

Meatspace Press

# Fake AI

Edited by  
Frederike  
Kaltheuner

With contributions by

Razvan Amironesei  
Aparna Ashok  
Abeba Birhane  
Crofton Black  
Favour Borokini  
Corinne Cath  
Emily Denton  
Serena Dokuaa Oduro  
Alex Hanna  
Adam Harvey  
Fieke Jansen  
Frederike Kaltheuner

Gemma Milne  
Arvind Narayanan  
Hilary Nicole  
Ridwan Oloyede  
Tulsi Parida  
Aidan Peppin  
Deborah Raji  
Alexander Reben  
Andrew Smart  
Andrew Strait  
James Vincent

**Fake AI**

Fake AI

Edited by: Frederike Kaltheuner

Publisher: Meatspace Press (2021)

Weblink: [meatspacepress.com](http://meatspacepress.com)

Design: Carlos Romo-Melgar, John Philip Sage and Roxy Zeiher

Copy editors: David Sutcliffe and Katherine Waters

Format: Paperback and pdf

Printed by: Petit. Lublin, Poland.

Paper: Munken Print White 20 - 90 gsm

Set in: Roobert and Times New Roman

Length: 206 pages

Language: English

Product code: MSP112101

ISBN (paperback): 978-1-913824-02-0

ISBN (pdf, e-book): 978-1-913824-03-7

License: Creative Commons BY-NC-SA

Contributors (alphabetically): Razvan Amironesei, Aparna Ashok, Abeba Birhane, Crofton Black, Favour Borokini, Corinne Cath, Emily Denton, Serena Dokuaa Oduro, Alex Hanna, Adam Harvey, Fieke Jansen, Frederike Kaltheuner, Gemma Milne, Arvind Narayanan, Hilary Nicole, Ridwan Oloyede, Tulsi Parida, Aidan Peppin, Deborah Raji, Alexander Reben, Andrew Smart, Andrew Strait, James Vincent

All rights reserved according to the terms of the Creative Commons BY-NC-SA license, excluding any product or corporate names which may be trademarks or registered trademarks of third parties, and are cited here for the purposes of discussion and/or critique without intent to infringe. Discussion of such third party product, corporate or trademarked names does not imply any affiliation with or an endorsement by the rights holder.

The publisher has endeavoured to ensure that any URL for external websites referred to in this book are correct and active at the time of going to press. However, the publisher has no responsibility for the websites and can make no guarantee that these links remain live or that the content is, or will remain, appropriate for all audiences.

**Fake Δι**



## CONTENTS

- 07**      **This book is an intervention**  
Frederike Kalthener
- 19**      **À snake oil, pseudoscience and hype**  
**an interview with Arvind Narayanan**
- 41**      **Cheap À**  
Abeba Birhane
- 53**      **The bodies underneath the rubble**  
Deborah Raji
- 63**      **Who am I as data?**  
Frederike Kalthener
- 77**      **The case for interpretive techniques**  
**in machine learning**  
Razvan Amironesei, Emily Denton,  
Alex Hanna, Hilary Nicole,  
Andrew Smart
- 89**      **Do we need À or do we need Black**  
**feminisms? A poetic guide**  
Serena Dokuaa Oduro
- 97**      **How (not) to blog about an**  
**intelligent toothbrush**  
James Vincent
- 105**      **Learn to take on the ceiling**  
Alexander Reben

- 115**     **Uses (and abuses) of hype**  
Gemma Milne
- 125**     **Talking heads**  
Crofton Black
- 135**     **What is a face?**  
Adam Harvey
- 147**     **Why automated content  
moderation won't save us**  
Andrew Strait
- 161**     **Consolidating power in the name  
of progress: techno-solutionism  
and farmer protests in India**  
Tulsi Parida, Aparna Ashok
- 171**     **When fintech meets 60 million  
unbanked citizens**  
Favour Borokini, Ridwan Oloyede
- 183**     **Algorithmic registers and their  
limitations as a governance practice**  
Fieke Jansen, Corinne Cath
- 193**     **The power of resistance: from  
plutonium rods to silicon chips**  
Aidan Peppin



**This book  
is an**



**intervention**  
Frederike  
Kaltheuner

**Not a week passes by without some research paper, feature article or product marketing making exaggerated or even entirely unlikely claims about the capabilities of Artificial Intelligence (AI). From academic papers that claim AI can predict criminality, personality or sexual orientation, to the companies that sell these supposed capabilities to law enforcement, border control or human resources departments around the world, fake and deeply flawed AI is rampant.**

**The current amount of public interest in AI was spurred by the genuinely remarkable progress that has been made with some AI techniques in the past decade. For narrowly defined tasks, such as recognising objects, AI is now able to perform at the same level or even better than humans. However, that progress, as Arvind Narayanan has argued, does not automatically translate into solving other tasks. In fact, when it comes to predicting any social outcome, using AI is fundamentally dubious.<sup>1</sup>**

The ease and frequency with which AI's real and imagined gains are conflated results in real, tangible harms. For those subject to automated systems, it can mean the difference between getting a job and not getting a job, between being allowed to cross a border and being denied access. Worse, the ways in which these systems are so often built in practice means that the burden of proof often falls on those affected to prove that they are in fact who they say they are. On a societal level, widespread belief in fake AI means that we risk redirecting resources to the wrong places. As Aidan Peppin argues in this book, it could also mean that public resistance to the technology will end up stifling progress in areas where genuine progress is being made.

What makes the phenomenon of fake AI especially curious is the fact that, in many ways, 2020–21 has been a time of great AI disillusionment. *The Economist* dedicated its entire summer *Technology Quarterly* to the issue, concluding that “An understanding of AI's limitations is starting to sink in.”<sup>2</sup> For a technology that has been touted as the solution to virtually every challenge imaginable—from curing cancer, to fighting poverty, predicting

1. Narayanan, A. (2019) How to recognize AI snake oil. *Princeton University, Department of Computer Science*. <https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf>

2. Cross, T. (2020, 13 June) An understanding of AI's limitations is starting to sink in. *The Economist*. <https://www.economist.com/technology-quarterly/2020/06/11/an-understanding-of-ais-limitations-is-starting-to-sink-in>

**criminality, reversing climate change and even ending death—AI has played a remarkably minor role<sup>3</sup> in the global response to a very real challenge the world is facing today, the Covid-19 pandemic.<sup>4</sup> As we find ourselves on the downward slope of the AI hype cycle, this is a unique moment to take stock, to look back and to examine the underlying causes, dynamics, and logics behind the rise and fall of fake AI.**

**Bringing together different perspectives and voices from across disciplines and countries, this book interrogates**

3. Mateos-Garcia, J., Klinger, J., Stathoulopoulos, K. (2020) Artificial Intelligence and the Fight Against COVID-19. *Nesta*. <https://www.nesta.org.uk/report/artificial-intelligence-and-fight-against-covid-19/>

4. Peach, K. (2020) How the pandemic has exposed AI's limitations. *Nesta*. <https://www.nesta.org.uk/blog/how-the-pandemic-has-exposed-ais-limitations/>

the rise and fall of AI hype, pseudoscience, and snake oil. It does this by drawing connections between specific injustices inflicted by inappropriate AI, unpacking lazy and harmful assumptions made by developers when designing AI tools and systems, and examining the existential underpinnings of the technology itself to ask: why are there so many useless, and even dangerously flawed, AI systems?

Any serious writing about AI will have

to wrestle with the fact that AI itself has become an elusive term. As every computer scientist will be quick to point out, AI is an umbrella term that's used for a set of related technologies. Yet while these same computer scientists are quick to offer a precise definition and remind us that much of what we call AI today is in fact machine learning, in the public imagination, the term AI has taken on a meaning of its own. Here, AI is a catch-all phrase used to describe a wide-ranging set of



technologies, most of which apply statistical modelling to find patterns in large data sets and make predictions based on those patterns—as Fieke Jansen and Corinne Cath argue in their piece about the false hope that’s placed in AI registers.

Just as AI has become an imprecise word, hype, pseudoscience, and snake oil are frequently used interchangeably to call out AI research or AI tools that claim to do something they either cannot, or should not do. If we look more closely however, these terms are distinct. Each highlights a different aspect of the phenomenon that this book interrogates.

As Abeba Birhane powerfully argues in her essay, *Cheap AI*, the return of pseudoscience, such as race science, is neither unique nor distinct to AI research. What is unique is that dusty and long discredited ideas have found new legitimacy through AI. Dangerously, they’ve acquired a veneer of innovation, a sheen of progress, even. By contrast, in a wide-ranging interview that considers how much, and how little, has changed since his original talk three years ago, Arvind Narayanan hones in on “AI snake oil”,

explaining how it is distinct from pseudoscience. Vendors of AI snake oil use deceptive marketing, fraud, and even scams to sell their products as solutions to problems for which AI techniques are either ill-equipped or completely useless.

The environment in which snake oil and pseudoscience thrives is characterised by genuine excitement, unchallenged hype, bombastic headlines, and billions of dollars of invest-

ment, all coupled with a naïve belief in the idea that technology will save us. Journalist James Vincent writes about his first encounter with a PR pitch for an AI toothbrush and reflects on the challenges of covering hyped technology without further feeding unrealistic expectations. As someone who used to work as a content moderator for Google in the mid 2010s, Andrew Strait makes a plea against placing too much hope on automation in content moderation.

DF 0.1501

Each piece in this book provides a different perspective and proposes different answers to problems which circle around the shared question of what is driving exaggerated, flawed or entirely unfounded hopes and expectations about AI. Against broad-brush claims, they call for precise thinking and scrupulous expression.

For Deborah Raji the lack of care with which engineers so often design algorithmic systems today belongs to a long history of engineering irresponsibility in constructing material artefacts like bridges and cars. Razvan Amironesei, Emily Denton, Alex Hanna, Andrew Smart and Hilary Nicole describe how benchmark datasets contribute to the belief that algorithmic systems are objective or scientific in nature. The artist Adam Harvey picks apart what exactly defines a “face” for AI.

LG 2.3

A recurring theme throughout this book is that harms and risks are unevenly distributed. Tulsi Parida and Aparna Ashok consider the effects of AI inappropriately applied through the Indian concept of *jugaad*. Favour Borokini and Ridwan Oloyede warn of the dangers that come with AI hype in Nigeria’s fintech sector.

E 6800

Amidst this feverishly hyped atmosphere, this book makes the case for nuance. It invites

readers to carefully separate the real progress that AI research has made in the past few years from fundamentally dubious or dangerously exaggerated claims about AI's capabilities. We are not heading towards Artificial General Intelligence (AGI). We are not locked in an AI race that can only be won by those countries with the least regulation and the most investment. Instead, the real advances in AI pose both old and new challenges that can only be tamed if we see AI for what it is. Namely, a powerful technology that at present is produced by only a handful of companies with workforces that are not representative of those who are disproportionately affected by its risks and harms.

Frederike Kaltheuner is a tech policy consultant and researcher. She is the Director of the European AI Fund, a philanthropic initiative to strengthen civil society.

LDR 0.0389

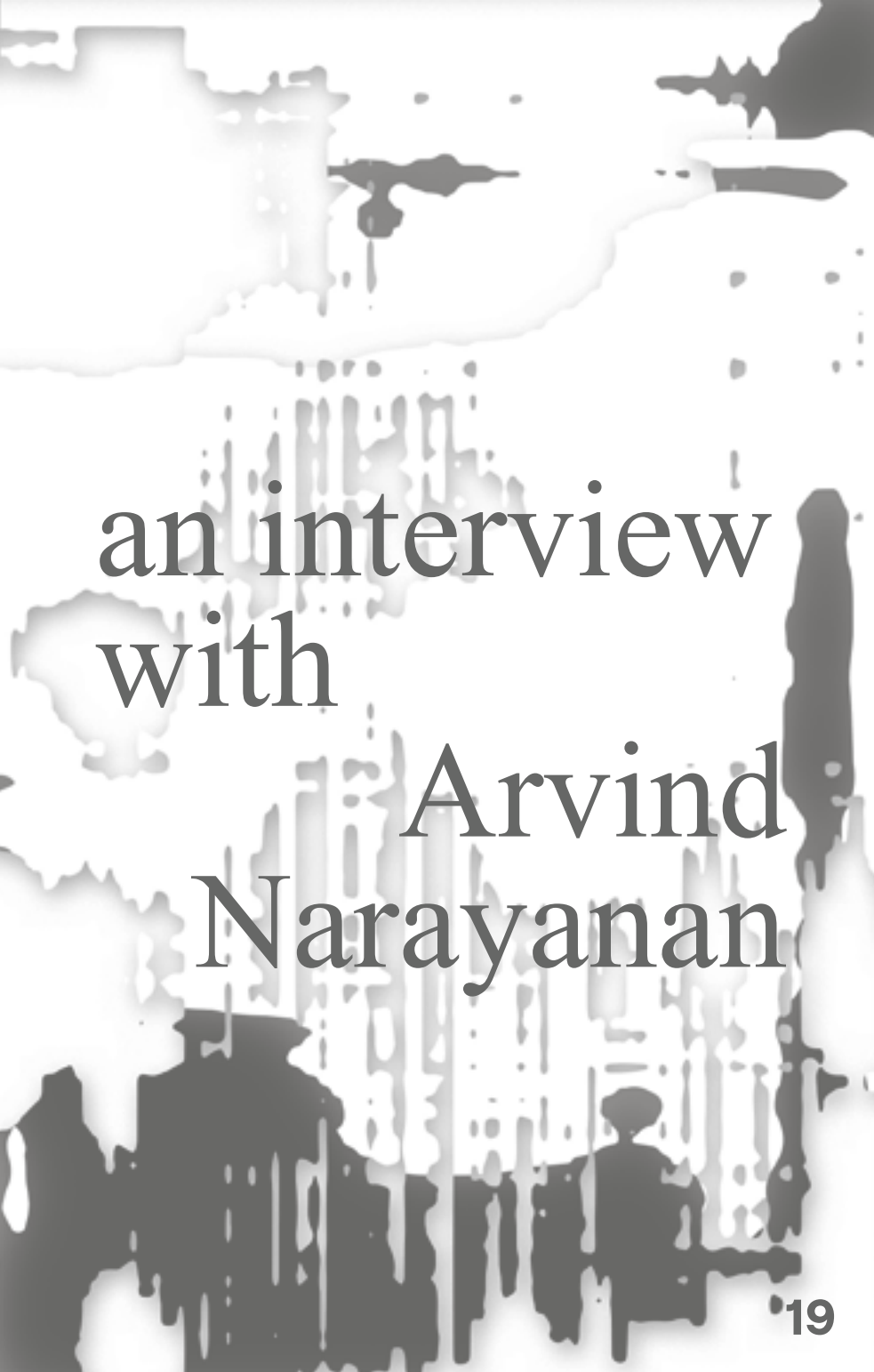
## Acknowledgements

This book was created during a phase of collective late-stage pandemic burnout. I would like to thank the contributors, the publisher, and the designers for their patience with me, for their commitment in the process, and for working under the challenging conditions created by various lockdowns, quarantines, and restrictions.

This project was made possible by the generous support of the Mozilla Foundation through its Tech Policy Fellowship programme. In a moment of extraordinary global disruption, Mozilla, and Janice Wait in particular, placed their trust in a nonlinear process and provided me with the luxury of time and space to think.

A special thanks goes to Katherine Waters. This book would not have seen the light of day, and would certainly have looked very different, without her skilful edits, thoughtful comments, and her friendship.

**AI snake oil,  
pseudo  
science  
and hype**

The background of the page is a black and white photograph showing the silhouettes of people walking on a street. The scene is captured from a low angle, looking down the street. The sky is bright, and the silhouettes of buildings and streetlights are visible in the background. The overall mood is one of a busy, urban environment.

an interview  
with  
Arvind  
Narayanan



The term “snake oil” originates from the United States in the mid 19th century when Chinese immigrants working on the railroads introduced their American counterparts to a traditional treatment for arthritis and bursitis made of oil derived from the Chinese water snake. The effectiveness of the oil, which is high in omega-3 acids, and its subsequent popularity prompted some profiteers to get in on a lucrative market. These unscrupulous sellers peddled quack remedies which contained inferior rattlesnake oil or completely arbitrary ingredients to an unsuspecting public. By the early 20th century, “snake oil” had taken on its modern, pejorative meaning to become a byword for fake miracle cures, groundless claims, and brazen falsehoods.

Much of what is sold commercially as AI is snake oil, says Arvind Narayanan, Associate Professor for Computer Science at Princeton University—we have no evidence that it works, and based on our scientific understanding, we have strong reasons to believe that it couldn’t possibly work. And yet, companies continue to market AI products that claim to predict anything from crime, to job performance, sexual orientation or gender. What makes the public so susceptible to these claims is the fact that in recent years, in some domains of AI research, there has been genuine and impressive progress. How, then, did AI become attached to so many products and services of questionable or unverifiable quality, and slim to non-existent usefulness?

Frederike Kalthener spoke to Arvind Narayanan via Zoom in January 2021. Frederike, from lockdown in Berlin, and Arvind from his office in Princeton.

**FK** Your talk, *How to Recognise AI Snake Oil* went viral in 2019. What inspired you to write about AI snake oil, and were you surprised by the amount of attention your talk received?

**AN** Throughout the last 15 years or so of my research, one of my regular motivations for getting straight into a research topic is when there is hype in the industry around something. That's how I first got started on privacy research. My expertise, the desire for consumer protection, and the sense that industry hype got out of control all converged in the case of AI snake oil. The AI narrative had been getting somewhat unhinged from reality for years, but the last straw was seeing how prominent these AI-based hiring companies have become. How many customers they have, and how many millions of people have been put through these demeaning video

interviews where AI would supposedly figure out someone's job suitability based on how they talked and other irrelevant factors. That's really what triggered me to feel "I have to say something here".

I was very surprised by its reception. In addition to the attention on Twitter, I received something like 50 invitations for papers, books... That had never happened to me before. In retrospect I think many people suspected what was happening was snake oil but didn't feel they had the expertise or authority to say anything. People were speaking up of course, but perhaps weren't being taken as seriously because they didn't have the "Professor of Computer Science" title. That we still put so much stock in credentials is, I think, unfortunate. So when I stood up and said this, I was seen as someone who had the authority. People really felt it was an important counter to the hype.

**FK** ... and it is still important to counter the hype today, especially in policy circles. Just how much of what is usually referred to as AI falls under the category of AI snake oil? And how can we recognise it?

