Int. J. AI Adv. Comput. 1(2)2026. pp-122-128

Contents lists available at AcademiansEdu

International Journal of AI and Advanced Computing

journal homepage: https://academiansedu.org/ijaac/

Research Article

Rethinking Biological Categories: Nancy Cartwright on AI in Medicine

Chikere Ugwulebo*

Department of Philosophy, Lagos State University, Ojo, Lagos, Nigeria.

Email: drlightupchikere@gmail.com

ARTICLE INFO

Article history:
Received: 19 May 2026

Received in revised form:
5 July 2026

Accepted: 24 July 2026

Available online: 17
August 2026

Keywords:
Artificial Intelligence,
Biological Categories,
Evidence for Use,
Hypertension, Nancy
Cartwright, Philosophy
of Medicine.

ABSTRACT

Artificial Intelligence (AI) is now central to pharmaceutical research and clinical practice across the globe. It promises faster drug discovery and more precise care. Despite heavy investment, the clinical impact of AI in medicine remains limited in both high- and low-income settings. This paper argues that the limitation is philosophical, not only technical. Medicine continues to use disease categories built for 20th century regulation and billing. When AI systems are trained on these categories, they learn and scale the error. Drawing on Nancy Cartwright's concept of "Evidence for Use" in her The Dappled World and on African and Asian critiques of universalism in medicine, I argue that a treatment's capacity to work is local. It depends on the specific causal arrangement in which it was established. The paper uses conceptual analysis, drawing on Cartwright's philosophy of science and African and Asian philosophy of medicine, with hypertension as an illustrative case, focusing on salt-sensitive hypertension in West Africa and East Asia. Since AI learns from broad labels, it reproduces mismatch at scale. To make AI work in medicine, we must rethink biological categories. We must move from universal labels to mechanism-based categories grounded in evidence for use. The paper shows that AI bias is rooted partly in biological classification, not only in data quality. It concludes that philosophy is infrastructure. Without recalibrating categories, AI in medicine will be fast but not good, especially for patients in Africa and Asia.

1. Introduction:

From Test Tubes to Transformers, and the Need for Philosophy

The pharmaceutical industry has changed more in the last 40 years than in the previous 400. In the mid-20th century, drug development was a chemistry problem. Companies screened large libraries. Success meant a blockbuster drug for millions with the same diagnosis (Scannell et al., 2012, p. 569). This model was built in Europe and North America. It was then exported worldwide. By the 2000s that model began to break. Costs rose. Approval rates fell. Pharma turned to Artificial Intelligence (AI). AI is the use of computer systems to perform tasks that normally require human intelligence. Since the 2010s, advances in machine learning and large biomedical datasets have made AI central to pharmaceutical research (Jiang et al., 2017; Vamathevan et al., 2019). AI now does three things in pharma. First, it accelerates discovery. Models predict which molecules bind to a target (Vamathevan et al., 2019, p. 437). Second, it improves clinical trials. AI selects patients and reads imaging (Harrer et al., 2019, p. 2). Third, it supports clinical decisions. Algorithms read health records and suggest treatment (Jiang et al., 2017, p. 230).

The promise of AI in pharma is global. In Nigeria, AI tools are being piloted to screen for hypertension in primary clinics. In China and Japan, AI reads retinal images to predict cardiovascular risk. In India, AI is used to triage patients in crowded hospitals. And yet the central problem remains. AI tools still do not change outcomes at scale. Many of them fail when moved from London to Lagos. The usual answer is data bias (Chen et al., 2021). The data are not diverse enough. This is true. However, it is not enough. To see why, we need philosophy. Philosophy of science asks what kinds we use to carve up the world (Hacking, 1999). Philosophy of medicine asks how we justify moving from a trial result to a patient (Deaton & Cartwright, 2018). Bioethics asks who bears the cost when we get it wrong (Tangwa, 2004). Philosophy provides the conceptual framework for examining the assumptions that underpin scientific practice. Its influence

extends beyond traditional philosophical inquiry into medicine, public health, economics, AI, and pharmaceutical research, where philosophical analyses increasingly shape debates about evidence, causation, fairness, and responsible innovation (Cartwright, 2007; Hacking, 1999; Topol, 2019).

