the-bots-are-coming. Accessed 29 May 2019.

United Republic of Tanzania Ministry of Health and Social Welfare. (2015). *Health Sector Strategic Plan July 2015-June 2020*. http://www.moh.go.tz/en/strategic-plans. Accessed 28 February 2020.

wa'Thiongo, N. (1986). *Decolonising the Mind*. James Curry.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics*, 2(6), doi:10.1126/scirobotics.aan6080.

Winfield, A. F. T., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions A*, 376(20180085).

World Health Organisation (WHO). 2020. Guidelines on reproductive health research and partners' agreement. *Sexual and Reproductive Health*. https://www.who.int/reproductivehealth/topics/ethics/partners_guide_serg/en/. Accessed 21 February 2020.