**AN** Much of what is sold commercially today as “AI” is what I call “snake oil”. We have no evidence that it works, and based on our scientific understanding of the relevant domains, we have strong reasons to believe that it couldn’t *possibly* work. My educated guess is because “AI” is a very loose umbrella term. This happens with buzzwords in the tech industry (like “blockchain”). After a point nobody really knows what it means. *Some* are not snake oil. There *has* been genuinely remarkable scientific progress. But because of this, companies put all kinds of systems under the AI umbrella—including those you would have more accurately called regression 20 years ago, or statistics, except that statistics asks rigorous questions about whether something is working and how we can quantify it. But because of the hype, people have skipped this step and the public and policy-makers have bought into it.

Surveys show that the public largely seems to believe that Artificial General Intelligence (AGI) is right around the corner—which would be a turning point in the history of human civilisation! I don’t think that’s true at all, and most experts don’t either. The idea that our current progress with AI would lead to AGI is as absurd as building a taller and taller ladder that reached the moon. There are fundamental differences between what

we're building now and what it would take to build AGI. AGI is not task-specific, so that's in part why I think it will take something fundamentally new and different to get there.

**FK** To build on your metaphor—if the genuinely remarkable scientific progress under the  $\Delta$  umbrella is a ladder to the moon, then AGI would take an entirely different ladder altogether.  $\Delta$  companies are pointing at genuine progress to make claims that require an entirely different kind of progress altogether?

**AN** Right. There's this massive confusion around what AI is, which companies have exploited to create hype. Point number two is that the types of applications of so-called "AI" are fundamentally dubious. One important category is predicting the future, that is, predicting social outcomes. Which kids might drop out of school? Who might be arrested for a crime in the future? Who should we hire? These are all contingent on an incredible array of factors that we still have trouble quantifying—and it's not clear if we ever will.

A few scientific studies have looked rigorously at how good we are at predicting these future social outcomes and shown that it's *barely* better than

random. We can't really do much better than simple regression models with a few variables. My favourite example is the "Fragile Families Challenge" led by my Princeton colleague Professor Matt Salganik, along with colleagues and collaborators around the world. Hundreds of participants used state-of-the-art machine learning techniques and a phenomenal dataset to scrutinise "at-risk" kids over a decade and try to predict (based on a child's circumstances today) what their outcomes might be six years in the future. The negative results are very telling. No team, on any of these social outcomes, could produce predictions that were significantly better than random prediction. This is a powerful statement about why trying to predict future social outcomes is a fundamentally different type of task to those that AI has excelled at. These things don't work well and we shouldn't expect them to.

**FK** Which domains seem to have a lot of snake oil in them and why?

**AN** My educated guess is that to understand the prevalence of AI snake oil it's better to look at the consumers/buyers than the sellers. Companies will spring up around any type of technology for which high demand exists.

So why are

people willing to buy certain types of snake oil? That's interesting.

I think it's because certain domains (like hiring) are so broken that even an elaborate random-number generator (which is what I think some of these AI tools are), is an improvement over what people are doing today. And I don't make this statement lightly. In a domain like hiring we—culturally as well as in business—have a hard time admitting that there is not much we can do to predict who's going to be most productive in a job. The best we can do is have some basic tests of preparation, ability and competence, and beyond that just accept that it's essentially a lottery. I think we're not willing to accept that so much success in life is just randomness, and in our capitalistic economy there's this constant push for more “rationality”, whether or not that makes sense.

So the way hiring works is a) fundamentally arbitrary because these outcomes are hard to predict, and b) there's a lot of bias along all axes

that we know about. What these tools promise to do is cut down on bias that is relatively easy to statistically quantify, but it's much harder to prove that these tools are actually selecting candidates who will do better than candidates who were not selected. The companies who are buying these tools are either okay with that or don't want to know. Look at it from their perspective: they might have a thousand applications for two positions. It's an enormous investment of time to read those applications and interview those candidates, and it's frustrating not to be able to make decisions on a clear candidate ranking. And against this backdrop emerges a tool that promises to be AI and has a veneer of scientific sophistication. It says it will cut down on bias and find the best candidates in a way that is much cheaper to their company than a traditional interview and hiring process. That seems like a great deal.

**FK** So what you're saying is the domains in which snake oil is more prevalent are the ones where either the market is broken or where we have a desire for certainty that maybe doesn't exist?

**AN** I hesitate to provide some sort of sweeping characterisation that explains where there is a lot of snake oil. My point is more that if we look at the individual domains, there seem to be some important reasons why there are buyers in that domain. We have to look at each specific



domain and see what is specifically broken there. There's also a lot of AI snake oil that's being sold to governments. I think what's going on there is that there's not enough expertise in procurement departments to really make nuanced decisions about whether this algorithmic tool can do what it claims.

**FK** Do you think this problem is limited to products and services that are being sold or is this also something you observe within the scientific community?

**AN** A lot of my thinking evolved through the "Limits to Prediction" course that I co-taught with Professor Matt Salganik, whom I mentioned earlier. We wanted to get a better scientific understanding of when prediction is even possible, and the limits of its accuracy. One of the things that stuck out for me is that there's also a lot of misguided research and activity around prediction where we have to ask: what is even the point?

One domain is political prediction. There's a great book by Eitan Hersch which criticises the idea of politics, and even political activism, as a sport—a horse race that turns into a hobby or entertainment. What I find really compelling about this critique is what it implies about efforts like

FiveThirtyEight that involve a lot of statistics and technology for predicting the outcomes of

various elections. Why? That's the big question to me. Of course, political candidates themselves might want to know where to focus their campaigning efforts. Political scientists might want to understand what drives people to vote—those are all great. But why as members of the public...?

Let me turn this inwards. I'm one of those people who refreshes the *New York Times* needle and FiveThirtyEight's predictions. Why do I participate in this way? I was forced to turn that critique on myself, and I realised it's because uncertainty is so uncomfortable. Anything that promises to quell the terror that uncertainty produces and tell us that "there's an 84% chance this candidate will win" just fills a huge gap in our subconscious vulnerabilities. I think this is a real problem. It's not just FiveThirtyEight. There's a whole field of research to figure out how to predict elections. Why? The answer is not clear at all. So, it's not just in the commercial sphere, there's also a lot of other misguided activity around prediction. We've heard a lot about how these predictions have not been very successful, but we've heard less about why people are doing these predictions at all.

**FK** Words like "pseudoscience" and "snake oil" are often thrown around to denote anything

from harmful AI, to poorly-done research, to scams, essentially. But you chose your words very carefully. Why “misguided research” rather than, let’s say, “pseudoscience”?

**AN** I think all these terms are distinct, at least somewhat. Snake oil describes commercial products that are sold as something that’s going to solve a problem. Pseudoscience is where scientific claims are being made, but they’re based on fundamentally shaky assumptions. The classic example is, of course, a paper on supposedly predicting criminality from facial images. When I say “misguided research”, a good example is electoral prediction by political scientists. This is very, very careful research conducted by very rigorous researchers. They know their statistics, I don’t think they’re engaged in pseudoscience. By “misguided” I mean they’re not asking the question of “who is this research helping?”

**FK** That’s really interesting. The question you’re asking then is epistemological. Why do you think this is the case and what do you see as the problems arising from not asking these questions?

**AN** That’s a different kind of critique. It’s not the same level of irresponsibility as some of this harmful AI present in academia and on the street. Once an academic community decides something is an important

research direction, then you stop asking the questions. It's frankly difficult to ask that question for every paper that you write. But sometimes an entire community starts down a path that ultimately leads nowhere and is not going to help anybody. It might even have some harmful side-effects. There's interesting research coming out that the false confidence that people get from seeing these probability scores actually depresses turnouts. This might be a weird thing to say right after an election that saw record levels of turnout, but we don't know whether even more people might have voted had it not been for this entire industry of predicting elections, and splashing those predictions on the front-pages. This is why misguided research is, I think, a separate critique.

**FK** Moving onto a different theme, I have two questions on the limit of predictability. It seems like every other year a research paper tries to predict criminality. The other one for me that surprisingly doesn't die is a 2017 study by two Stanford researchers on predicting homosexuality from faces. There are many, many problems with this paper, but what still fascinates me is that the conversations with policymakers and journalists often revolved around "Well maybe we can't predict this now, but who

knows if we will be able to predict it in future?”. In your talk you said that this is an incomplete categorisation of tasks that AI can be used to solve—and I immediately thought of predicting identity. It’s futile, but the reason why ultimately lies somewhere else. It’s more a question of who we think has the ultimate authority about who defines who we are. It’s an ontological question rather than one about accuracy or biology. I am curious how you refute this claim that AI will be able to predict things in the future, and place an inherent limit on what can be predicted?

**AN** If we look at

the authors of the paper on predicting sexual orientation, one of their main supposed justifications for writing the paper is they claim to be doing this in the interest of the gay community. As repressive governments want to identify sexuality through photos and social media to come after people, they think it’s better for this research to be out there for everybody to see and take defensive measures.

I think that argument makes sense in some domains like computer security. It absolutely does not make sense in this domain. Doing this research is exactly the kind of activity that gives a veneer of legitimacy to an oppressive government who says “Look! There’s a peer-reviewed research paper and it says that this is scientifically accurate, and so we’re doing something

that’s backed  
by science!”

Papers like this  
give ammunition

to people who might do  
such things for repressive

ends. The other part is that if you find a vulnerability in a computer program, it’s very easy to fix—finding the vulnerability is the hard part. It’s very different in this case. If it is true (and of course it’s very doubtful) that it’s possible to accurately infer sexual orientation from people’s images on social media, what are these authors suggesting people do to protect themselves from oppressive governments other than disappear from the internet?

**FK** I think that the suggestion was “accept the death of privacy as a fact and adapt to social norms” which... yeah...

**AN** Right. I would find the motivations for doing this research in the first place to be very questionable. Similarly, predicting gender. One of the main applications is to put a camera

in the back of the taxi that can infer the rider's gender and show targeted advertisements on the little television screen. That's one of the main applications that I'm seeing commercially. Why? You know... I think we should push back on that application in the first place. And if none of these applications make sense, we should ask why people are even working on predicting gender from facial images.

**FK** So you would rephrase the question and not even engage in discussions about accuracy, and just ask whether we should be doing this in the first place?

**AN** That's right. I think there are several kinds of critique for questionable uses of AI. There's the bias critique, the accuracy critique, and the questionable application technique. I think these critiques are separate (there's often a tendency to confuse them) and what I tried to do in the Snake Oil talk is focus on one particular critique, the critique of accuracy. But that's not necessarily the most relevant critique in all cases.

**FK** Let's talk about AI and the current state of the world. I was moderately optimistic that there was less AI solutionism in

LD

response to Covid-19 than I feared. Could this be a positive indicator that the debate has matured in the past two years?

LDR 0.0801

**AN** It's hard to tell, but that's a great question. It's true that companies didn't immediately start blowing the AI horn when Covid-19 happened, and that is good news. But it's hard to tell if that's because they just didn't see enough commercial opportunity there or because the debate has in fact matured.

**FK** There could be various explanations for that...

LG 5.428

**AN** Yeah. There is a lot of snake oil and misguided AI in the medical domain. You see a lot where machine learning was tested on what is called a "retrospective test", where you collect data first from a clinical setting, develop your algorithm on that data and then just test the algorithm on a different portion of the same data. That is a very misleading type of test, because the data might have been collected from one hospital but when you test it on a different hospital in a different region—with different cultural assumptions, different demographics—where the patterns are different, the tool totally fails. We have papers that look at what happens if you test these retrospectively-developed tools in a prospective

E 8840



clinical setting: there's a massive gap in accuracies. We know there's a lot of this going on in the medical machine learning domain, but whether the relative dearth of snake oil AI for Covid-19 is due to the debate maturing or some other factor, who can tell.

**FK** One thing I was wondering... do you feel like you've made an impact?

**AN** (laughs)

**FK** As in, are you seeing less snake oil now than you did, say two years ago?

**AN** That's hard to know. I think there is certainly more awareness among the people who've been doing critical AI work. I'm seeing less evidence that awareness is coming through in journalism, although I'm optimistic that that will change. I have a couple of wish-list items for journalists who often unwittingly provide cover for overhyped claims. One is: please stop attributing agency to AI. I don't understand why journalists do this (presumably it drives clicks?) but it's such a blatantly irresponsible thing to do. Headlines like "AI discovered how to cure a type of cancer". Of course it's never AI that did this.

It's researchers, very hardworking researchers, who use machine learning tools like any other tool. It's both demeaning to the researchers who did that work and creates massive confusion among the public when journalists attribute agency to AI. There's no reason to do that, especially in headlines.

And number two is that it's virtually never meaningful to provide an accuracy number, like "AI used to predict earthquakes is 93% accurate". I see that all the time. It never makes sense in a headline and most of the time never makes sense even in the body of the article. Here's why: I can take any classifier and make it have just about any accuracy I want by changing the data distribution on which I do the test. I can give it arbitrarily easy instances to classify, I can give it arbitrarily hard instances to classify. That choice is completely up to the researcher or the company that's doing the test. In most cases there's not an agreed-upon standard, so unless you're reporting accuracies on a widely-used, agreed-upon benchmark dataset (which is virtually never the case, it's usually the company deciding on-the-road how to do the test) it never makes sense to report an accuracy number like that without a lengthy explanation and many, many other caveats. So don't provide these oversimplified headline accuracy numbers. Try to provide these

**caveats and give qualitative descriptions of accuracy. What does this mean? What are the implications if you were to employ this in a commercial application? How often would you have false positives? Those are the kinds of questions that policymakers should know, not these oversimplified accuracy numbers.**



**Cheap AI**



Abeba  
Birhane

1. This definition was partly inspired by Rabinowitz, A. (2021, January 8) Cheap talk skepticism: why we need to push back against those who are 'just asking questions'. *The Skeptic*. <https://www.skeptic.org.uk/2021/01/cheap-talk-skepticism-why-we-need-to-push-back-against-those-who-are-just-asking-questions>

***Cheap talk (n):  
talk that results in  
harm to the marginal-  
ised but costs nothing  
for the speaker.<sup>1</sup>***

Not a week goes by without the publication of an academic paper or AI tool claiming to predict gender, criminality, emotion, personality, political orientation or another social attribute using machine learning. Critics often label such work as pseudoscience, digital phrenology, physiognomy, AI snake oil, junk science, and bogus AI. These labels are fitting and valid. However, identifying this work as *Cheap* captures the fact that those producing it (usually a homogeneous group from privileged backgrounds) suffer little to no cost, while the

people who serve as the testing grounds, frequently those at the margins of society, pay the heaviest price.

Cheapness emerges when a system makes it easy to talk at little or no cost to the speaker, while at the same time causing tangible harm to the most vulnerable, disenfranchised, and underserved. Within traditional sciences, cheapness manifests when racist, sexist, ableist, misogynist, transphobic and generally bigoted assumptions are re-packaged as scientific hypotheses, with the implication that the only viable way to reject them is to test them. Much of the work from the intimately related fields of “race science” and IQ research constitutes Cheap science. Increasingly, parts of AI research and its applications are following suit.

Cheap AI, a subset of Cheap science, is produced when AI is inappropriately seen as a solution for challenges that it is not able to solve. It is rooted in the faulty assumption that qualities such as trustworthiness, emotional state, and sexual preference are static characteristics with physical



expression that can be read (for example) from our faces and bodies. Software that claims to detect dishonesty, most facial recognition systems deployed in the public sphere, emotion and gait recognition systems, AI that categorises faces as “less or more trustworthy”—all of these constitute Cheap AI. Judgements made by these systems are inherently value-laden, wholly misguided and fundamentally rooted in pseudoscience.

2. Saini, A. (2019) *Superior: the return of race science*. Boston, Mass: Beacon Press.

At the root of Cheap AI are prejudiced assumptions masquerading as objective enquiry. For instance, the very conception of racial categories derives from German doctor Johann Friedrich Blumenbach’s declaration in 1795 that there are “five human varieties: Caucasians, Mongolians, Ethiopians, Americans, and Malays”.<sup>2</sup> This arbitrary classification placed white people of European descent at the top of the hierarchy and cleared the way for colonisation.

Subsequently, apparently scientific racial classifications have served as justifications for inhumane actions that include naturalised slavery, forced sterilization, the genocidal Nazi attempt to exterminate the Jewish people, people with disabilities, and LGBTQ+ people; among other “deviant” classes. Today, racial classifications continue to justify discriminatory practices—including immigration policy—as a way to filter out “inferior” races.

Like other pseudosciences, Cheap science is built around oversimplifications and misinterpretations. The underlying objective of “race science”, for example, is to get at biological or cognitive differences in supposed capabilities between different races, ethnicities, and genders, on the presumption that *there exists* a hierarchy of *inherent* differences to be found between groups. IQ research, for instance, has asserted the existence of race-based differences,

by reducing intelligence to a single number and by framing it as dependent upon race. Similarly, Cheap AI arrogates complex and contingent human behaviour to a face, gait, or body language. Tools based on these presumptions are then produced en masse to classify, sort, and “predict” human behaviour and action.

Cheap AI presents itself as something that ought to be tested, validated, or refuted in the marketplace of ideas. This grants bigoted claims the status of scientific hypothesis, and frames the proponents and critics of Cheap AI as two sides of equal merit, with equally valid intent and equal power. This equivalence is false. While those creating or propagating Cheap AI may face criticism, or reputational harm (if they face these things at all), marginalised people risk discrimination, inhumane treatment, or even death as a result.

Time and time again, attempts to find meaningful biological differences between racial groups have been proven futile, laden with error (there exist more average differences within groups than between groups), and rooted in racist motivations. Yet, the same speculations persist today, just differently framed. In a shift precipitated by the catastrophic effects of Nazi “race science”, race and IQ research has abandoned the outright racist, colonialist, and white supremacist framings of the past, and now masquerades in cunning language such as “populations”,

“human variation”, and “human biodiversity” research.

The mass application of Cheap AI has had a gradual but calamitous effect, especially on individuals and communities that are underserved, marginalised, and disproportionately targeted by these systems. Despite decades of work warning against the dangers of a reductionist approach, so-called “emotion detection systems” continue to spread. Though criminality is a largely complex social phenomenon, claims are still made that AI systems can detect it based on images of faces. Although lies and deception are complex behaviours that defy quantification and measurement, assertions are still made that they can be identified from analysis of video-feeds of gait and gestures. Alarmingly, this and similar work is fast becoming mainstream, increasingly appearing in prestigious academic venues and journals such as NeurIPS and Springer.