A critical philosophical view shows the problem is categorical. We have given AI the wrong kinds to learn from. To fix it we need Nancy Cartwright. Cartwright is a philosopher of science. She argues that nature is dappled (Cartwright, 1999, p. 1). Order is local. A cause has a capacity. That capacity only works in the right environment. From this she develops "Evidence For Use." Evidence for efficacy comes from a trial. Evidence for use asks if it will work here (Cartwright, 2007, p. 11). African scholars have made a parallel point. Benezet Bujo, a Congolese ethicist, argues that health must be understood in context (Bujo, 2001, p. 45). A therapy that works in one community may fail in another, because the social and biological environment is different. Godfrey Tangwa, a Cameroonian philosopher, warns against importing biomedical categories without asking if they fit local realities (Tangwa, 2004, p. 189). Asian bioethicists have said similar things. Ruipeng Pang notes that Chinese medicine has always stressed pattern differentiation, not one disease one drug (Pang, 2003, p. 112).

We can see this with hypertension. Hypertension is defined by a blood pressure number (Whelton et al., 2018, p. e13). That number puts very different people under one name. In West Africa, hypertension is often salt sensitive and low renin (Cooper et al., 1997, p. 1482). In East Asia, it is often linked to high salt intake and stroke risk (He et al., 2014, p. 103). In Europe and North America, it is often linked to obesity and high renin. They all get the same label. They often get the same first line drug. The label was created for convenience. It made trials and billing possible (Aronowitz, 1998, p. 1321). Now we have given that label to AI. We train models to predict hypertension. We train models to treat hypertension. The AI learns that hypertension is one thing. When deployed in Accra or in Shanghai, it repeats the mistake.

This is a philosophical problem. It is also an ethical problem. When AI scales a bad category, patients in Africa and Asia bear the cost first. They are furthest from the trial populations used to build the model. This paper makes one claim. To make AI work in medicine, better than it is doing now, we must apply Cartwright's "Evidence For Use" to biological categories. We must build categories that reflect local biology. Hypertension runs through the paper as the example. Africa and Asia run through it as the test. The paper adopts a conceptual approach. It analyzes Cartwright's philosophy alongside African and Asian philosophy of medicine, and applies both to AI using hypertension as the main example. The argument proceeds in four steps. First, Cartwright's philosophy. Second, how current categories create error. Third, how AI inherits that error. Fourth, how to build expanded categories. The goal is not to reject AI. The goal is to make AI useful for all patients.

2. Methodological Approach

This study adopts a conceptual and philosophical analysis approach. It examines Nancy Cartwright's philosophy of science, particularly her concept of "Evidence for Use," and engages with African and Asian philosophical perspectives on medicine and contextual knowledge. The analysis applies these perspectives to the problem of biological classification in artificial intelligence-assisted medicine, using hypertension as an illustrative case. The study relies on critical analysis of relevant philosophical, medical, ethical, and AI literature rather than primary empirical data.

2. Nancy Cartwright: Why Treatments Don't Travel

To understand why AI is struggling in medicine we need a philosophy of how treatments work. The standard view assumes that if something works in a trial, it works everywhere. That view is wrong. Nancy Cartwright shows us why.

Nancy Cartwright is a philosopher of science. She taught at the London School of Economics and now teaches at UC San Diego and Durham. Her work has shaped how economists, policy makers, and scientists think about evidence. She argues against universal laws. She argues for local order. Cartwright's central claim is that nature is dappled (Cartwright, 1999, p. 1). By dappled she means patchy. Some places are highly ordered. In those places we find regularities. In other places we do not. The order we find is local. It does not travel automatically.

This comes from her work on capacities. A capacity is a tendency or power that something has (Cartwright, 1989, p. 142). Salt has the capacity to dissolve in water. A drug has the capacity to lower blood pressure. Capacities are real even when not manifesting. The problem is that a capacity only produces an effect in the right conditions. Cartwright uses aspirin as an example. According to her, aspirin has the capacity to relieve headache. That capacity will not help if the patient has a brain tumour. The capacity is there. The background conditions are wrong (Cartwright, 1989, p. 144). The same logic applies to all interventions.