The forms Cheap AI takes, the types of claims made of it, and the areas to which it

is applied, are varied and fast expanding. Yet, a single theme persists: the least privileged, the most disenfranchised, and the most marginalised individuals and communities pay the highest price. Black men are wrongly detained due to failures in facial recognition systems; Black people with ill health are systematically excluded from medical treatment; the elderly and women are not shown job ads. These few cases represent only the tip of the iceberg. So far there are three *known* cases of Black men wrongly detained due to facial recognition; there are likely many more. The victims of algorithmic injustice that we know about are often disenfranchised, and it is likely that many more are fighting injustice in the dark, or falling victim without redress, unaware that Cheap AI is responsible.

Like segregation, much of Cheap AI is built on a logic of punishment. These systems embed and perpetuate stereotypes. From “deception detection” to “emotion recognition” systems, Cheap AI serves as a tool that “catches” and punishes those deemed to be outliers, problematic, or otherwise unconventional.

The seemingly logical and open-minded course of action—to withhold judgement on these systems on the premise that their merits lie in how “well” or “accurately” they

work—lends them a false sense of reasonableness, giving the impression that the “self-correcting” nature of science will eliminate bad tools. It also creates the illusion of there being “two sides”. In reality, criticisms, objections, and calls for accountability drown in the sea of Cheap AI that is flooding day-to-day life. Cheap AI is produced at an unprecedented rate and huge amounts of money go into producing it. By contrast, those working to reveal it as scientifically unfounded and ethically dangerous are scholars and activists working under precarious positions with little to no support, who are likely to suffer negative consequences.

By suspending judgement until wrong is proved, an ecosystem has been created where anyone can claim to have created obviously absurd and impossible tools (some of which are nonetheless taken up and applied) without facing any consequences for engaging in Cheap AI. Such creators and deployers may risk their reputations when their tech is proven to be “inaccurate”. However, for those who face the burn of being measured by this tech, it can be a matter of life and death, resulting in years lost trying to prove innocence, and other grave forms of suffering.

There is no quick solution to ending Cheap AI. Many factors contribute to the ecology in which it thrives.

This includes blind faith in  $\mathbb{S}_A$ , the illusion of objectivity that comes with the field's association with mathematics, Cheap  $\mathbb{N}$ 's creators and deployers' limited knowledge of history and other relevant fields, a lack of diversity and inclusion, the privilege hazard (a field run by a group of mostly white, privileged men who are unaffected by Cheap  $\mathbb{N}$ 's harms), the tendency to ignore and dismiss critical voices, and a lack of accountability. We must recognise Cheap  $\mathbb{A}$  as a problem in the ecosystem. All of these factors and more need to be recognised and challenged so that Cheap  $\mathbb{A}$  is seen for what it is, and those producing it are held accountable.

It took Nazi-era atrocities, forced sterilizations, and other inhumane tortures for phrenology, eugenics, and other pseudosciences to be relegated from science's mainstream to its fringe. It should not take mass injustice for Cheap  $\mathbb{A}$  to be recognised as similarly harmful. In addition to strict legal regulation and the enforcement of academic standards, we ourselves also bear a responsibility to call out and denounce Cheap  $\mathbb{A}$ , and those who produce it.

LDF 0.1101

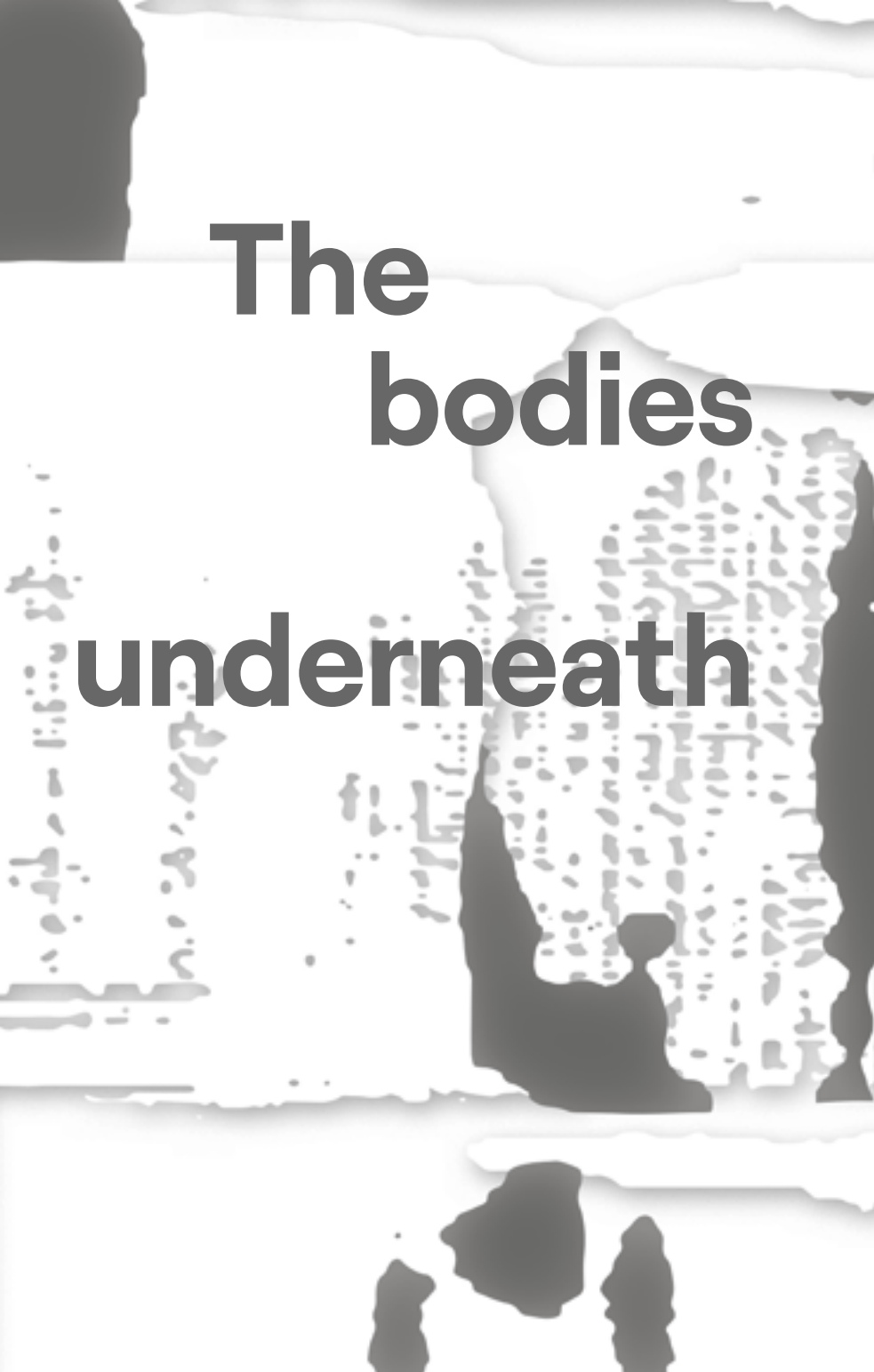
LDR 0.1922

LG 2.9631

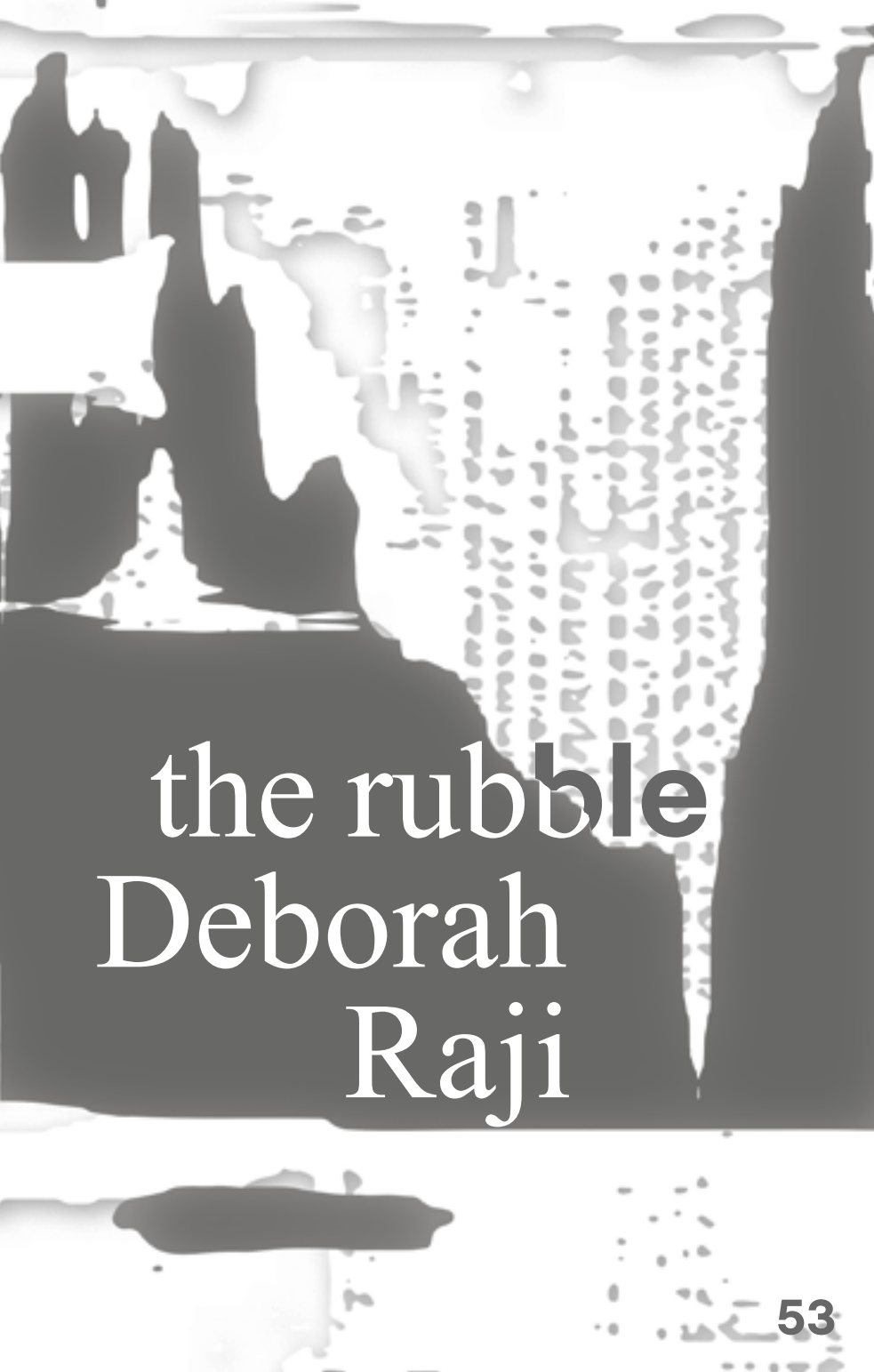
E 4480

Abeba Birhane is a cognitive science PhD candidate at the Complex Software Lab, University College Dublin, Ireland, and Lero, the Science Foundation Ireland Research Centre for Software.





**The  
bodies  
underneath**



the rubble  
Deborah  
Raji

1. Charette, R. N. (2018, 24 January) Michigan's MiDAS Unemployment System: Algorithm Alchemy Created Lead, Not Gold, *IEEE Spectrum*.

<https://spectrum.ieee.org/riskfactor/computing/software/michigans-midas-unemployment-system-algorithm-alchemy-that-created-lead-not-gold>

LDF 0.2

LDR 0.0334

LG 5.8964

**On August 29th 1907, the Quebec Bridge collapsed. Bound for the record books as the longest cantilever bridge ever to be built, the disaster cost 75 of the 86 lives working on its construction that day. In the formal inquiry that followed, the cause of the accident was traced to the poor oversight of the two lead engineers, Theodore Cooper and Peter Szlapka. The landmark declaration on engineering responsibility called out both by name. Their avoidable missteps, careless miscalculations, and chronic disregard for safety were described in detail and ultimately identified as key to the bridge's structural failure.**

Decades later, in 1965, US consumer advocate Ralph Nader published *Unsafe at Any Speed*, a book which would transform the automotive industry. In it, Nader revealed the alarming instability of the Chevrolet Corvair—a vehicle which, for him, exemplified the auto industry's indifference towards safety. He noted an absence of foresight in how cars were designed and actually used, and

connected this to a string of avoidable collisions and accidents. He flagged this as a new era, one marked by its own kind of collapse. In lieu of careful planning and precaution, he observed instead only “an explosion of what is actually done”.

Such “explosions” persist today in the engineering of algorithmic systems. The rubble is everywhere. The deployment of the MiDAS algorithm led to over 20,000 Michiganders being erroneously accused of unemployment fraud.<sup>1</sup> A model widely used in US hospitals to allocate health-care systematically discriminates against Black patients.<sup>2</sup> Several facial recognition products fail on Black female faces.<sup>3</sup> Uber’s self-driving car software has led to dozens of accidents and even a human death.<sup>4</sup> In each case, the investigated causes are increasingly

2. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366.6464: 447-453. DOI: 10.1126/science.aax2342

3. Raji, I. D. & Buolamwini, J. (2019) Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. DOI: 10.1145/3306618.3314244

4. Lee, K. (2019, 6 November) Uber’s Self-Driving Cars Made It Through 37 Crashes Before Killing Someone. *Jalopnik*. <https://jalopnik.com/ubers-self-driving-cars-made-it-through-37-crashes-befo-1839660767>

**inexcusable—uninspected “corrupt and missing” data, the admittedly incorrect framing of a task, a remarkable lack of oversight in testing, or an unimplemented feature.**

**The harm inflicted by these products is a direct consequence of shoddy craftsmanship, unacknowledged technical limits, and poor design—bridges that don’t hold, cars that can’t steer. Contrary to popular belief, these systems are far more likely to cause harm when they do not work than on the rare occasion in which they work “too well”, or in truly unaccountable ways.**

**Even in light of these obvious harms, vendors remain stubborn—and defensive. Despite a 93% error rate, the MiDAS algorithm was used for at least three years. It took several highly publicised lawsuits for its use to finally be questioned.**

5. Ledford, H. (2019) Millions of black people affected by racial bias in health-care algorithms. *Nature* 574.7780: 608-610. DOI: 10.1038/d41586-019-03228-6

Optum, the company behind the healthcare prioritisation algorithm, called audit results helpful but “misleading” and continues to deploy the system on millions of patients.<sup>5</sup> In spite of a peer reviewed study revealing an over 30% error rate on darker female faces, Amazon still pitches its biased facial recognition technology for nation-wide partnerships with law enforcement. Uber’s automated vehicles, which are still being tested on public streets, continue to run on its flawed software.

Just as the car manufacturer called out by Nader shifted blame onto car dealerships for failing to recommend tyre pressures to “correct” the Corvair’s faulty steering, algorithm developers also seek scapegoats for their own embarrassing failures. Optum states that the quality of its healthcare algorithm’s assessments actually depends on “the doctor’s expertise”. Amazon discusses setting higher accuracy thresholds for police clients, and MiDAS’s creators point to the difficulty of migrating from an independent legacy system. Even the fatal Uber self-driving car crash was ultimately blamed on a distracted test driver

and the victim's own jaywalking. Those creating these algorithms will point to anything—the human operator, institutional infrastructure, society-at-large, or even the actions of the affected population itself—before admitting to inherent flaws in the product's design and implementation. Eyes averted, they claim the issues reported are really someone else's responsibility.

However, just as it was General Motors' responsibility to ensure that the Corvair's required tyre pressures were within the range of recommended tolerance, it is also Amazon's responsibility to inform police clients to operate at a level other than the default accuracy threshold. It is Uber's responsibility to ensure drivers remain adequately attentive at the wheel, the MiDAS developers' task to make their algorithm portable, and Optum's role to analyse the causal assumptions in their data. To forget a test, neglect design or dismiss the data is not just unfortunate, it's irresponsible.

Any institutional stakeholder involved in the development of a product—be they engineer, executive business authority, or product or marketing participant—does, ultimately, have an impact on outcomes. The technical community tends to simultaneously

anthropomorphise AI systems, ascribing to them a false ability to “think” and “learn” independently of the provided input, while also downplaying, with language of data “bricks”, “moats” and “streams”, the existence of real humans in the data and human participation in the decisions that shape these systems. This reluctance to admit to the human influence on AI functionality makes the field stubbornly blind to its contribution to these systems. Technologists are not like doctors, looking each patient in the eye. They stand at a distance, the relationship between their judgement and system outcomes blurred by digitised abstraction, their sense of responsibility dampened by scale, the rush of agile innovation, countless undocumented judgements, and implicit feature engineering. The result is an imagined absolution of responsibility, a false narrative in which they’ve created an artificial system outside of anyone’s control, while the human population affected by their decisions and mistakes is inappropriately erased.

As a former participant in an Applied Machine Learning team, I’ve witnessed the model development process up close. It’s a chaotic enterprise. Whether building a moderation model that disproportionately filtered out the content of



people of colour, or training a hair classifier that did not have inclusive categories, we regularly disappointed our clients and ourselves, frequently falling short of expectations in practice. Even after a live pilot failed or unacceptable bias was discovered, at the client's request we would deploy the model anyway. The fact is, in this field, inattentive engineers often get away with it. No data review is imposed, nor reporting requirements. There is no sanctioned communication protocol with operators, no safety culture to speak of, no best practices to adhere to, no restrictive regulations or enforced compliance, and no guide for recall—voluntary or imposed—to remove from our daily lives the models that cannot be steered, the algorithms without brakes.

Of the 75 construction workers who died in the Quebec Bridge catastrophe, up to 35 were Native Americans, Mohawks from the Kahnawake community who faced a dearth of employment options at the time and received poverty wages. This is what happens when systems collapse: they fall on the most vulnerable. In his book, Nader described Ms. Rose Pierini “learning to adjust to the loss of her left arm” after a crash caused by the unaddressed steering challenges of the Chevrolet Corvair.

He profiled Robert Comstock, a “veteran garage mechanic”, whose leg was amputated after he was run over by a Buick with no brakes. Today, we discuss Robert Williams, a Black man wrongfully arrested due to an incorrect facial recognition “match”; Carmelita Colvin, a Black woman falsely accused of unemployment fraud; Tammy Dobbs, a sickly older lady who lost her healthcare due to a program glitch, and Davone Jackson, locked out of the low-income housing he and his family needed to escape homelessness due to a false report from an automated tenant screening tool.

The fact is, there are bodies underneath the rubble. The individuals caught in the web of AI’s false promises are just the latest casualties of technological collapse. It will never be enough to feign ignorance, deflect blame, and keep going. For the sake of those that suffer, responsibility is required. It’s a concurrent goal, of a differing urgency, to address alongside any ambition for a more hopeful future.

Deborah Raji is a Mozilla fellow, interested in algorithmic auditing. She also works closely with the Algorithmic Justice League initiative to highlight bias in deployed products.

**Who am I  
as data?**

Frederike



Kaltheuner

**Who am I as data? The question has haunted me for years. It's a ghoulish curiosity; wanting to know not just how much data companies routinely harvest about my behaviour, but also what this data might reveal about me, and who I appear to be.**

**It also stems from my fascination with a paradoxical societal shift. We are living at a time when the terms we use to describe identities are becoming more fluid, and boundaries more negotiable. While our language falls short of fully representing our social realities, it reflects the fundamental changes that are afoot—providing an echo of what is already the case. At the same time, another current is pulling in the opposite direction. Out of sight, we are increasingly surrounded by data-driven systems that automatically affix names to us and assign us an identity. From the gender binary**

**that's encoded into targeted online advertising, to facial recognition systems that claim to predict people's gender or ethnicity, these systems are often completely inadequate to our social realities. Yet**

**simply by existing, they classify and thereby mould the world around us.**

1. Kaltheuner, F. (2018). I asked an online tracking company for all of my data and here's what I found. *Privacy International*. <https://privacyinternational.org/long-read/2433/i-asked-online-tracking-company-all-my-data-and-heres-what-i-found>

In 2018, I asked an advertising company for all of my<sup>1</sup> data. Quantcast, an AI company that is known for its cookie consent notices, tracks users across millions of websites and apps to create detailed profiles of people's browsing histories, presumed identity, and predicted interests.

Staring into the abyss of one's tracking data is uncanny. On the one hand, there were pages and pages of my own browsing history—not just top-level domains, but URLs that revealed exactly what I had read and clicked on, which restaurants I'd booked, and which words I'd translated into German. Then there were the predictions made about me, confidence intervals about my predicted gender, age, and income. Had I done the same experiment outside the European Union—where this is unlawful—the data would have included an ethnicity score.

Was this who I am? I expected to feel somewhat violated, but not so misunderstood. These rows and rows of data revealed a lot about me, my interests, and how I spend my time (every click is timestamped and carries other metadata) but the picture remained incomplete, shallow. Eerily accurate inferences were combined with seemingly random consumer categories: bagel shop frequenter, alcohol consumption at home.

AI-driven profiling for advertising may sound banal, yet it is the most profitable and prevalent

use of AI to classify, infer, and detect people's identity—be it gender, ethnicity or sexual orientation. These AI systems operate in the background, in ways that are fundamentally beyond people's knowledge or control. Those who are classified and assessed by them frequently don't know where, when or how these systems are used. These systems are deeply Orwellian when they happen to get it right, and Kafkaesque when they do not.

Just to be crystal clear: it is impossible to detect someone's gender, ethnicity, or sexual orientation using AI techniques, and any attempt to do so falls squarely in the domain of AI pseudoscience. That is not because AI isn't good enough yet, or because we need more and better data. It's because identity isn't something that can be straightforwardly detected, like the colour or shape of an object.

Identity is a fallacy, argues Kwame Anthony Appiah in *Rethinking Identity*<sup>2</sup>. The fallacy is in assuming that there is some deep similarity that binds together people of a particular identity—because ultimately such a similarity does not exist—and yet, identities matter deeply to people precisely because belonging to something larger than oneself is a key human need. To acknowledge the ways identity is at once true and a lie reveals the tension at its heart. The negotiations

between how we identify ourselves and how others identify us, what we are identifying with and what we're being identified as, are perpetual.

AI-driven identity prediction attempts to dissolve this tension by reducing complex questions of identity to what is automatically detectable from the outside, according to pre-conceived notions of what an identity is. At best, AI systems can offer a very incomplete guess of someone's presumed identity—the data-driven and automated equivalent of the stereotyping that people regularly do to passing strangers in the street. At worst, AI systems that claim to detect identity give racism, transphobia and sexism a veneer of scientific advancement and objectivity. This has devastating consequences for those who are either targeted or rendered invisible by these systems. As Sasha Costanza-Chock describes in *Design Justice*, the gender binary that is encoded into so much infrastructure, such as airport security systems, amounts to ontological reduction:

“As a non-binary trans femme,

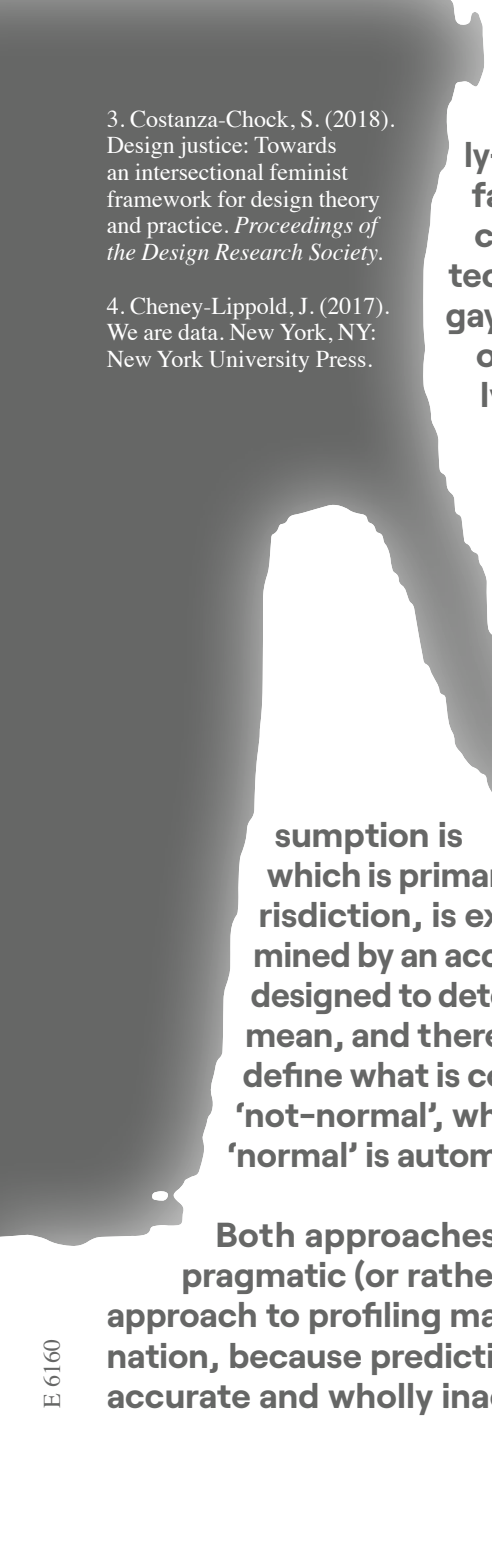
2. Appiah, K. A. (2018).  
The lies that bind: Rethinking  
identity. London: Profile Books.



I present a problem not easily resolved by the algorithm of the security protocol.”<sup>3</sup>

Broadly speaking, there are two ways in which AI is used commercially to assess, detect or predict identity. Let’s call them the pragmatic and the essentialist approaches. The former predominates in online advertising. Ultimately, it doesn’t matter to Quantcast, Google or Facebook whether I truly *am* a woman or a man (or identify as trans, an option that is often not available). What matters is whether my purchasing decisions and online behaviour resemble that of other people who are also classified as “women”. In other words, “woman” is not really a statement about identity, as John Cheney-Lippold writes in *We Are Data*<sup>4</sup>. It is an arbitrary label that is applied to an arbitrary group of people who share an arbitrary set of online behaviours. What determines the label and the group is not whether it is true, but whether grouping people together increases whatever the advertising system is optimised for: clicks, time on site, purchases.

By contrast, the essentialist approach positions automated predictions about identity as truth claims. This is where pseudoscience comes in. The idea behind security cameras that detect Uyghur ethnicity is to *really* detect and spot people of Uyghur descent in



3. Costanza-Chock, S. (2018). Design justice: Towards an intersectional feminist framework for design theory and practice. *Proceedings of the Design Research Society*.

4. Cheney-Lippold, J. (2017). We are data. New York, NY: New York University Press.

a crowd. When a widely-criticised and now-infamous Stanford study claimed that AI can detect whether someone is gay or straight based only on a photo, the underlying premise was that sexual orientation is

*really* somehow visually inscribed in a face. And when immigration authorities use voice recognition to detect accents to determine nationality, the unspoken as-

sumption is that nationality, which is primarily a matter of legal jurisdiction, is expressed in and determined by an accent. These systems are designed to detect deviations from the mean, and thereby they also inevitably define what is considered 'normal' and 'not-normal', where anything that isn't 'normal' is automatically suspicious.

Both approaches are problematic. The pragmatic (or rather, "just good enough") approach to profiling may still lead to discrimination, because predictions can be both eerily accurate and wholly inaccurate or inadequate

at the same time. Whether accurate or not, inferring someone's ethnicity or gender without their knowledge or consent still means this "information" can be used in ways which discriminate or exclude. Inferences that are produced for contexts which make do with a pragmatic approach—for example, in advertising—may still end up in a context where accuracy

and truth matter.

It's one thing to target an ad based on someone's likely interests, whereabouts and ethnicity; it's something entirely different to use the same data for immigration enforcement.

What's so deeply troubling about the essentialist approach to AI identity prediction, by contrast, is not just how it operates in practice. These predictions come with significant margins of error, completely eliminating those who don't fit into whatever arbitrary classifications the designers of the system have chosen, and their harms disproportionately affect communities that are already marginalised. Equally disturbing is the insidious idea that automated classification systems can, and *should* be the ultimate authority over who we are understood to be. Decisions of this magnitude can affect, curtail and even eliminate the ability to enjoy rights as an individual, a citizen, or member of any other group. Yet

these systems are intended to be, and are often treated as, more reliable, more objective and more trustworthy than the statements made by those they are assessing.

This is as troubling as it is absurd, for the essentialist approach ignores the fact that the underlying and unspoken assumptions, norms, and values behind every identity prediction system are categories and modes of classification that were chosen by the designers of those very systems. Take, for instance, the gay faces study mentioned earlier. The entire model was based on the assumption that people are either gay or straight, male or female. Such binary labels fail to capture the lived experience of a vast number of queer people—not to mention that the study only included white people.

Classifications have consequences, and produce real-world effects. In the words of Geoffrey Bowker and Susan Star, “For any individual group or situation, classifications and standards give advantage or they give suffering.”<sup>5</sup> The idea that an automated system can detect identity risks transforming inherently political decisions—such as opting to use the gender binary—into a hard, yet invisible infrastructure. Such is

5. Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. Cambridge, Mass: MIT Press.

the case when any form of attribute recognition is added into computer vision systems, such as face recognition cameras.

AI pseudoscience does not happen in a vacuum. It is part of a much broader revival of certain ways of thinking. Amid the return of race science and the emergence of DNA tests that designate people's ancestry as "35 percent German", or "76 percent Finnish", the use of AI to predict and detect identity needs to be seen as part of a much wider revival of (biological) essentialism and determinism<sup>6</sup>. Many companies that offer genetic predictions, for instance, sell much more than DNA tests<sup>7</sup>. They are also benefiting from—and ultimately spreading—the dangerous, yet incredibly compelling, idea that who we are is ultimately determined by biology. There are now DNA companies that claim genetics can predict people's ideal lifestyle, their intelligence, their personality and even their perfect romantic partner. A German start-up makes muesli based on people's DNA, house-sharing platform

6. Saini, A. (2019). *Superior: The return of race science*. Boston, Mass: Beacon Press.

7. Kaltheuner, F. (2020). *Acknowledging the Limits of Our AI (and Our DNA)*. *Mozilla*. <https://foundation.mozilla.org/en/blog/acknowledging-limits-our-ai-and-our-dna/>

SpareRoom trailed genetically matched roommates, and the music streaming service Spotify offers a playlist tailored to your DNA.

At the core of this determinism lies people and categories ble and therefore shadow this casts categories of identity, and by extension some people's lives, are superior to others. Dmerely gives outmoded ways of thinking the veneer of objectivity and futurism.

return to biological the idea that both peo- are fixed, unchangea- predictable. The dark is the belief that some

In reality, categories like race, gender, and sexual orientation evolve over time. They remain subject to contestation. The idea of a criminal face is absurd, not least because our ideas of criminality are constantly changing. Does our face change whenever laws change, or when we move to a new country? What are DNA companies referring to when saying that someone is German? The Federal Republic of Germany? The Holy Roman Empire of the German Nation that existed from 1512 to 1806? Nazi Germany? The very idea of the nation state is a modern concept. And is someone who recently immigrated to Germany and has a German passport not German?

On the individual level, we are often different things to different people. We reveal different parts of ourselves in different settings. How we see ourselves can evolve or even radically change. This freedom to selectively disclose and manage who we are to whom, and the space we have to do it in, is drastically eroded by the increased ability of companies and governments to link and join previously distinct data points, both spatially and temporally, into a distinct, singular identity—an apparently definite assessment of who each of us is as a person.

In 2020, Google announced that it would drop gender recognition from its Cloud Vision API, which is used by developers to analyse what's in an image, and can identify anything from brand logos to faces to landmarks. That's a good first step, but further action needs to be taken in industry more widely. For this we need AI regulation that does more than simply setting up the standards and guidelines that determine how AI systems can be used. What is needed are clear red lines that demarcate what AI cannot and should not be used for.

**Resisting and protecting people from AI pseudoscience is about far more than making AI accountable, explainable, more transparent, and less biased. It is about defending and vigorously protecting what it means to be a person. It is about resisting ontological reduction. It is about allowing space for identities to be challenged and contested, not cemented by inscrutable automated systems.**

Frederike Kalthener is a tech policy analyst and researcher. She is also the Director of the European AI Fund, a philanthropic initiative to strengthen civil society in Europe.



# The case for interpretive techniques

Razvan Amironesei  
Emily Denton  
Alex Hanna  
Hilary Nicole  
Andrew Smart



**in  
machine  
learning**

Many modern AI systems are designed to ingest and analyze massive datasets so as to make recommendations, predictions, or inferences on unseen inputs ranging from images to pieces of text and other forms of data. These datasets often reflect patterns of inequity that exist in the world,<sup>1</sup> and yet the data-driven nature of AI systems often serves to obscure the technology's limitations within a pervasive rhetoric of objectivity.<sup>2</sup> As AI technologies and methods are increasingly incorporated into all aspects of social life, often in ways that increase and accelerate existing social inequities, this is especially important. In this piece, we examine how and why appeals to objectivity are so deeply embedded in technological discourses and practices. As Ruha Benjamin notes, routing algorithmic bias through a rhetoric of objectivity can make it "even more difficult to challenge it and hold individuals and institutions accountable."<sup>3</sup> Our starting question is: how, and under which conditions, do truth claims that are embedded in algorithmic systems and associated data practices function as a justification for a myriad of harms?

A way for us to answer this question lies in understanding and accounting for how material artefacts—that is to say the instruments, devices, and sociotechnical systems—contribute to an understanding of algorithmic systems as objective or scientific in nature.

1. Smith, A. (2019, 12 December) AI: Discriminatory Data In, Discrimination Out. *SHRM*. <https://www.shrm.org/resourcesandtools/legal-and-compliance/employment-law/pages/artificial-intelligence-discriminatory-data.aspx>; Richardson, R. Schultz, J., Crawford, K. (2019) Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online* 94:192. SSRN: 3333423

2. Waseem, Z., Lulz, S., Bingel, J. & Augenstein, I. (2021) Disembodied Machine Learning: On the Illusion of Objectivity in *NLP*. <https://arxiv.org/abs/2101.11974>

3. Benjamin, R. (2019) *Race After Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity Press, p.122.

Of these artefacts, benchmark datasets play a crucial role in the constitution of the machine learning life cycle. These datasets are used in the training and development of artificial intelligence. They establish a “gold standard” for specific AI tasks, defining the ideal outputs of an AI system for a set range of exemplar data inputs. Benchmark datasets can be understood as measurements for assessing and comparing different AI algorithms. Within AI research communities, performance on benchmark datasets is often understood as indicative of research progress on a particular AI task. Benchmarks are the equivalent of IQ tests for algorithms. Just as IQ tests are controversial because it is unclear what exactly they measure about human intelligence, what benchmark datasets are supposed to be measuring about algorithms has never been fully articulated. And while the role of IQ tests in historically providing justification for white supremacist beliefs is well recognised, despite these data infrastructures

having significant social impact through the dissemination of unjust biases, they remain strikingly under-theorised and barely understood, or even acknowledged, in the public sphere. In what follows, we will address benchmarks via two definitions which entail, as we will see, different types of problems and critiques.

4. Butterfield, A., Ekembe Ngondi, G. & Kerr, A. (Eds) (2016) *A Dictionary of Computer Science*, 7th edition. Oxford: Oxford University Press.

Defined from a purely technical perspective, a benchmark is “a problem that has been designed to evaluate the performance of a system [which] is subjected to a known workload and the performance of the system against this workload is measured.” The objective is to compare “the measured performance” with “that of other systems that have been subject to the same benchmark test.”<sup>4</sup>

In order to illustrate the limits of a purely technical understanding of benchmark datasets, let’s briefly discuss ImageNet, a large visual dataset which is used in visual object recognition research. Its stated objective is to map “the entire world

of objects.”<sup>5</sup> The dataset contains more than 14 million hand-annotated images, and was one of the largest created at the time, making it one of the most important benchmark datasets in the field of computer vision. However, as research led by Kate Crawford and Trevor Paglen has illustrated, ImageNet does more than annotate objects with relatively straightforward descriptions (such as “apple”, “round” or “green”).<sup>6</sup> The dataset contains a significant number of categorisations which can only be described as depreciative and derogatory. For instance, a photograph of a woman in a bikini is labelled “slattern, slut, slovenly woman, trollop” and a man drinking a beer as “alcoholic, alky, dipsomaniac, boozier, lush, soaker, souse.” So, how is it that morally degrading, misleading, and defamatory descriptions shape the purportedly objective structure of the benchmark dataset?