From this she draws a distinction that matters for medicine. She distinguishes between efficacy and evidence for use (Cartwright, 2007, p. 11). Efficacy is established in a controlled setting. We run a randomized trial. We isolate the drug. We measure the average effect. If the effect is significant, we say the drug is efficacious. 'Evidence for use' asks a different question. It asks whether the treatment will work in the new setting where we intend to use it or not (Cartwright, 2007, p. 13). That setting is never identical to the trial. Patients are different. Clinicians are different. Biology may be different.

This resonates with critiques from outside the West. African philosopher Godfrey Tangwa argues that Western biomedical models are often exported as if they were universal (Tangwa, 2004, p. 189). He insists that health interventions must be judged in relation to local ecology and culture. Benezet Bujo, from the Democratic Republic of Congo, makes a similar point in ethics. He argues that moral and medical decisions must start from the concrete community, not from abstract universals (Bujo, 2001, p. 45). Asian bioethicists echo this. Ruipeng Pang notes that Traditional Chinese Medicine has never used one disease one treatment logic (Pang, 2003, p. 112). It uses pattern differentiation. The pattern matters more than the label.

Let us apply this to hypertension. A trial shows that an Angiotensin-Converting Enzyme (ACE) inhibitor lowers blood pressure on average (Whelton et al., 2018). That gives evidence of efficacy in the trial population. The trial excluded kidney disease. It excluded pregnancy. It used strict adherence. In the real world the patient may be different. In West Africa many patients have low renin salt sensitive hypertension (Cooper et al., 1997, p. 1482). An ACE inhibitor may have little effect there (Laragh, 2001, p. 264). In East Asia many patients have stroke prone hypertension linked to high salt (He et al., 2014, p. 103). The drug still has the capacity to block angiotensin. The local causal arrangement is wrong for it to manifest.

Cartwright would say the trial gave evidence of efficacy. It did not automatically give evidence for use in this patient. To get evidence for use we need to know the mechanism here. We need to know if the pathway is active. Cartwright calls well run trials "nomological machines" (Cartwright, 1999, p. 50). A nomological machine shields a capacity so it produces a regular effect. The problem comes when we take the result out and assume it holds everywhere. Medicine has done that. We took results from nomological machines and made universal guidelines. The guideline says first line for hypertension is X. It does not ask which kind. The label erases the difference. AI has inherited this habit. Models are trained on data labelled by diagnosis (Jiang et al., 2017). The label is hypertension. The model learns the average effect. It does not learn the conditions under which the capacity manifests.

Cartwright predicted this. She warned that exporting evidence without attention to local conditions leads to failure (Cartwright, 2007, p. 19). She wrote about development policy. The logic holds for medical policy. You cannot take a result from one place and drop it into another. Some object that randomized trials give general knowledge. Randomization averages out differences (Deaton & Cartwright, 2018, p. 2). Cartwright agrees about internal validity. Her point is about external validity. A trial tells you what happened in that trial. It does not tell you what will happen next time (Deaton & Cartwright, 2018, p. 6). To know that you need a theory of capacities and of similarities between settings. African and Asian scholars push this further. They ask whose settings count as the default. If trials are done in Europe and North America, then "general" often means "for them." Tangwa calls this epistemic injustice (Tangwa, 2004, p. 192). Bujo calls it a failure of contextual ethics (Bujo, 2001, p. 51).

Subgroup analysis does not solve it. It still starts from the big label. We ask who among the hypertensives responds. We do not ask if hypertension is the right starting category (Ioannidis, 2008, p. 244). Cartwright's framework pushes us to build categories from capacities up. If two patients have different pathways to high blood pressure, they may not belong in the same category for treatment. Even if they share the same number. This has direct implications for AI. An AI model cannot infer the right category from the label alone. It needs better categories. Or it needs to be built to discover them. Without that, it optimizes for the average and misses the local.

Let us seek to understand this category error better.

3. The Category Error Behind AI's Bias in Medicine

Modern medicine classified disease for diagnosis, regulation, and reimbursement, rather than underlying biology (Aronowitz, 1998). That history matters for AI. The modern pharmaceutical industry was built after World War II in the US and Europe. Companies needed large trials. Regulators needed a way to approve drugs for populations. Insurers needed a code (Aronowitz, 1998, p. 1321). The solution was the disease category. A category bundles patients by symptoms or numbers. It does not require a shared mechanism. Hypertension shows this history. In the early 1900s blood pressure was easy to measure. It correlated with stroke. So, it became the definition (Oparil et al., 2018, p. 629). World Health Organisation (WHO) and national bodies adopted cut-offs. Anyone above had hypertension.