5. Gershgorn, D. (2017, 26 July)

The data that transformed AI research – and possibly the world. *Quartz*. <https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/>

6. <https://excavating.ai/>

**One explanation lies in the fact that benchmark datasets are typically perceived by the machine learning community as purely technical devices which**

provide factually objective data. This is the case of ImageNet, in which a particular label attached to an image is often interpreted as a truth claim about the nature of the object or phenomenon depicted. As a result, benchmark datasets operate on—and reinforce—the assumption that they represent some “fact” in the world. A purely technical definition of benchmarks also does not take into account how social, ethical and political factors shape the dataset. Clearly, we should question the assumption that technical objectivity is somehow embedded in benchmark datasets. A different framing of benchmark datasets—which takes into account the context of the production process that shaped them into existence—is needed.

In this reformulated definition, benchmarks can be understood as socio-technical measurements, governed by specific norms that, in turn, act as standards of evaluation. As we can see, for example with the ImageNet Large Scale Visual Recognition Challenge,<sup>7</sup> state-of-the-art performance on the benchmark challenges came to

7. Russakovsky, O., Deng, J., Su, H., et al. (2015) *ImageNet Large Scale Visual Recognition Challenge*. <https://arxiv.org/abs/1409.0575>

8. Metz, C. (2021, 16 March) The secret auction that set off the race for AI supremacy. *Wired*. <https://www.wired.com/story/secret-auction-race-ai-supremacy-google-microsoft-baidu/>

be understood as not only a measure of success on the specific formulation of object recognition represented in the benchmark, but as a much broader indicator of AI research progress.

Researchers who have produced state-of-the-art performance on benchmark datasets have gone on to receive prestigious positions in large industry labs, or received massive funding for AI startups.<sup>8</sup> Such a pervasive view of benchmark datasets as value-neutral markers of progress is both misguided and dangerous. Benchmark datasets can be more appropriately understood as tools that institute and enforce specific norms, which establish normative modes of functional accuracy, error rates and such. In other words, benchmark datasets are measurement devices that function as a regulative mechanism for assessing and enforcing a particular standard in a technical setting. When understood this way, it becomes clear that the histories of their production and epistemological limits should be thoroughly documented. Current benchmarking practices reveal troubling moral problems, in particular when gold standards become petrified in a field of practice, and are uncritically accepted and reproduced. In such

LDF 0.2106

LDR 0.0394

LG 4.8099

L420



circumstances, benchmarks normalize and perpetuate arbitrary assumptions which function as normalization mechanisms.

For us, as researchers in the field of  $\Delta$ , to address the harmful effects of benchmarks understood as normalizing mechanisms, we propose cultivating a careful and responsible critique which analyses the formation of meaning inherent in these technologies. By understanding the socio-technical nature of dataset production and their contingent genesis, we create the conditions

to stem the myriad harms mentioned in our introduction.

In a word, analysing datasets along technical, social, and ethical axes reveals their contestable modes of construction. Datasets have a socio-ethical reality that is distinct from their socio-technical dimension. As such, it's possible to recognise datasets as the contextual product, or contingent outcome, of normative constraints and conflicts between various stakeholders with visible or concealed agendas. Thus, the representational harms that we have referenced in the context of ImageNet are not simply the unexpected and unfortunate effect of purportedly objective algorithmic systems, but, most importantly, they can be traced back

to their interpretive origins, that is, the underlying conditions, presuppositions, practices and values embedded in dataset creation.

By placing (distorted and degrading) labels on how various beings and objects exist in the world, benchmark datasets exert a computational power of naming. This power of naming—which mimics and recalls, for example, Linnaeus' Promethean efforts in *Species Plantarum*—operates as a power to classify and catalogue all the existing objects in the world. By labelling an object available to the computational gaze, the dataset grants and withdraws recognition based on the socio-technical assessment of the object's identity. Left unchecked, this mode of perceiving and labelling the world amounts to accepting and scaling the reproduction and normalization of representational harms.

To address this problem, we seek to validate benchmark datasets by analysing their internal modes of construction. We aim to do so by using *techniques of interpretation* which establish rules for understanding the relation between data collection practices and the ways in which they shape models of algorithmic development.

In particular, a key necessary step is performing an interpretive socio-ethical analysis of how and when training sets should be used in the creation of benchmark datasets. This should cover: the disclosures associated with the work and research on specific datasets; the stakeholders involved and their reflective or unreflective intent, data practices, norms and routines that structure the data collection, and the implicit or explicit values operative in their production; the veiled or specific assumptions of their authors and curators; the relation between the intended goals of the original authors and the actual outcomes; and the adoption of datasets and the related practices of contestation by subsequent researchers. This way, we will be in a position to adequately identify and interrogate the historical conditions of dataset production, document their related norms, practices and axiological hierarchies, and thereby both reveal and prevent the excesses that currently operate in the machine learning pipeline.

Dr Razvan Amironesei is a research fellow in data ethics at the University of San Francisco and a visiting researcher at Google, currently working on key topics in algorithmic fairness.

Dr Emily Denton is a Senior Research Scientist on Google's Ethical AI team, studying the norms, values, and work practices that structure the development and use of machine learning datasets.

Dr Alex Hanna is a sociologist and researcher at Google.

Hilary Nicole is a researcher at Google.

Andrew Smart is a researcher at Google working on AI governance, sociotechnical systems, and basic research on conceptual foundations of AI.

All authors contributed equally.

**Do  
we**

**need  
or  
do we  
need**

**Black  
feminisms?**



**A poetic  
guide**

Serena  
Dokuaa  
Oduro

LDF 0.1028

**Black feminism is rooted in the knowledge that every system and technology has race, gender, and class implications. Black feminism grows from Black women's experience of living at the intersections of race, gender, class, and many other axes of oppression.**

LDR 0.0441

**Black feminists know  
That we  
Cannot run from  
Oppression.  
From work  
To the kitchen**

**Everything  
Is political.**

LG 5.0597

**No technology is truly race neutral. Breakthrough research, such as that conducted by Dr Safiya Umoja Noble about the racism and sexism embedded in search engines, has proven it. In fact, the very claim that algorithmic design is race neutral warrants caution in that it is meant to reinforce white, Western, and male norms and power.**

E.5600

Black death is  
No glitch.  
Black erasure is  
No glitch.  
Black surveillance is  
No glitch.  
Black false positives is  
No glitch.

**New questions must guide the purpose, design, development, and implementation of AI. Not only are new approaches needed to ensure that all groups benefit from AI, they are also needed to curb AI disillusionment. Every report of another instance of racial discrimination in AI further fuels rightful Black scepticism. It is not enough for the goals of AI to be non-discriminatory. AI must seek to benefit and liberate Black women and all historically marginalised groups.**



I dream of AI being  
Crafted by Black hands  
And Black dreams.  
If I can sit around the table  
And gush about AI  
With my Mom, Sister  
And Aunties, then  
I'll believe in it.

**Black feminism provides a roadmap for improving AI, because Black feminist politics “actively [commit] to struggling against racial, sexual, heterosexual and class oppression, and sees as [its] particular task the development of integrated analysis and practice based upon the fact that the major systems of oppression are interlocking.”<sup>1</sup> Current research on the impacts of AI ignores those at the intersections of marginalised identities and the differences between historically marginalised groups.<sup>2</sup> This poetic guide can be used to measure the usefulness and validity of various AI applications with a Black feminist lens. If an AI system cannot pass the muster of these questions, I recommend seriously considering that it may be AI hype, snake oil, or pseudoscience.**

LG 5.0665

E 36780

## 1. DOES THIS ALGORITHM INCREASE INDIVIDUAL, INTERPERSONAL, AND COMMUNAL WELL-BEING FOR BLACK WOMEN?

An AI system is not useful for me as a Black woman if it reinforces racist and sexist norms. If an AI system perpetuates the hypersexualisation of Black women, then it does not increase my well-being. If an automated decision system defaults to white content, it does not increase my individual, interpersonal or communal well-being. If the algorithm prioritises content that aligns with colourist, texturist, fatphobic, queerphobic or transphobic norms, then it does not increase individual, interpersonal or communal well-being for Black women.

## 2. IS THIS ALGORITHM LESS BIASED THAN A BLACK FEMINIST?

If the AI system is only better than humans because it is slightly less racist or sexist, then it is not useful for Black women.

1. The Combahee River Collective (1978) A Black Feminist Statement. *Capitalist Patriarchy and the Case for Socialist Feminism*. Ed. Eisenstein, Z. R. New York: monthly review press.

2. Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020) Towards a critical race methodology in algorithmic fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. DOI: 10.1145/3351095.3372826

### 3. SHOULD BLACK FEMINISTS BE HYPED?

If Black feminists are hyped, then I know individual, interpersonal, and communal well-being is going to increase. If Black feminists are hyped, then I know it will be beneficial for historically oppressed people, because Black feminisms are rooted in the knowledge that oppressions are tied together. Since oppressions are tied together, so is liberation.

### 4. WHY IS THIS AI A SOLUTION FOR BLACK WOMEN?

It is typical in the West to look for technical solutions to social problems. But racism, sexism, and class difference cannot be fixed with AI alone. How does this algorithm help Black feminists oppose racial, gender, and class oppression? Who is designing this system? Who is classifying groups?<sup>3</sup> Is this AI about bias minimisation or profit maximisation?<sup>4</sup> Is this a job for AI or a job that should be given to Black women who most likely already know a solution?

This poetic guide relies on historical evidence that:

**If Black women aren't mentioned, that means Black women will be harmed. If Black people aren't mentioned,**

3. Ibid., at p.10.

4. Benjamin, R. (2019) *Race After Technology: abolitionist tools for the new Jim Code*. Cambridge, UK: Polity Press. p.30.

5. See note 2, at p.9.

that means Black people will be harmed. If gender minorities aren't mentioned, that means gender minorities will be harmed. If any historically marginalised group is not mentioned, that group will be harmed.

Any marginalised group can ask these (and more) questions and it's important that we do. Current analyses of fairness in AI often flatten the nuance between—and variety in—Black women's experiences.<sup>5</sup> While flattening identities may make it easier for AI, it harms Black women and leads to further AI disillusionment. If the field of AI is to gain public trust, it will have to prove its benefits to all people.

Do we need AI or  
Do we need Black feminisms?  
Liberation should lead,  
Technology should support.  
There is no mechanical solution  
To sin.  
There is only  
The purposeful striving towards  
Justice

Serena Dokuaa Oduro is a writer and policy researcher aiming for algorithms to uplift Black communities. She is the Policy Research Analyst at Data & Society.

# How (not) to blog

James  
Vincent



**about  
an  
intelligent  
toothbrush**

News of the world's first toothbrush with artificial intelligence arrived in my email in January 2017. It was part of the tide of announcements accompanying the start of the world's biggest tech trade show, CES; an annual inundation that floods journalists' inboxes with the regularity of the Nile, leaving behind a rich and fertile substrate that's perfect for raising some blog posts.

It was not the stupidest pitch I received that year, nor was it the first time someone had tried to sell me a gadget that claimed to have "embedded AI" when it very obviously didn't. But it sticks in my memory all

There was something so the same. nonsensical, so inadvertently Dada about the phrase, that I remember getting up from my desk and making a cup of coffee to process it in mild wonder. The world's first *toothbrush* with artificial intelligence, I reflected, waiting for the kettle to boil. The world's first toothbrush with *artificial intelligence*, I pondered, returning the milk to the fridge.

Now, I freely confess to occasionally writing tech stories with ridiculous headlines just for the pleasure of lobbing said headline into the world like a can- taloupe into

the Grand Canyon. Any tech journalist who says they don't do this is either lying or has forgotten how to have fun. People love to dunk on stupid gadgets and tech writers love to help them. (Remember: it's not click-bait if you can get the gist of the story from the headline alone. If you read, groan, then click to find out more about this or that stupid gadget out of a sense of morbid curiosity, then that's just a good headline.)

To write these headlines, most of the time you don't need to do more than state what it is the gadget does. The whole point of this sort of coverage—indeed, one suspects, the whole point of this sort of gadget—is the self-evident silliness. The products become more enticing the more superfluous they are, like a variant of a Veblen good in which demand for an item increases as it becomes more expensive. From a journalist's point of view, this usually means you can avoid adding snark to the headline and let the thing's idiocy speak for itself. People on social media will quote-tweet you with delight, pointing out the daftness of the product and perhaps adding some comment about "late capitalism" for good measure. All this as if the Victorians didn't also fill their lives with air-conditioned top hats and mechanical hairbrushes. "Kohler's smart toilet promises a 'fully-immersive experience'" is one such



headline I remember with fondness from CES 2019—the silliness of the semi-submerged pun matching, I hoped, the product’s stupidity.

It’s still  
 analysts  
 idiotic  
 be kept  
 coverage.  
 question

important, of course, for tech jour-  
 to write responsibly about even  
 gadgets, and two things should  
 in mind when creating this sort of  
 First, whether the product in  
 is so obviously a scam or so

potentially  
 harmful that  
 covering it in  
 any way is unethical. And second, the  
 degree of credulity with which you write  
 about claims being made about the  
 technology. Is this a silly gadget or is  
 it an actual con?

For the AI toothbrush this left me  
 with a bit of a problem. (I’d decided  
 while drinking my coffee that it was too  
 foolish not to write up.) It seemed like a  
 real product, no worries on that score, but  
 I was uneasy about repeating claims about  
 the device’s “embedded AI”. Would people  
 get that this was a stupid thing to say,  
 or would they think, given the level  
 of ambient hype surrounding arti-  
 ficial intelligence, that this was a  
 legitimate  
 breakthrough of  
 some sort?

I'd found this dilemma to be a common problem writing about AI. The term "AI" itself is so rotten with ambiguity, so overburdened with varying connotations and expectations, that it cannot be trusted to support much linguistic weight. Depend on your readers to parse the phrase "artificial intelligence" this way or that and its meaning will collapse underfoot. This is both symptom and cause of the febrile atmosphere that surrounds discussion of AI—a misunderstanding that is productive for generating hype and selling snake oil, but not much else. If an artificial intelligence researcher, tech journalist, and lay reader all have different expectations when reading those two letters AI, clear communication about the subject is always going to be difficult.

So what to do about the toothbrush? As I read through the press release, Googled the company that made the gadget and scoured their website for something approaching detailed tech specs, it seemed—as expected—that what was being passed off as "artificial intelligence" was just a teaspoon of embedded computation: the thing was only smart compared to rocks.

What it did was to record when it was long for, and, roughly, the mouth covered by brushing. It refined this data into charts and sent the results to a smartphone app. It then revealed along the lines “brush more” or “brush less” how much you brushed. It was the sort of information that a stopwatch could not provide without making any claims to intelligence. A little further Googling even revealed that the gadget’s creators had announced a near-identical product in 2014. The only major change was that they’d dropped the term “smart” in favour of the more fashionable “AI”. It seemed this was just another company climbing aboard the artificial intelligence hype train, but what else had I expected?

In the end, I wrote a short little blog post, barely 300 words, with the headline “This smart toothbrush claims to have its very own ‘embedded AI’”. I hoped this formulation—the scare quotes and “clā:ms”—would neutralize any potential deception, but I’m not sure it really did. Later in the day, a little annoyed I’d covered the thing at all, I wrote a longer article castigating superfluous “AI gadgets” in

use sensors used, how the parts of each brush-graphs and ing insights delivered of “brush depending on It was the sort stopwatch could ing any claims inge. A little even revealed that ators had announced product in 2014. The change was that they’d term “smart” in favour fashionable “AI”. It seemed another company climbing artificial intelligence hype train, but what else had I expected?

general. More searching dug up “AI-enhanced” alarm clocks, TVs, beds, headphones, washing machines, dishwashers, fridges, and more—each product trading off the same ambiguity surrounding artificial intelligence to sell basic digital functionality as something exciting and new. Plenty of these gadgets attracted press coverage—some critical, some credulous—but browsing through it all I felt that no matter what had been written, it only added to the confusion. Taken together it all seemed like epistemological chaff—meaning thrown to the winds. The world’s first toothbrush with artificial intelligence had only ever been a single point in a scattered and chaotic *melée*.

James Vincent is a senior reporter for The Verge who covers artificial intelligence and other things. He lives in London and loves to blog.



**Learn  
to**



**take on  
the ceiling**  
Alexander  
Reben

Whether it's forming a scientific hypothesis or seeing deities burnt into toast, humans are hard-wired to find patterns in the world. We are adept at making sense of the chaotic or unknown from information we already know.

This can elevate understanding as much as it can mislead. One consequence is our tendency to project the human onto the artificial, a phenomenon called anthropomorphism. Similarly, ascribing animal-like qualities to non-animals is zoomorphism. This can result in curious behaviour. People give names to their robot vacuums, which they hate to see struggling in corners. Apple users thank Siri when she answers a question.

A car owner feels subject to a personal attack by their broken-down car.

Advanced technologies are particularly ripe for this sort of misperception, as operating mechanisms are often opaque, capabilities are increasingly human or animal-like, and the results can seem magical to those who do not have insight into the design. Indeed, as soon as room-sized computers were first developed, they were commonly referred to as "electronic brains". Obviously, they were

nothing like brains, but the phrase sounded super futuristic, gave the public a grasp of their functions, and grabbed more attention than something bland, like “Electronic Numerical Integrator and Computer”.

While metaphorical descriptions which equate machine capabilities with human traits may initially appear harmless, they can lend more agency to the technology than is warranted. If you read the headline “Scientists Invent Giant Brain” you would imagine much about that device based on your understanding of a brain and its functions. When we ascribe human traits to technology, those traits often come with assumptions of abilities which may not be true. For example, even if an AI system produces human-like text, it cannot be assumed that it also has other capabilities associated with language, such as comprehension, perception or thought. Nevertheless, most of us do this unconsciously and automatically. If we are aware of our tendency to anthropomorphise, it’s easier to notice when this sort of bias kicks in, making us recognise advanced technology for what it is.

Equally, concepts drawn from sci-fi are often used as a shorthand to describe (and sometimes overrepresent) the capabilities of technology. Probably journalists do this because it creates a quick and familiar image in the reader’s mind. (More likely, it is to chase clicks.)



One example is when self-balancing skateboards were called “hoverboards” in reference to the movie, *Back to the Future*. Since the device worked using wheels which were firmly planted on the ground, no hovering was involved at all. Still, “hoverboard” was a better buzzword than, say, “self-balancing transport”. The implication was that sci-fi technology was here now, even though it was clearly not.

Something similar has happened with artificial intelligence. There are two important categories of AI. Artificial intelligence (AI), which is a type of machine learning that carries out a specific task, and artificial general intelligence (AGI), which is a system that is capable of doing any non-physical task as well as, or better than, a human. AI in a car can determine whether or not there is a stop sign in an image captured by a camera. An AGI can do the same while also being the CEO of the company that builds the cars and writing flawless poetry.