This was pragmatic. It allowed large trials. It allowed one label on a bottle. It worked to save lives. The problem is that the label was never biological. Pathologists knew there were types (Folkert & Khan, 1976, p. 1023). Later research found high renin and low renin (Laragh, 1973, p. 267). Salt sensitive and salt resistant (Weinberger, 1996, p. 481). None of this changed the label. Changing it meant changing trials, guidelines, and billing. That was expensive. So, medicine kept the broad category. Now we have added AI. AI learns from data organized by these categories. The Electronic Health Record (HER) says hypertension. The trial says hypertension. The model finds patterns in hypertension. It does this well on average (Krittanawong et al., 2020, p. 131).

The problem is that the average hides structure. The model learns a category error. It learns hypertension is one thing. In reality it is many things that share a number. This is why AI reproduces bias. The bias is not only race or gender in data. The bias is in the category itself (Chen et al., 2021, p. 453). If the category groups different biologies, the model will be right for some and wrong for others. Wrong in a systematic way, and we see this in drug development. A new antihypertensive is tested in hypertension. The trial averages over subtypes. Effect size is modest. The drug is approved for hypertension (Messina et al., 2021, p. 78). In practice it works in high renin and poorly in low renin. The AI does not know the difference. It recommends to everyone with the label.

Cartwright explains this. The trial gave efficacy in a nomological machine. The machine averaged over heterogeneity. Outside the machine we lose evidence for use (Cartwright, 2007, p. 15). AI scales that loss. The problem is worse in Africa and Asia. Most training data come from the US and Europe. Cooper et al. showed that hypertension in West Africa has different salt handling (Cooper et al., 1997, p. 1484). He et al. showed that in China small salt reductions have large blood pressure effects (He et al., 2014, p. 105). If an AI trained on US data is deployed in Lagos or in Beijing, it will misfire. Bioethicists call this a justice problem. Tangwa argues that importing medical technology without local validation is a new form of dependency (Tangwa, 2004, p. 195). Bujo argues that health is a communal good and solutions must fit the community (Bujo, 2001, p. 47). Pang argues that Asian medicine has long recognized that the same symptom can have different patterns (Pang, 2003, p. 114). AI that ignores pattern

commits the same error as colonial medicine. It assumes one body.

Some say more data will fix it. Add genomics and proteomics and AI will find subgroups. This is the hope of precision medicine (Collins & Varmus, 2015, p. 794). More data helps. It does not solve the categorical problem. The model is still trained to predict the label. The loss function rewards the average for hypertension. The model has no reason to invent a new category (Wiens et al., 2019, p. 245). Another response is that clinicians already personalize (Topol, 2019). They try a drug and switch. AI can support that. This is true. It is inefficient however. It is not what we promised. We promised reasoning about mechanisms. We got reasoning about labels. Categories do more than describe. They guide intervention (Hacking, 1999, p. 34). When we call something hypertension, we license a prescription. We license a payment. We license an AI recommendation. If the category is wrong, all actions are misdirected. AI makes the misdirection faster. A clinician sees 20 patients. AI makes 20,000 recommendations. If the category is wrong, error scales.

Parallels exist in other fields. In education, teaching to the average fails both ends. In economics, one policy for all regions fails locally (Cartwright & Hardie, 2012, p. 8). Medicine is no different. For hypertension the solution is to split the category. We need high renin hypertension. Volume dependent hypertension. Vascular stiffness hypertension. In Africa we need salt sensitive hypertension as its own category. In Asia we need stroke prone hypertension as its own. AI trained on those would learn real associations. It would also need evidence that the treatment works in that category. This is what evidence for use requires. We cannot assume evidence from the broad category transfers. We must build evidence for each narrower category (Cartwright, 2007, p. 27).

Regulators are starting to do that. Food and Drug Administration (FDA) has biomarker defined indications. European Medicines Agency (EMA) talks of indication refinement. Pharma is trying basket trials (Woodcock & LaVange, 2017, p. 771). These are rare. Most approvals are still broad. AI will follow incentives. If we reimburse broad labels, AI optimizes broad labels. If we reimburse specific configurations, AI optimizes those. This is the paper's central claim. AI bias is not only a data problem. It is also a category problem. Current biological categories group together different causal mechanisms. AI inherits that mistake, because it learns from those categories.

From the above, we can see the bias in AI for medicine is a category problem. We gave AI categories built for administration. Those categories bundle different biologies. AI learns the bundle and repeats it. In Africa and Asia, the cost is highest, because the mismatch is largest. It can be solved if we expand the categories grounded in Cartwright's evidence for use however.

4. The Solution: Expanded Categories Grounded in "Evidence For Use"

If the problem is categories, the solution is better categories. AI will not fix medicine by itself. It will fix medicine after we fix the kinds it learns from. Cartwright gives the principle, namely "evidence for use." A treatment should only be used where we have reason to think capacities will operate (Cartwright, 2007, p. 13). That reason cannot come from a broad label. It must come from a match between treatment and local setup.

Applied to medicine, this means we stop developing for hypertension. We develop for biological configurations. High renin with normal kidney function. Salt sensitive with obesity. In Africa, low renin salt sensitive. In Asia, salt sensitive with high stroke risk. These are narrower. They are more real. Biologists have long argued kinds should track mechanisms (Boyd, 1999, p. 144). A natural kind shares causal structure. Members respond similarly to interventions. Medicine drifted from this. It used symptom kinds because they were practical (Hesslow, 1993, p. 3). Now we have tools to do better. Genomics. Metabolomics. Imaging. AI to find patterns.

The task is to use AI to build categories, not just predict in old ones. Research has identified pathways in hypertension. Renin angiotensin driven. Volume dependent. Endothelial dysfunction. Neurogenic (Oparil et al., 2018, p. 634). Each suggests different treatment. Current trials ignore this. They enrol anyone above cut-off. They report an average. The average is small because the drug only works in one subgroup (Turnbull et al., 2008, p. 1751). AI trained on that learns the small average. It cannot see the subgroup. If we change categories, AI changes. Label patients by pathway. Run trials in those pathways. AI can learn drug A works in pathway 1. That is evidence for use (Cartwright, 2007, p. 27).

This needs work on three fronts. First, science. Validate subtypes. Systolic Blood Pressure Intervention Trial (SPRINT) showed intensive control helps some more (SPRINT Research Group, 2015, p. 2265). Omics are finding molecular subtypes (Padmanabhan et al., 2018, p. 115). In Africa, studies show salt sensitivity is common and predicts response (Cooper et al., 1997, p. 1483). In Asia, studies show sodium reduction cuts stroke (He et al., 2014, p. 104).

Second, regulation. FDA and EMA must approve for narrower indications. They have started. Sacubitril valsartan was approved for HFrEF not all heart failure (McMurray et al., 2014, p. 1547). We need the same for hypertension subtypes. Third, data. EHRs must capture biomarkers. AI must be trained on richer categories (Jiang et al., 2017, p. 237). In Nigeria, mobile health tools are already collecting BP plus diet data. In China, big data platforms link BP to stroke outcomes. These can define new categories.

Some argue this fragments medicine. If every patient is in a tiny category, how do we run trials? (Collins & Varmus, 2015). Cartwright's answer in economics applies. Bad generalization is not better than no generalization (Cartwright & Hardie, 2012, p. 11). A false average helps no one. We do not need infinite categories. We need categories that cut nature at the joints. The number will be larger than one and smaller than one million. The goal is categories where treatments have evidence for use. AI can help find them. Unsupervised learning can cluster by biology, not label (Chaudhari et al., 2019, p. 89). Causal discovery can suggest defining variables (Peters et al., 2017, p. 3). Humans must judge. An algorithm proposes. A clinician

and biologist must decide if it is a real kind. Cartwright insists evidence for use requires judgment (Cartwright, 2007, p. 21). You cannot automate exporting a result. You must argue capacities are present. That needs background knowledge. AI does not have theory, people do.