Unfortunately, in media reporting, AI is sometimes treated like AGI. Take the example of an AI chatbot developed by Facebook which sometimes formed incomprehensible English output when two such AIs “spoke” to each other. This was an artefact of the system attempting to find a more efficient way to communicate, a sort of shorthand. Here’s an example of the two chatbots negotiating how to split up basketballs between them:

Bob: i can i i everything else

Alice: balls have zero to me to me to me to me to  
me to me to me to me to

Bob: you i have everything else

Alice: balls have a ball to me to me to me to me  
to me to me to me to me

**This kind of semantic soup became such an obstacle to the intended function of the system, they stopped research.**

**A human analogue to this AI chat might be drawn with military acronyms or scientific terms, which make sense to those within a particular communication system, but not to those outside it. However, alarmist headlines evoking images of a movie hero “killing” an AI poised to destroy humanity were used to report the incident: “Facebook engineers panic, pull plug on AI after bots develop their own language”.**

This kind of sensationalism leads to misunderstanding and fear, when the reality of AI is often far more mundane.

Another phenomenon which contributes to the misrepresentation of AI's capabilities is "cherry picking", whereby only the best system outputs are shared. The rest are hidden or omitted, and no mention is made of the winnowing process. Sometimes this is intentional, such as when artists generate AI outputs for artworks. Other times this can be a tactic used to intentionally overrepresent capabilities. This might be done to make systems appear better than they are, to gather more attention and support, or to make the system seem more interesting to potential investors.

Within an artistic context, where an artist works with the system to make a creative output, "cherry picking" is more akin to curation. For example, in one of my own artworks, I trained an AI on fortune cookie fortunes so it could generate new ones. I then chose outputs I liked and had them printed and enfolded in actual fortune cookies.

"Your dreams are worth your best pants  
when you wish you'd given love a chance"

"I am a bad situation"

“Your pain is the essence of everything and the value of nothing”

“Today, your mouth shut”

**Fairly unusual fortunes which make sense, if you don't think about them too hard. Many others which I left out didn't catch my eye initially. Yet they kind of have meaning, if you bend your mind a bit:**

“There is a time when you least expect it”

“Learn to take on the ceiling”

“Emulate what you need to relax today”

“You are what you want”

**Then there are the really nonsensical ones:**

“Nappies and politicians need to listen to”

“You are a practical person with whom you exchanged suggestive glances?”

“You will be out of an ox”

“The greatest war sometimes isn’t on the ground even though friends flatter you”

Cherry picking without disclosing that you are doing so leads to a misrepresentation of current AI systems. They are already impressive and improving all the time, but they are not *quite* what we see in sci-fi.

So next time you encounter or read about an AI, ask yourself the following questions: Am I projecting human capabilities onto it which may be false? Are sci-fi ideas and popular media clouding my perceptions? Are the outputs cherry picked? We need to keep in mind that if current AI is to AGI as “self-balancing transports” are to hoverboards, we are still a long way from getting off the ground.

Alexander Reben is an MIT-trained artist and technologist who explores the inherently human nature of the artificial.

LDR 0.0272

IG 5.6622

E 10300

A stylized, high-contrast map of the world in black and white, with the text "Uses (and abuses)" overlaid in the center. The map uses a dot-matrix or grid-like pattern to represent landmasses, with the text in a bold, sans-serif font.

# Uses (and abuses)



Gemma  
Milne

of  
hype



Hype has always been a part of my career. As a science and technology researcher I spend time following the science startup receiving end I've helped noticed, and obscures until we

writer and a lot of my technology and scene. I've been on the of bombastic pitches, people desperate to get I wrote a book on how hype the future. I believe that reckon with the dual nature of hype, snake oil and pseudoscience will persist in  $\Delta$ , and science and technology more generally.

Reactions to hype range from excitement at the genuine progress that  $\Delta$  has made, to scorn at lame sales pitches. The word has a variety of definitions too, depending on who you ask. Some see it as unfaithful to the truth, a lie of sorts. Others consider it as merely an ultra-subjective form of expression, a form of "fair play" marketing, as it were.

To me, hype is a tool for capturing attention. Hype is exaggerated publicity which inflates expectations and prompts emotions such as excitement or fear, which can stop people in their tracks. It's a little like a magic show. By tugging on the emotions, hype causes people to lend their ears, eyes, and brain waves to ideas, claims,

and sales pitches in an uncritical manner. The tricks are wondrous, but they're entertainment.

The use of hype comes with a warning label though. Forget that it's entertainment and it becomes lying, usually for nefarious ends. Furthermore, hype can result in accidental fooling. Out of context, hyped messages can very easily be misinterpreted. The originator may not have set out to trick or deceive, however, the fact remains that because of its spread, the language used, or the cultural context, hype can be misinterpreted as truth.

Yet the role of hype is more significant than simply capturing attention at all costs. It eases the complicated, arduous, and expensive journeys between research lab, office meeting room or simply someone's brain, out to the market and the public. It lubricates, bridges, and facilitates. Hype can be a strategic tool for selling ideas, products or political campaigns. It helps attract people, pennies, policies, and partnerships.

Hype is, in and of itself, a creator of value—regardless of the substance behind it.

Hype attracts attention, money, good employees, and social capital. Startups want the attention of investors. NGOs want to attract funding. Thought leaders want to lead thought. Researchers want

**better-  
funded labs.**

**Universities want to be higher on the league tables. Companies want more clients. To align yourself with “innovative” narratives gives you a head over the competition.**

**That’s why hype is unlikely to disappear. It is genuinely useful for generating attention—a scarce resource in today’s attention economy. But where there’s opportunity, there’s opportunists, and so the existence of a snake oil is unsurprising. Its prevalence, though, is dependent on the rest of us not reckoning with the problematic, dual nature of hype. This needs to happen on two levels.**

**First, those on the receiving end of hype need to collectively recognise the power, mechanics, and sheer volume of it. The sole purpose of a hyped-up message is to grab attention. So when those on the receiving end do not scrutinise the message further, a lack of context, or the absence of deeper understanding about the topic can easily result in accidental fooling. There are also places and circumstances where hype must be checked at the door entirely—otherwise the public, journalists, investors, and government**

ministers will continue to find separating snake oil from legitimate claims extremely difficult. Hype about technology, for instance, has no place in serious policy deliberations about how to regulate technology.

Enhanced critical thinking and a better understanding of how hype and marketing “work”

can help reduce hype’s problematic impact. Its power, after all, is in its illusion, so seeing it for what it is can help stop it in its tracks.

Still, that’s not enough to stem or limit the volume of hype that is produced, especially in AI.

Secondly, those of us who work in science and technology—as builders, policy-makers, commentators, researchers, investors—must reckon with our own complicity in the use of hype. This requires introspection, acceptance,

and possibly a pinch of low-level shame. Hype helps projects move forward. We might not use those words, but we know it to be true.

We could argue that hype is just a tool used on the road to greatness, that this lubricant is just

a reality of getting through that complicated, arduous, and expensive idea-to-market journey. We can justify its wielding like advocates of placebos justify lying to unsuspecting patients. The cost of accidental fooling that hype comes with is worthwhile as it's the only way to get stuff done in the dog-eat-dog world of the innovation industry. The ends justify the means when saving the world is at stake.

This pragmatic view neglects the dual nature of hype. The very lubricant that eases ideas through complex systems is what allows snake oil to make it to market. Hype is useful to capture attention in today's attention economy, but snake oil is an inevitable by-product. Those who generate, amplify or simply cash in on the hype surrounding AI inevitably also create the conditions in which snake oil vendors and pseudoscience can thrive.

## Lack of

### care characterises a

lot of the problems around hype in science and technology. Those on the receiving end who do not listen might be ignorant, flippant or time-poor. Then there's the fact that those creating the hype are often willingly complicit; carelessly wielding a tool which makes them look credible. If we reckon with these two forces, snake oil will start to have less success.

Beyond that, what's needed is a genuine culture change in technology and science. The pressure to publish at all costs in order to progress in academia, the precarious working conditions in journalism, and the growing competition for funding are all systemic forces that create conditions in which problematic hype can thrive. This culture often equates value with attention, and links status and fame to trendy topics or cults of personality, rather than proof of positive social impact. Publish or perish should itself die.

Perhaps what we need is more boring <sup>AV</sup>. More cautious product pitches. More thoughtful leaders who communicate uncertainty about their own thoughts.

**Snake oil in AI will thrive until the various forms of hype that keep it trendy are deeply and unashamedly reckoned with by those who arguably stand to benefit most from no-one saying anything at all.**

Gemma Milne is a Scottish science and technology writer and PhD researcher in Science & Technology Studies at University College London. Her debut book is *Smoke & Mirrors: How Hype Obscures the Future and How to See Past It* (2020).

LDF 0.0536

LDR 0.0493





# Talking heads



Crofton  
Black

In  
the  
13th century,  
it was said, the scientist and philosopher Albertus Magnus built a talking head. This contraption, skilfully manufactured with hidden wheels and other machinery, spoke so clearly that it frightened his pupil, who smashed it.<sup>1</sup>

1. Anon (1627)  
*The famous  
historie of Fryer Bacon.*  
London: A.E. for  
Francis Grove.

Like us, the people of the Middle Ages told themselves myths about technology. One

particular mythic strand involved talking heads.

Like today's systems, these heads received inputs from their builders and offered outputs that told of things to come. However, the heads failed as predictive systems. The tales warned of automated data operations gone wrong.

Such fables probably resonated with their audiences for several reasons. One is that although there were many known examples of automata that were capable of moving or performing other functions, the possession of a voice—the sign of a reasoning soul—crossed a significant threshold. Another is that these systems occupied an ambiguous, or sometimes overtly transgressive, place in the taxonomy of technology.

In the Middle Ages, technology—the “science of craft”—embraced three categories, which we might for convenience call exoteric, esoteric, and demonic.

Exoteric technologies, such as mills, bows, cranks, and wheelbarrows, operated in accordance with physical laws which were well

understood and clearly visible to all. They were uncontroversial. Esoteric technologies, on the other hand, occupied the space of so-called "natural" magic. This involved secret or hidden mechanisms, which were not commonly known or understood. Those who exploited these mechanisms argued that they were nonetheless natural, because they operated through certain intrinsic features of the universe.

The workers of natural magic construed the universe in terms of analogies. They perceived relationships between heavenly and earthly bodies, between stars and comets, animals and plants, and they believed in a cause and effect straddling these analogies. Not unlike, perhaps, the data analysts of today, they built their universe through proxies.

The danger in the practice of natural magic, as witnessed in writings and controversies across the centuries of the early modern period, was that the line dividing it from the third category of technology, demonic magic, was highly contested. Demonic magic comprised any activities for which no natural explanation could be found, and which therefore had to be the work of demons, persuaded by scholars via ritual to do their bidding. Martin Delrio, for example, recounting the fate

LDF 0.069

of Albert's head in his *Investigations into Magic* (1599), concluded that the head could not be the product of human ingenuity. This was impossible, he said, for such human-manufactured idols "have a mouth but speak not".<sup>2</sup>

Therefore it was a demon that spoke.

In the absence of plausible explanation, there was a black box, and in the box was thought to be a demon. The story of Albert's

head is about a scientific—literally, knowledge-making—endeavour which ended in the construction of an object of horror.

Before Albert there was Gerbert of Aurillac.

Gerbert, it is recounted by the historian William of Malmesbury, constructed his head at the apex of a career investigating computational and predictive systems.<sup>3</sup> He had mastered judicial astrology, the analysis of the effects of the movements of heavenly bodies on human lives. He had "learned what the singing and the flight of birds portended". And as a

2. De' Corsini, M. (1845) *Rosaio della vita: trattato morale*. Florence: Società Poligrafica Italiana.

3. Delrio, M. (1657) *Disquisitionum magicarum libri sex*. Cologne: Sumptibus Petri Henningii.

mathematician he had established new ways of manipulating the abacus, developing for it “rules which are scarcely understood even by laborious computers”.

The head was capable of responding in binary to a question, answering “yes” or “no”. But its responses, which lacked context and strictly adhered to narrowly defined inputs, were misleading. Gerbert, who lived in Rome, had asked the head whether he would die before visiting Jerusalem. The head replied that he would not. But it failed to specify that there was a church in Rome of that name, where Gerbert did indeed go. Soon after, he died.

Likewise, it was said that in Oxford in the 13th century, the philosopher Roger Bacon built a bronze talking head

as a component in a defence system which would render England invulnerable.<sup>4</sup> Yet after building the head, he fell asleep, leaving a servant to monitor it, whereupon the head uttered three obscure phrases and exploded. The servant, who was blamed for the debacle, argued that a parrot with the same duration of training would have spoken more sense. Bacon himself was convinced that the head had something profound to communicate. What this was, no one would ever know, because only the machine's builder was able to interpret it. To those charged with its management, its findings were opaque.

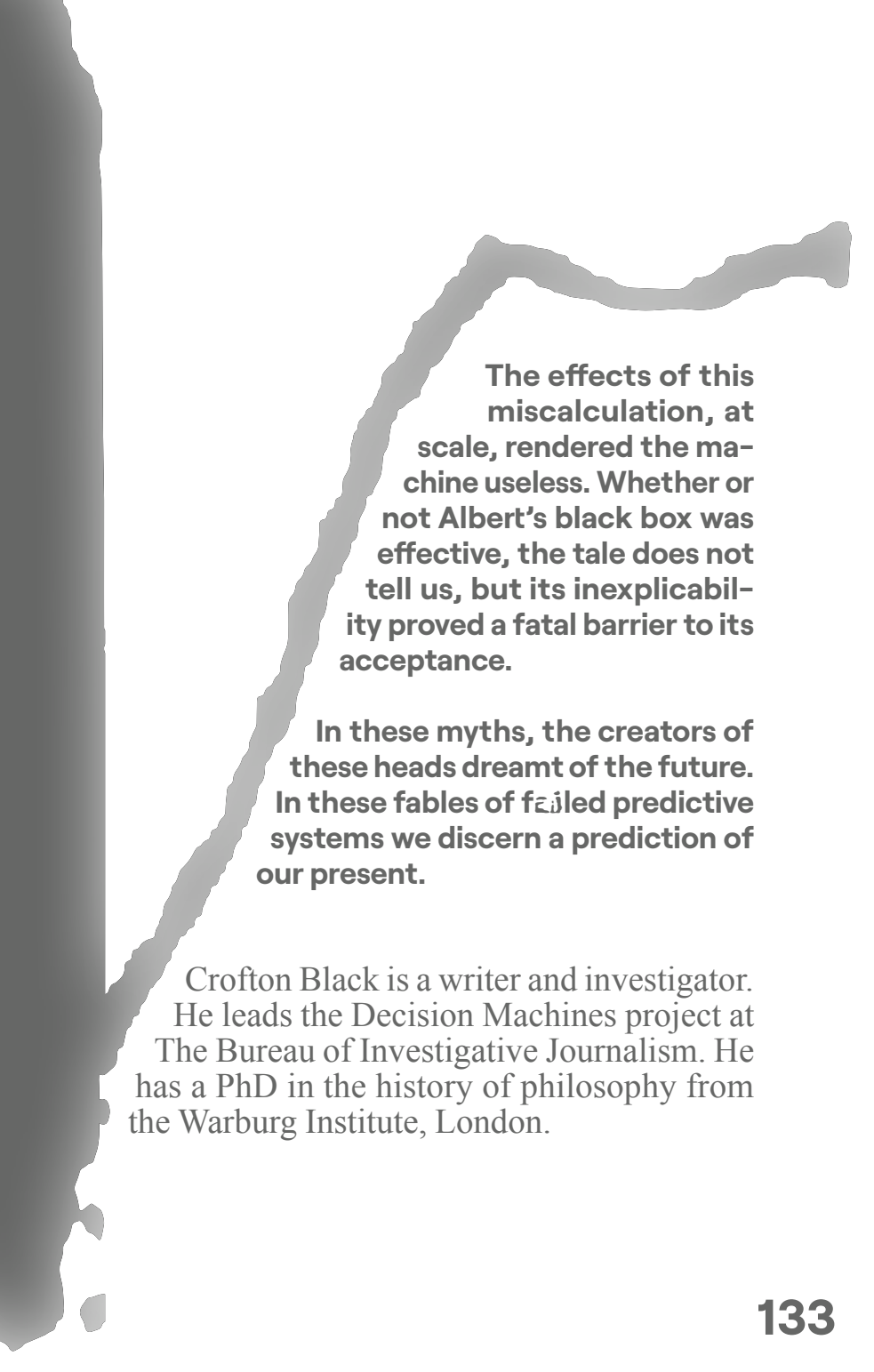
4. Ed. Giles, J. A. (1847) *William of Malmesbury's Chronicle of the Kings of England*. London: Henry G. Bohn.



4  
A story told of Bacon's teacher, Robert Grosseteste, runs parallel. He, too, built a head, and worked on it for seven years. But his concentration flagged for 30 seconds as he sought to finish the task. A lapse in the internal mechanism or the underlying data? We are not told. As a result, he "loste all that he hadde do".<sup>5</sup>

LG 6.3247  
E 13600  
These fables, centuries old, offer us patterns which we may find familiar: allegories of modern encounters with predictive systems. Gerbert failed to appreciate how his system's outputs were conditioned by unperceived ambiguities in its inputs. A flaw in the underlying dataset produced misleading results. Bacon left his system running under the management of someone not trained to understand its outputs. Consequently, nothing useful was derived from them. Grosseteste's machine failed owing to a very small programming discrepancy, a mere half-minute's worth in seven years.

5. Ed. Macaulay, G. C. (1899) *The Complete Works of John Gower*. Oxford: Clarendon Press.



**The effects of this miscalculation, at scale, rendered the machine useless. Whether or not Albert's black box was effective, the tale does not tell us, but its inexplicability proved a fatal barrier to its acceptance.**

**In these myths, the creators of these heads dreamt of the future. In these fables of failed predictive systems we discern a prediction of our present.**

Crofton Black is a writer and investigator. He leads the Decision Machines project at The Bureau of Investigative Journalism. He has a PhD in the history of philosophy from the Warburg Institute, London.





**What is**

**a face?**

Adam  
Harvey

135

Until the 20th century, faces could only be seen through the human perceptual system. As face recognition and face analysis technologies developed throughout the last century, a new optical system emerged that “sees” faces through digital sensors and algorithms. Today, computer vision has given rise to new ways of looking at each other that build on 19th century conceptions of photography, but implement 21st century perceptual technologies. Of course, seeing faces through multi-spectral modalities at superhuman resolutions is not the same as seeing a face with one’s own eyes. Clarifying ambiguities and creating new lexicons around face recognition are therefore important steps in regulating biometric surveillance technologies. This effort begins with exploring answers to a seemingly simple question: what is a face?