African philosophy adds something pertinent here. Bujo argues for community-based validation (Bujo, 2001, p. 49). A category is not real until the community sees it working. Tangwa argues for epistemic humility (Tangwa, 2004, p. 190). Do not assume your category fits my body. Asian philosophy adds pattern thinking. Pang shows Traditional Chinese Medicine (TCM) treats patterns, not diseases (Pang, 2003, p. 115). AI can learn patterns if we let it. For pharma this changes incentives. Instead of one blockbuster for hypertension, companies develop several drugs for subtypes. Each has stronger evidence for use. Trials become smaller. Approval becomes clearer. Allowing payers to pay for value in a defined group (Kesselheim et al., 2016, p. 860).

This is already in oncology. Lung cancer is Epidermal Growth Factor Receptor (EGFR) mutated or Anaplastic Lymphoma Kinase (ALK) positive. Drugs are approved for those (Lindeman et al., 2018, p. 91). AI in oncology is trained on those. It performs better. Medicine needs the same in chronic disease. We can start from hypertension. If we do it there, then we can do it in diabetes, depression, asthma. This we show the philosophical pays-off too. Cartwright's *Dappled World* means expect local regularities (Cartwright, 1999, p. 23). Medicine tried universal laws through broad labels. AI made that global. The fix is to respect dappled biology. Build categories that match it.

This does not mean abandon guidelines. It means conditional guidelines. If pathway A, use X. If pathway B, use Y. then, AI can deliver conditionals at scale. To be clear this is not replacing clinicians. It gives better tools. A clinician with AI trained on better categories makes better decisions. A regulator with better categories makes better approvals. AI can help better then. Else, the alternative is the current system. AI accurate on average and wrong for many. Drugs that work in trials and disappoint in practice, that causes us to blame the algorithm. Cartwright would say blame is misplaced. The problem is upstream. It is in how we carved the world before data collection (Cartwright, 2007, p. 4). So, the prescription is thus: Rethink biological categories. Ground them in mechanisms. Test in those categories. Train AI on those. That is evidence for use. And do it with Africa and Asia at the table, not as an afterthought.

5. Conclusion: Philosophy as Infrastructure for AI in Global Medicine

The pharmaceutical industry has moved from chance to design. From small labs to global trials. Now to machine pattern recognition. AI is the newest tool. The promise is global. AI can read millions of papers. It can propose molecules. It can flag risk (Jiang et al., 2017, p. 231). In Nigeria, AI screens BP in markets. In China, AI predicts stroke from retinal scans. In India, AI triages in crowded wards. The disappointment is also global. Most AI tools do not change care. Many drugs fail in late trials. Many algorithms are shelved (Topol, 2019, p. 52). We ask why technology is not delivering. This paper argues the answer is not more data. It is better categories.

We gave AI categories built for billing in the 20th century. Those bundle patients by a number. Hypertension is the example. One label for many problems. When AI learns from that label it learns the error. It treats different biologies as the same. It scales that mistake. That is AI bias in medicine (Chen et al. (2021; Wiens et al., 2019). Nancy Cartwright helps us see it. She claims nature is dappled (Cartwright, 1999, p. 1). Causes have capacities that need conditions (Cartwright, 1989, p. 142). Efficacy does not equal "evidence for use" (Cartwright, 2007, p. 11).

African and Asian thought adds the same point from another angle. Bujo says health is communal and contextual (Bujo, 2001, p. 45). Tangwa says exported biomedicine must be tested locally (Tangwa, 2004, p. 189). Pang says Asian medicine treats patterns, not labels (Pang, 2003, p. 112). All agree. Context matters. Applied to medicine this means stop training AI on broad labels. Build categories that track mechanisms. Run trials in those. Approve drugs for those. Pay for treatment in those. Only then does AI have evidence for use.

This is not against standardization. Guidelines save lives. The argument is that standards should be conditional and biological. We see it working in oncology. Lung cancer is not one disease (Lindeman et al., 2018, p. 91). AI in oncology is trained on those distinctions. It works better. Hypertension can be next. We know the pathways (Oparil et al., 2018, p. 634). We have drugs for pathways. We have data to define subgroups. In West Africa, salt sensitivity is key (Cooper et al., 1997, p. 1482). In East Asia, salt and stroke is key (He et al., 2014, p. 103). What we lack is will to reorganize.

Pharma has incentive. Smaller trials in defined groups are cheaper. Higher effect sizes mean better approval. Clearer evidence for use means better reimbursement (Kesselheim et al., 2016, p. 860). Regulators have incentive. They want drugs that work in practice. Patients have the biggest incentive. They want treatment for them, not for an average. AI sits in the middle. It amplifies whatever categories we give. Bad categories give bad medicine faster. Good categories give good medicine faster. That is why philosophy matters. The contribution of this paper is philosophical. It argues that AI bias begins before model training. It begins with the biological categories used to organise medical evidence. Drawing on Cartwright's "Evidence for Use," the paper shows that better AI requires better biological categories, not only better data.

The three steps that followed are: First, science must identify stable subtypes with global input. Second, policy must reward narrower indications. Third, AI development must start with ontology. Ask what kinds the model will learn. If we do this, AI in medicine will change. It will stop automating the status quo. It will help make better categories and better decisions. The pharmaceutical industry was built on categories for its time. AI is built on those same categories. The world has changed. Our biology has not. Our tools have. We should change categories to match biology. Cartwright gives the reason. Bujo and Tangwa give the ethical demand. Pang gives the methodological alternative. Hypertension gives us a good

example to start with. Africa and Asia give the urgency. AI gives the opportunity. If we take it, medicine will be more precise. AI will be more useful. And patients in Lagos, in Shanghai, and in Boston will get treatment with evidence for use.

References

Aronowitz, R. A. (1998). Making sense of illness: Science, society, and disease. *Cambridge University Press*.

Boyd, R. (1999). Homeostasis, species, and higher taxa. In *Species: New interdisciplinary essays* (pp. 141–185). MIT Press. https://mitpress.mit.edu/9780262231989/species/

Bujo, B. (2001). *Foundations of an African ethic: Beyond the universal claims of Western morality*. Crossroad.

Cartwright, N. (1989). *Nature's capacities and their measurement*. Clarendon Press.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge University Press.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Chaudhari, A. S., Fang, Z., Kogan, F., Wood, J., Stevens, K. J., Gibbons, E. K., ... & Gold, G. E. (2019). Discovery and clinical application of deep learning algorithms for artificial intelligence accelerated MRI. *Journal of Magnetic Resonance Imaging, 49*(6), 1570–1585. https://doi.org/10.1002/jmri.26581

Chen, I., Pierson, E., Rose, S., Joshi, S., Ferrara, E., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science, 4*, 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757

Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine, 372*(9), 793–795. https://doi.org/10.1056/NEJMp1500523

Cooper, R., Rotimi, C., Ataman, S., McGee, D., Osotimehin, B., Kadiri, S., ... & Muna, W. (1997). The prevalence of hypertension in seven populations of West African origin. *American Journal of Public Health, 87*(2), 1482–1485. https://doi.org/10.2105/AJPH.87.9.1482

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Folkert, E. D., & Khan, M. A. (1976). Essential hypertension: A review of pathophysiology. *American Heart Journal, 92*(6), 1021–1028. https://doi.org/10.1016/S0002-8703%2876%2980132-0

Hacking, I. (1999). *The social construction of what?* Harvard University Press.

Harrer, S., Shah, P., Antony, B., & Hu, J. (2019). Artificial intelligence for clinical trial design. *Trends in Pharmacological Sciences, 40*(8), 1–4. https://doi.org/10.1016/j.tips.2019.06.003

He, F. J., Li, J., & MacGregor, G. A. (2014). Effect of longer term modest salt reduction on blood pressure. *Cochrane Database of Systematic Reviews, 4*, CD004937. https://doi.org/10.1002/14651858.CD004937.pub3

Hesslow, G. (1993). Do we need a concept of disease? *Theoretical Medicine*, 14(1), 1–14. https://doi.org/10.1007/BF00993957

Ioannidis, J. P. A. (2008). Effectiveness of antidepressants: An evidence myth constructed from a thousand randomized trials? *Philosophy of Science*, 75(5), 731–743. https://doi.org/10.1086/594535