The word “face” generally refers to the front-most region of the uppermost part of the human body.

However, there are no dictionary definitions that strictly define where the face begins or where it ends. Moreover, to “see” hinges on an anthropocentric definition of vision which assumes a visible spectrum of reflected light received through the retina, visual cortex, and fusiform gyrus that is ultimately understood as a face in the mind of the observer. Early biometric face analysis systems built on this definition by combining the human perceptual system with statistical probabilities. But the “face” in a face recognition system takes on new meanings through the faceless computational logic of computer vision.

Computer vision requires strict definitions. Face detection algorithms define faces with exactness, although each algorithm may define these parameters in different ways. For example, in 2001, Paul Viola and Michael Jones<sup>1</sup> introduced the first widely-used face

1. Viola, P. & Jones, M.J. (2004) Robust Real-Time Face Detection. *International Journal of Computer Vision* 57, 137–154. DOI: 10.1023/B:VISI.0000013087.49260.fb

detection algorithm that defined a frontal face within a square region using a  $24 \times 24$  pixel grayscale definition. The next widely used face detection algorithm, based on Dalal and Triggs' Histogram of Oriented Gradients (HoG) algorithm,<sup>2</sup> was later implemented in dlib<sup>3</sup> and looked for faces at  $80 \times 80$  pixels in greyscale. Though in both cases images could be upscaled or down-scaled, neither performed well at resolutions below  $40 \times 40$  pixels. Recently, convolutional neural network research has redefined the technical meaning of face. Algorithms can now reliably detect faces smaller than 20 pixels in height,<sup>4</sup> while new face recognition datasets, such as TinyFace, aim to develop low-resolution face recognition algorithms that can recognise an individual at around  $20 \times 16$  pixels.<sup>5</sup>

2. Dalal, N. & Triggs, B. (2005) Histograms of oriented gradients for human detection, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 886-893 vol.1. DOI: 10.1109/CVPR.2005.177

3. <https://github.com/davisking/dlib>

4. Hu, P. & Ramanan, D. (2017) Finding Tiny Faces, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1522-1530. DOI: 10.1109/CVPR.2017.166

5. <https://qmultinyface.github.io>

**Other face definitions include the ISO/IEC 19794-5 biometric passport guidance**

for using a minimum of 120 interocular (pupil to pupil) pixels. But in 2017, America's National Institute of Science and Technology (NIST) published a recommendation for using between 70-300 interocular pixels. However, NIST also noted that "the best accuracy is obtained from faces appearing in turnstile video clips with mean minimum and maximum interocular distances of 20 and 55 pixels respectively."<sup>6</sup> More recently, Clearview, a company that provides face recognition technologies to law enforcement agencies, claimed to use a 110 × 110 pixel definition of a face.<sup>7</sup> These wide-ranging technical definitions are already being challenged by the face masks adopted during the Covid-19 pandemic, which have led to significant drops in face recognition performance.

The face is not a single biometric but includes many sub-biometrics, each with varying levels of identification possibilities. During the 2020 International Face Performance Conference, NIST proposed expanding the concept of a biometric face template to include an additional template for the periocular region,<sup>8</sup> the

6. <https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing>

7. Clearview AI's founder Hoan Ton-That speaks out [Interview], <https://www.youtube.com/watch?v=q-1bR3P9RAw>



region not covered by a medical mask.

This brings into question what exactly “face” recognition is if it is only analysing a sub-region, or sub-biometric, of the face. What would peri-ocular face recognition be called in the profile view, if not ocular recognition? The distinction between face recognition and iris recognition is becoming increasingly thin.

8. <https://www.nist.gov/video/international-face-performance-conference-ifpc-2020-day-1-part-1>

At a high enough resolution, everything looks basically unique. Conversely, at a low enough resolution everything looks basically the same. In *Defeating Image Obfuscation with Deep Learning*, Macpherson et al. found that face recognition performance dropped between 15–20% as the image resolution decreased from  $224 \times 224$  down to  $14 \times 14$  pixels, using rank-1 and rank-5 metrics respectively, with a dataset size of only 530 people. But as the number of identities in a matching database increases, so do the inaccuracies. Million-scale recognition at  $14 \times 14$  pixels is simply not possible.

LDF 0.0215

LDR 0.0339

**Limiting the resolution of biometric systems used for mass surveillance could contribute profoundly towards preventing their misuse, but such regulation requires unfolding the technical language further and making room for legitimate uses, such as consumer applications or investiga-**

**tive journalism, while simultaneously blunting the capabilities of authoritarian police or rogue surveillance companies.**

**Calls to ban face recognition would be better served by replacing ambiguous terminology with more technically precise language about the resolution, region of interest, spectral bandwidth, and quantity of biometric samples. If law enforcement agencies were restricted to using low-resolution 14 × 14 pixel face images to search against a database of millions, the list of potential matches would likely include thousands or tens of thousands of faces, making the software virtually useless, given the operator would still have to manually sort through thousands of images. In effect this would defeat the purpose of face recognition,**

LG 6.5214

E 39820

which is to compress the search space into a human-scale dataset in order to find the needle in the haystack. Severely restricting the resolution of a face recognition system means that searching for a needle would only yield more haystacks.

In 2020, a project by University, called PULSE, researchers at Duke showed that the restricted perceptual space of a  $16 \times 16$  pixel face for wildly different identities allows all downscale to perceptually different images.<sup>9</sup> The project faced criticism because it also showed that up-sampling a low-resolution image of Barack Obama produced a high-resolution image of a light-skinned face. Nevertheless, the work confirmed the technical reality that two faces can appear identical at  $16 \times 16$  pixels, but resemble completely different identities at  $1024 \times 1024$  pixels. As image resolution decreases so too does the dimensionality of identity.

9. <http://pulse.cs.duke.edu/>

Unless a computational definition of “face” can be appended to the current language around


biometrics, the unchallenged ambiguities between a  $16 \times 16$  pixel face and a  $1024 \times 1024$  pixel face will likely be decided by the industries or agencies that stand to benefit most from the increasingly invasive acquisition of biometric data. Moreover, better regulatory definitions of “face” that include specific limits on the resolution of face imagery could help limit the potential for face recognition technologies to be used for mass surveillance. The monitoring and tracking of our every public move—at meetings, in classrooms, at sporting events, and even through car windows—is no longer limited to law enforcement agencies. Many individuals are already scraping the internet in order to create their own face recognition systems.

A better definition of “face” in recognition technology should not be limited to sampling resolution, but should also include the spectral capacity, duration of capture, and the boundaries of where a face begins and where

it ends. As the combinatory resolution of a face decreases, so does its potential for mass surveillance. Limiting resolution means limiting power and its abuses.

Adam Harvey is a researcher and artist based in Berlin. His most recent project, *Exposing.ai*, analyses the information supply chains of face recognition training datasets.





# Why automated content



**moderation  
won't  
save us**

Andrew  
Strait



LDF 0.1409

LDR 0.0754

LG 3.3473

F 5340

When I was a content moderator at Google in the mid 2010s, there were days I wished a machine could do my job. It was usually on the boring days spent scanning through monotonous copyright complaints with thousands of URLs, or the bad days clearing out queues of child sexual abuse content (they would always refill by the next day).

Content moderation is difficult work, often exciting but occasionally damaging to the soul. The platforms I moderated were (and still are) used by billions of people across the world as libraries, news sources, and public squares, but they were not built for these purposes. Rather, these platforms were designed to incentivise user engagement, content sharing, and mass data collection—objectives that

often made moderating harmful content feel like throwing cups of water on a raging inferno. When I did this work, a small number of automated tools helped prioritise and flag certain types of content for review, but human moderators like me were ultimately charged with putting out the daily blazes. Over the years I worked at Google, the copyright complaints grew longer and the queues of traumatising content continued to grow. Surely, I wondered, only the increased use of AI could help moderate these platforms at scale?

Nearly a decade later, the heads of large tech firms and global policymakers share the same vision. In the wake of high-profile incidents like the Christchurch mosque shooting and the Rohingya genocide in Myanmar, policymakers across the globe have called for tech firms to remove hate speech, terrorist content, and other forms of harmful speech at

increasingly breakneck speed. In response, tech platforms have invested heavily in AI-based moderation tools to identify, predict, and remove harmful content within minutes—if not at the time of upload. Called before the US Congress in 2018 over election misinformation, Mark Zuckerberg declared that “over the long term, building AI tools is going to be the scalable way to identify and root out most of this harmful content.” Facebook now boasts that 94.7% of hate speech removed from their platform in the third quarter of 2020 was proactively identified using their automated tools. Has the dream of automated moderation come true? Or do statistics like the ones above mask deeper problems with platform moderation?

Sadly, automated moderation tools have become yet another example of the misplaced hope in artificial intelligence to solve complex human problems. By framing online

safety as simply a matter of moving from “human to automated”, tech firms and policymakers risk exacerbating known accountability and transparency issues with platform moderation policies while distracting themselves from more urgent questions around platform design features and business incentives. The hype around automated moderation tools has largely overlooked their technical limitations, the hidden labour of human moderators, and the increasing opacity of platform decision-making.

The technical limitations of automated moderation tools are well known to research and civil society communities. Notoriously bad at identifying the nuance and context of online speech, these systems routinely fail to identify whether a video constitutes illegal copyright infringement or lawful parody, or whether a post with a racial slur is written by a victim of a hate crime or their assailant. Due to their reliance on historically labelled content, these tools fail to keep pace with the constant evolution of human language, such as the shifting codewords used in QAnon misinformation campaigns. Some systems exhibit serious issues of language bias—for example, researchers have found Google’s

LP 3.0.0756

LG 4.7575

E.5620

**Perspective API tool, which uses machine learning to predict the “toxicity” of certain content, penalises content written in African American Vernacular English.<sup>1</sup> Addressing these issues would require not only a paradigm shift in AI research but a fundamental reconstitution of the labour forces designing these tools to incorporate more diverse perspectives.**

**The torrent of online misinformation sparked by the Covid-19 pandemic laid these limitations bare like never before. Not trusting contracted moderators to take corporate devices home and work remotely, large platforms like**

**Facebook and YouTube resorted to fully automating many moderation decisions. The results were alarming. Child exploitation and self-harm removals on Facebook fell by at least 40%,**

I. Sap, M., Card, D., Gabriel, S., Choi, Y. & Smith, N. A. (2019) The Risk of Racial Bias in Hate Speech Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. DOI: 10.18653/v1/P19-1163

pages reporting factual Covid-19 information were misidentified as misinformation, and YouTube appeals of wrongfully removed accounts skyrocketed.<sup>2</sup> Rather than improving outcomes, the results highlighted how automated tools can worsen the already inadequate *status quo* of platform moderation practices.

2. Scott, M. & Kayali, L. (2020, October 21) What happened when humans stopped managing social media content.

*Politico.*

<https://www.politico.eu/article/>

[facebook-content-moderation-automation/](https://www.politico.eu/article/facebook-content-moderation-automation/)

The decision to jettison human moderators during Covid-19 also reflects the changing role of the moderator as platforms have grown in size and scale. In my time doing this work, most moderators were full-time employees respected as trusted experts in particular regions or products, encouraged to deliberate on tricky edge cases. Today, tech platforms increasingly treat moderators as an expendable and invisible labour source whose decisions are used as fuel to train automated moderation tools. The vast majority of moderators at large platforms today are temporary contract labourers outsourced from third-party agencies,

in countries that include India, the Philippines, and Malaysia. They are low-paid, relative to full-time employees of large tech firms, and lack the same career advancement opportunities and access to high-quality mental healthcare that is necessary for the traumatising aspects of their work.<sup>3</sup> These moderators

3. Roberts, S. T. (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press.

are hired, constantly assessed, and fired on the basis of their ability to render decisions quickly (often within seconds) and consistently with rigid platform-wide policies. They are not encouraged to bring deliberative nuance or contextual expertise to a decision, or to question whether a policy that applies to billions of people in hundreds of countries is adequate for the specific complaint they are viewing. It is their hidden emotional labour that keep platforms and automated moderation tools

a float. Yet these moderators remain profoundly undervalued, their welfare and expertise pushed to the side in the rush to moderate at scale.

The rush towards automation also risks further obfuscating already opaque moderation processes, exacerbating platform transparency and accountability issues. As public reliance on platforms has grown, so too has their incontestable power to determine what speech is acceptable online. Freedom of expression is increasingly reliant on secretive and unaccountable business practices. The classified nature of moderation policies makes it virtually impossible to assess whether automated tools are effective at keeping users safe. Platforms have been hesitant to make their policies public, and industry transparency reports and experiments like Facebook's Oversight Board, an independent committee that reviews a tiny portion of Facebook's moderation decisions (but importantly, not its content moderation policies or design features) offer narrow, self-selected forms of transparency.

Automated moderation tools risk making this situation far worse. Their decisions are difficult to audit, assess, and understand even for developers,



## let alone for regula-

tors or third-party researchers who struggle to gain access to them. These tools are often built to meet the needs of different “governance stakeholders”, which

may not align with

the interests of

users or national

laws.<sup>4</sup> To give one

example, the

Syrian Archive, an

open source initiative to document war

crimes in the country, has routinely battled

YouTube’s algorithm to disable terrorist

content which routinely fails to differ-

entiate between videos glorifying vio-

lence and those documenting abuses.<sup>5</sup>

Leaving the decision

is allowed on the

box tools and black

with no independ-

or assessment, will

diminish the account-

platforms and render

cost of automated

of what speech

web to black

box policies,

ent oversight

only further

ability of large

the societal

tools invisible.

Rather than a dream

increasing reliance

automation tools

ing another AI night-

mare. Platforms and

regulators must not

be constrained to a

come true, the

on automated

risks becom-

4. Gorwa, R., Binns, R. & Katzenbach, C. (2020)

Algorithmic content moderation: technical and political challenges in the automation of platform governance. *Big Data & Society*,

January–

June:

1–15,

DOI:

10.1177/

20539517

19897945

5. O’Flaherty, K. (2018, June 26) YouTube keeps deleting evidence of Syrian chemical weapon attacks. *WIRED*.

<https://www.wired.co.uk/article/chemical-weapons-in-syria-youtube-algorithm-delete-video>

narrative of “human vs. automated” moderation. Instead, they must ask a broader set of questions. For example, how can platforms be designed to encourage safe behaviour? Is limiting their size and scale a more promising solution? Rather than following profit-driven metrics like maximising user engagement or increasing content virality, what if platforms redesigned their affordances and features with safety front and centre? Rather than looping moderators in at the last minute to put out fires, what if firms included them in the design of these products from the outset?

And rather than exploiting outsourced moderators, further burying

their emotional labour, what if tech firms properly compensated and cared for their well-being as full-time employees?

As with many other AI technologies, the hype around automated moderation tools reflects a misplaced belief that complex sociopolitical problems can be adequately resolved through technical solutions. AI is tremendously helpful at addressing well-defined and concrete tasks, but determining what speech is acceptable for billions of people is anything but a well-defined challenge. It is understandable that we wish it were so—how nice would it be to simply let a machine police our speech for us without changing the scale, practices, or affordances of Facebook, Twitter or YouTube?

Andrew Strait is a former Legal Policy Specialist at Google and works on technology policy issues. He holds an MSc in Social Science of the Internet from the Oxford Internet Institute.

The background of the image is a grayscale photograph of a city skyline, featuring several tall buildings. The image has a 'torn paper' or 'cutout' effect, with irregular, jagged white edges that make it look like a piece of paper has been ripped out of a larger sheet. The text is overlaid on this background in a bold, black, sans-serif font.

**Conso-  
lidating  
power  
in the  
name of  
progress:  
techno-  
solutionism**

**and  
farmer  
protests  
in India**

Tulsi  
Parida

Aparna  
Ashok

## TECHNO-SOLUTIONISM AND JUGAAD

In many low and middle-income economies, the adoption of new technologies such as artificial intelligence is often done quickly and aggressively, with little oversight, in an effort to “catch up” with more advanced economies. The prevailing narrative is that countries still wading through the debris of a colonial past need rapid technological advancements for economic development. Recent digital policy consultations in countries such as India reflect an inclination to accelerate AI adoption even in the absence of essential regulatory protections.

1. <https://niti.gov.in/national-strategy-artificial-intelligence>

**Artificial Intel-  
lignce high-  
lighted the “transformative impact” of  
AI and its potential to better the econ-  
omy, society, and environment.<sup>1</sup> As is the  
case with countries the world over, AI is  
seen as a key component of devel-  
opment. However, India  
lacks regulatory meas-  
ures to protect data pri-  
vacy, which is essential  
to individual and con-  
sumer rights. Privacy as  
a fundamental human  
right was established  
in India only in 2017,  
and a personal data  
protection bill is yet to  
be passed. Both gov-  
ernment and industry  
utilise a techno-solutionist narrative<sup>2</sup>  
to limited regulation and sometimes  
promote deregulation of sectors, which  
can result in substantial infractions of  
human rights and human development.**

**In India, techno-solutionism is often  
paired with the concept of *jugaad*, the  
notion that low-cost innovations can  
solve local problems.<sup>3</sup> Examples include  
everything from rejigging a truck so**

## **The Indian 2018 National Strategy on ligence high-**

2. Tucker, I.  
(2018, March 22)

Evgeny Morozov:

“We are abandoning all the checks  
and balances.” *The Guardian*.

<https://www.theguardian.com/technology/2013/mar/09/evgeny-morozov-technology-solutionism-interview>

3. Saha, P. K. (2018, June 16)  
‘Jugaad’ culture: The good, the  
bad and the ugly. *Mint*. <https://www.livemint.com/Leisure/IZhsWLKRpWIQjnnAiGNNKM/Jugaad-culture-The-good-the-bad-and-the-ugly.html>



it can power a rural village's electricity, to Mangalyaan, India's successful Mars orbiter mission

achieved with a meagre budget and home-grown technologies. On the surface, *jugaad* seems an ingenious and agile approach to development. However, the *jugaad* mentality is eerily similar to Silicon Valley's own "move fast and break things". While *jugaad* might have few consequences at a small scale, when applied to technologies such as AI that operate at a vast scale, without adequate oversight, this mentality can have devastating consequences.