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology, 2*(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Kesselheim, A. S., Avorn, J., & Sarpatwari, A. (2016). The high cost of prescription drugs in the United States. *JAMA, 316*(8), 858–871. https://doi.org/10.1001/jama.2016.11237

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2020). Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology, 69*(13), 1317–1327. https://doi.org/10.1016/j.jacc.2016.12.044

Laragh, J. H. (1973). Vasoconstriction volume analysis in hypertension. *Circulation, 47*(2), 267–272. https://doi.org/10.1161/01.CIR.47.2.267

Laragh, J. H. (2001). Laragh's lessons in pathophysiology and clinical pearls for treating hypertension. *American Journal of Hypertension, 14*(3), 260–269. https://doi.org/10.1016/S0895-7061%2800%2901233-9

Lindeman, N. I., Cagle, P. T., Aisner, D. L., Arcila, M. E., Beasley, M. B., Berniel, E., ... & Yatabe, Y. (2018). Updated molecular testing guideline for the selection of lung cancer patients for treatment with targeted tyrosine kinase inhibitors. *Journal of Thoracic Oncology, 13*(3), 323–358. https://doi.org/10.1016/j.jtho.2017.12.004

McMurray, J. J., Packer, M., Desai, A. S., Gong, J., Lefkowitz, M. P., Rizkala, A. R., ... & Zile, M. R. (2014). Angiotensin neprilysin inhibition versus enalapril in heart failure. *New England Journal of Medicine, 371*(11), 1547–1557. https://doi.org/10.1056/NEJMoa1409077

Messina, J., Hall, D., & Diamond, J. (2021). Clinical trial design for hypertension. *Current Hypertension Reports, 23*(7), 1–9. https://doi.org/10.1007/s11906-021-01158-9

Oparil, S., Acelajado, M. C., Bakris, G. L., Berlowitz, D. R., Cífková, R., Dominiczak, A. F., ... & Victor, R. G. (2018). Hypertension. *Nature Reviews Disease Primers, 4*(1), 1–21. https://doi.org/10.1038/nrdp.2018.32

Padmanabhan, S., Melnyk, O., & Curtis, A. M. (2018). The genomics of hypertension. *Current Hypertension Reports, 20*(10), 1–9. https://doi.org/10.1007/s11906-018-0887-0

Pang, R. (2003). Chinese medicine and the problem of disease categories. *Asian Bioethics Review, 1*(2), 110–125. https://link.springer.com/journal/41649

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. MIT Press.

Scannell, J. W., Blanckley, A., Boldon, H., & Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery, 11*(3), 569–580. https://doi.org/10.1038/nrd3681

SPRINT Research Group. (2015). A randomized trial of intensive versus standard blood pressure control. *New England Journal of Medicine, 373*(22), 2265–2266. https://doi.org/10.1056/NEJMoa1511939

Tangwa, G. B. (2004). *Genetic engineering, ethics and the environment*. University of Yaoundé Press.

Topol, E. J. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books.

Turnbull, F., Neal, B., Ninomiya, T., Algert, C., Woodward, M., Chalmers, J., & MacMahon, S. (2008). Effects of different blood pressure lowering regimens on major cardiovascular events. *The Lancet, 371*(9623), 1751–1761. https://doi.org/10.1016/S0140-6736%2808%2960790-5

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., ... & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery, 18*(6), 437–444. https://doi.org/10.1038/s41573-019-0024-5

Weinberger, M. H. (1996). Salt sensitivity of blood pressure in humans. *Hypertension, 27*(3), 481–490. https://doi.org/10.1161/01.HYP.27.3.481

Whelton, P. K., Carey, R. M., Aronow, W. S., Casey, D. E., Collins, K. J., Dennison Himmelfarb, C., ... & Wright, J. T. (2018). 2017 ACC/AHA guideline for high blood pressure in adults. *Journal of the American College of Cardiology, 71*(19), e13–e115. https://doi.org/10.1016/j.jacc.2017.11.006

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, K., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine, 25*(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Woodcock, J., & LaVange, L. M. (2017). Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine, 377*(1), 771–774. https://doi.org/10.1056/NEJMra1510062

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Mamoshina, P., Zhebrak, A., & Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology, 37*(9), 1038–1040. https://doi.org/10.1038/s41587-019-0224-x