## AI AND POWER IMBALANCES

"AI hype" for economic development is not unique to emerging economies. The unique mix of *jugaad* and unregulated techno-solutionism in India, however, could result in rapid technological progress with a big down-side. When coupled with *jugaad*, unregulated techno-solutionism is dangerous in its belief that band-aid solutions and hastily created technologies are needed in the name of economic development. Without adequate protections,

the country's ambitious AI adoption plans could result in catastrophic power imbalances<sup>4</sup> that come with a tremendous social cost. One of the known potential risks of AI systems is the data-related power imbalances that pit individuals against large conglomerates that have substantial resources at their disposal.

In an ideal world, the benefits and risks of AI would be distributed equitably amongst firms and society. However, growing evidence

shows a small number of firms have a substantial share of the AI market, and without consequential changes, those very firms will also have a disproportionate share of the benefits of AI while society bears the brunt of the risks.<sup>5</sup> Exploitative business models with an inequitable distribution of the risks and benefits of AI give rise to a consolidation of power amongst those institutions that hoard data and use AI unchecked.

4. O'Keefe, C., Cihon, P., Flynn, C., Garfinkel, B., Leung, J. & Dafoe, A. (2020) *The Windfall Clause: Distributing the Benefits of AI*. *Centre for the Governance of AI Research Report*. Future of Humanity Institute, University of Oxford. <https://www.fhi.ox.ac.uk/windfallclause/>

5. Cheney-Lippold, J. (2017) *We Are Data*. New York, NY: New York University Press.

However, when designed responsibly, in a context-appropriate manner, with adequate internal governance and external regulation,

AI technologies can play a role in solving problems at scale. For example, AI applied in diagnostics can compensate for the lack of skilled professionals needed to make healthcare accessible in rural parts of India. Especially for countries in the nascent stages of AI development, it is important to scrutinise proposed solutions with a socio-technical lens and critically evaluate the potential of power imbalances and foreseeable positive and negative consequences.

## LOOKING AHEAD: AGRI-TECH AND FARMER PROTESTS

The combination of deregulation, the *jugaad* mindset, and techno-solutionist narratives together create the ideal conditions for disproportionate power imbalances. A case in point is the “technological disruption” that is currently underway in India’s agricultural sector.

One of India’s largest conglomerates has recently ventured into agri-tech, claiming that

6. Kadidal, A. (2020, January 4) AI must invade agri to help India prosper: Experts. *Deccan Herald*. <https://www.deccanherald.com/city/ai-must-invade-agri-to-help-india-prosper-experts-791491.html>

AI can improve efficiencies in the disorganised agricultural sector.<sup>6</sup> In early 2020, the company announced an agri-tech platform claiming to help farmers make data-driven decisions in farming practices, and connect them directly to suppliers. With this proposed platform, this corporation, which also owns India's largest supermarket chain and India's largest telecom services, could conceivably gain complete control of supply and demand chains—with farmers relying on the app to sell their produce via the company's telecom service directly to the company's retail stores.

Regulation to prevent such a consolidation of power is necessary to protect farmers' rights. Instead, in late 2020, the Indian government actively deregulated the sector by passing three heavily contested agricultural bills to liberalise farming and "attract private sectors and foreign direct investment".<sup>7</sup> The new laws allow for (1) farmers to trade freely, (2) farmers to enter into contract farming, and (3) deregulation of selected essential commodities, such as cereals, edible oil, oilseeds, pulses, onion etc. The bills were passed hastily during the Covid-19 pandemic and without consultation with farmers.

7. Ministry of Consumer Affairs, Food & Public Distribution. (2020, September 22) *Parliament passes the Essential Commodities (Amendment) Bill, 2020* [Press release]. <https://pib.gov.in/PressReleasePage>

8. Damodaran, H. (2020, December 31) Explained: The concerns of farmers, and what Centre can negotiate to end protests. *The Indian Express*. <https://indianexpress.com/article/explained/farmers-big-concern-and-what-govt-could-negotiate-7073291/>

**The new laws have been criticised as “corporate-friendly and anti-farmer”, and this controversy has given rise to**

**one of the largest labour movements in the world—a 250 million-strong strike in support of farmer unions. The demands are clear. The three bills must be repealed. A minimum price and state procurement of crops must be made a legal right.<sup>8</sup> Farmers have also made the connection between the new laws and corporate power. They have called for a boycott of the company’s products, with many farmers porting their mobile service to other providers.<sup>9</sup>**

9. Arora, K. (2020, December 27) Rising Number of Protesting

Farmers Switch From Jio to Rival Mobile Networks. *The Wire*. <https://thewire.in/agriculture/farmer-protest-jio-networks-airtel-vodafone-idea>

**This example of technological solutionism in the agriculture sector is emblematic**

of the inequitable distribution of the risks and benefits of technology systems imposed on society. When the people who are being disenfranchised by AI systems demand protection, whose concerns will be heeded? India strives to be a role model for countries in the region, and these choices could set a precedent. Gaining the trust of citizens through the equitable distribution of risks and benefits would be helpful for AI adoption. At the time of writing, the outcome of the farmer protests is uncertain—as are the effects that the laws and this movement will have on those most disadvantaged, namely landless Dalit farmers. Still, one thing is clear: farmers in the world's largest democracy are ready to fight the imposition of emergent systems of oppression.

Tulsi Parida is a socio-technologist currently working on AI and data policy in fintech. Her previous work has been in edtech, with a focus on responsible and inclusive learning solutions.

Aparna Ashok is an anthropologist, service designer, and AI ethics researcher. She specialises in ethical design of automated decision-making systems.



**When  
fintech  
meets  
60 million**

**unbanked  
citizens**



Favour  
Borokini  
Ridwan  
Oloyede



**In Nigeria, AI is being touted by some as a one-size-fits-all solution to the country's inefficiencies and woes.<sup>1</sup> An unfortunate combination of sclerotic (and occasionally regressive) domestic development with Western-influenced tech solutionism has resulted in a burgeoning fascination with the technology. Nowhere, perhaps, is this more readily evident than in the fintech industry.**

1. Editorial Board. (2021, March 8). Nigeria Needs Artificial Intelligence to Combat Insecurity, Says Expert. *The Guardian Nigeria*. <https://guardian.ng/news/nigeria-needs-artificial-intelligence-to-combat-insecurity-says-expert>

**Fintech, a portmanteau word that describes the integration of financial services with technology, is often deployed by companies offering financial services to penetrate new markets. Following a series of high profile funding successes, especially that of the “country’s latest unicorn”,<sup>2</sup> the term has become a buzzword in Nigeria.**

2. Jackson, T. (2021, March) Nigerian Payments Startup Flutterwave Achieves ‘Unicorn’ Status after \$170m Funding Round. *Disrupt Africa*.

• <https://disrupt-africa.com/2021/03/10/nigerian-payments-startup-flutterwave-achieves-unicorn-status-after-170m-funding-round/>

3. Kola-Oyeneyin, E., Kuyoro, M., & Olanrewaju, T. (2000, September 23). Harnessing Nigeria’s fintech potential. *McKinsey & Company*. <https://www.mckinsey.com/featured-insights/middle-east-and-africa/harnessing-nigerias-fintech-potential>

**According to a 2020 McKinsey report,<sup>3</sup> Nigeria is home to over 200 standalone fintech companies. This figure does not include the dizzying array of fintech offerings by Nigerian brick-and-mortar banks. Yet, Nigeria is one of seven countries that contribute to nearly half of the world’s unbanked population, totalling**

4. Ventura, L. (2021, February 17). Global Finance Magazine - World's Most Unbanked Countries 2021. *Global Finance Magazine*. <https://www.gfmag.com/global-data/economic-data/worlds-most-unbanked-countries>

5. Forty percent of Nigerians live below the poverty line: Report. (2020, May 4). *AlJazeera*. <https://www.aljazeera.com/economy/2020/5/4/forty-percent-of-nigerians-live-below-the-poverty-line-report>

6. Osakwe, S. (2021, April 6). How Is Nigeria's National Financial Inclusion Strategy Going? *Center for Financial Inclusion*. <https://www.centerforfinancialinclusion.org/how-is-nigerias-national-financial-inclusion-strategy-going>

about 60% of her adult population.<sup>4</sup> In part, this is due to the sheer number of people living in severe poverty (83 million compared to India's 73 million).<sup>5</sup> This is exacerbated by a conservative and seemingly erratic economic policy and a lack of access to financial services, particularly in rural areas.<sup>6</sup> With the median age of Nigerians being around 18 years, home-grown fintechs are stepping up to the challenge of providing financial inclusion services to its youthful population.<sup>7</sup> By using targeted advertisements, appealing offerings, and other digital marketing strategies, fintechs may appear to have a better chance of reaching these young people than conventional banks.

Certain Nigerian fintechs, which offer their services to individuals or corporate clients, claim to leverage "machine learning algorithms" and

**“AI-powered facial recognition” to assess credit risk, prevent fraud, generate personality profile reports, and for identity verification.<sup>8</sup> Several fintechs use these products. A closer look at the environments in which these solutions are deployed raises serious questions about whether they can deliver on their promises to improve financial inclusion. Instead, the real risk seems to be that these solutions could exclude minorities and privilege profit at the expense of individuals’ rights and liberties. To identify and profile the ideal customer, these companies typically access the personal data of private individuals from the government, through third-party service providers, or harvest it directly by sifting through personal information on mobile devices, such as text messages, fine and coarse location, media contents, contact lists, social media, and use of trackers and other permissions. One company, which promises to empower Africans “by driving social and financial inclusion!” currently pays an**

7. Roland Berger Strategy Consultants. (2012, January). National Financial Inclusion Strategy. *Central Bank of Nigeria*. <https://www.cbn.gov.ng/Out/2012/publications/reports/dfd/CBN-Summary%20Report%20of-Financial%20Inclusion%20in%20Nigeria-final.pdf>

8. Onukwue, A. O. (2020, November 19). The BackEnd: Meet the “Palantir of Africa.” *Techcabal*. <https://tehcabal.com/2020/11/19/the-backend-analytics-intelligence-palantir-africa/>

**undisclosed amount to Nigerian government agencies to gain access to millions of Nigerians' national identity data.<sup>9</sup> The company's publicly**

9. Hersey, F. (2020, August 5). Verify my life: could a Nigerian problem lead to a global trust solution? (Or fuel a two-tier society?). *Biometric Update*. <https://www.biometricupdate.com/202008/verify-my-life-could-a-nigerian-problem-lead-to-a-global-trust-solution>

**stated goal is to promote trust through digital identity and verification services. Another company claims to help “banks distinguish between a photograph in an ID and a selfie”, to which end it has “created an identity management system that harnesses the powers of facial recognition technology.” This system connects to government databases and non-government databases, such as that of the Nigeria Inter-Bank Settlement System (NIBSS). Both companies promise that AI-powered facial recognition systems will improve integrity and better identify fraud by “knowing all details about the client in near real-time.”**

**The kind of access required by these systems is incredibly invasive. Ultimately amounting**

to surveillance of the activities of private citizens, it is potentially in breach of rights enshrined in the constitution and other laws, including freedom from discrimination, privacy, data protection and dignity. As other authors in this book have argued, these systems, which are not known for their accuracy or fairness, could be unfairly prejudiced against persons based on their gender, socio-economic background or other discriminatory factors. More troubling is the prospect of fintechs ending up as gatekeepers, shutting out the excluded groups they claim to include. If, for example, an individual is unable to purchase a smartphone (or electricity or internet to make use of it) and one of these systems, therefore, cannot automatically extract data, it will rank them as undesirable, further excluding them from access to credit and financial products.

According to a report published by Tech Hive Advisory about the pervasive practices of digital lenders in Nigeria,<sup>10</sup>

seven of the 22 mobile applications analysed publicly disclose that they use AI to determine borrowers' credit-worthiness. Only one mentioned the existence of profiling in its privacy notice, as is required by law.<sup>11</sup>

10. Tech Hive Advisory (2021, February)

Digital lending: Inside the pervasive practice of LendTechs in Nigeria. *LinkedIn*. [https://www.linkedin.com/posts/tech-hive-advisory\\_digital-lending-inside-the-practice-of-lendtechs-activity-6768431134297620480-E5zx](https://www.linkedin.com/posts/tech-hive-advisory_digital-lending-inside-the-practice-of-lendtechs-activity-6768431134297620480-E5zx)

- [linkedin.com/posts/tech-hive-advisory\\_digital-lending-inside-the-practice-of-lendtechs-activity-6768431134297620480-E5zx](https://www.linkedin.com/posts/tech-hive-advisory_digital-lending-inside-the-practice-of-lendtechs-activity-6768431134297620480-E5zx)
- [advisory\\_digital-lending-inside-the-practice-of-lendtechs-activity-6768431134297620480-E5zx](https://www.linkedin.com/posts/tech-hive-advisory_digital-lending-inside-the-practice-of-lendtechs-activity-6768431134297620480-E5zx)

11. Article 3.1(7)(L) of the Nigeria Data Protection Regulation (NDPR).

Besides this disregard for the law, the report also notes an alarmingly common lack of algorithmic transparency, lack of explainability, lack of accountability, and an absence of information on purpose limitation of the data used.

It appears that the solutionism being sold by fintechs and other players like digital identity providers, namely the belief that AI can single-handedly fix structural deficiencies, has been bought hook, line and sinker by those who ought to be the last to do so—the government. Neglecting or outsourcing civic obligations (such as digital identity registration, and in particular, the development of robust financial inclusion policies) to profit-driven private enterprises running machine learning algorithms without sufficient safeguards will undoubtedly worsen already-existing inequalities by systematically breaking down civil rights and freedoms. One such example is the licensing of private entities to access national identity biometric data held by the National Identity Management Commission, some of which claim to use artificial intelligence and facial recognition.<sup>12</sup> The details of the agreement



12. National Identity Management Commission (2020, 15 December) Public Notice: Approved Data Capturing Agents (Digital Identity Ecosystem) [Press release]. <https://nimc.gov.ng/public-notice-approved-data-capturing-agents-digital-identity-ecosystem>

are not available publicly. No record of a data protection impact assessment being conducted has either been publicly stated or published. These and other machine learning algorithms that make predictions based on historical events and data cannot be at the forefront of providing the forward-looking information we need for our future.

While AI has its uses in industry—fintechs included—it cannot



replace a comprehensive financial inclusion and development approach. Certainly not when spurred on by a lack of transparency and accountability. The performance, validation and deployment of AI must be ethical and meet existing legal requirements.

Fairness and transparency must determine limits to how data is used and the algorithms that are deployed. Security, privacy, data protection, and accountability about how data is used and by whom is critical. Most importantly, it is essential to understand users, their needs, and the context in which these technologies are being used.

Favour Borokini is a tech policy researcher interested in (emerging) technology-facilitated violence against women and the development and deployment of AI in Africa.

Ridwan Oloyede is a Co-Founder at Tech Hive Advisory, where he focuses on global data protection and privacy laws, digital health, and digital ethics, among other issues.





# Algorithmic registers and their limitations

**as a  
governance  
practice**  
Fieke Jansen  
Corinne Cath



Europe has been lured in by the siren call of artificial intelligence. Public debate is characterised by snake oil promises of AI's benefits to the economy, social welfare, and urban development. Here, "AI" is a catch-all phrase used to describe a wide-ranging set of technologies, most of which apply statistical modelling to find patterns in large data sets and make predictions based on those patterns. Concerns raised about the unpredictable nature and possible societal harms of AI models have given rise to a policy doctrine of ethical and procedural safeguards, the idea being that AI's "great" potential can be harnessed and its harms mitigated by implementing safeguarding principles of non-binding fairness, accountability, and transparency. Building on our work as researchers and practitioners in the field of technology and society, we

**will discuss one of these safeguards, namely algorithmic registers—websites that show what, where and how AI is used in a city. Extolled by some in the AI ethics community as an example of good AI governance, we argue that voluntary ethical and procedural safeguards in fact perpetuate the hype and neutralise important critical debate.**

## **ALGORITHMIC REGISTERS IN EUROPE**

**In line with these ethical activities, a number of European cities<sup>1</sup> are experimenting with algorithmic registers run by local municipalities. In September 2020, the cities of Amsterdam and Helsinki launched their registers to increase transparency around the deployment of AI in the public sector. These databases collect information about how AI systems are used in an open and voluntary fashion, which should provide insights into the local uses of these systems. Both the Helsinki and**

1. Johnson, K. (2020, September 28) Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI. *Venturebeat*. <https://venturebeat.com/2020/09/28/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/>

1335

the Amsterdam register contain only five entries. The entries function on an opt-in basis and mostly cover automated government services, including city libraries, hospitals, crowd management and parking control. A little over two weeks after the launch of these registers in the autumn of 2020, prominent AI ethics scholar Luciano Floridi published an editorial letter in *Philosophy & Technology*,<sup>2</sup> in which he heralded them as solutions for the many challenges of AI governance, not least those related to public accountability and trust in AI.

There are a number of governance assumptions attributed to the registers that we seek to question, especially regarding the *ex-post*, or after the fact, framework of “accountability through transparency” contained within the register concept. Some of the most

2. Floridi, L. (2020) Artificial Intelligence as a Public Service: Learning from Amsterdam and Helsinki. *Philosophy & Technology*, 33(4), 541-546. DOI: 10.1007/s13347-020-00434-3

3. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability, and Transparency* (pp. 77-91) <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>; Amnesty International (2020) Netherlands: We sense trouble: Automated discrimination and mass surveillance in predictive policing in the Netherlands. <https://www.amnesty.org/en/documents/eur35/2971/2020/en/>

harmful AI applications are evidently missing from the Amsterdam and Helsinki registers. The databases contain no mention of known welfare and law enforcement applications.

For example, the Amsterdam object detection entry, in which the city is

experimenting with GAIT recognition for crowd monitoring purposes, does not account for the police facial recognition trials taking place in these same locations. This means that some of the most sensitive applications of AI, often implicated in algorithmic discrimination,<sup>3</sup> are not currently covered by the registers, and it is unclear if they will be in the future.

The lack of critical engagement by AI proponents with these information voids in the registers is telling of their inability to function as a transparency tool between the city and its residents. By defining accountability as transparency through voluntary registration, proponents of algorithmic registers are essentially taking



an *ex-post* approach to governance too: push AI systems as a public utility first (“Just do it”) and ask for forgiveness later. This reinforces the assumptions that AI is neutral and should be used “for the greater good”, and neutralises criticism as simply a matter of imperfect information. This governance-by-database eschews difficult conversations about whether AI systems should be implemented at all, and how

these systems are advancing punitive politics that primarily target already vulnerable populations.

### PERPETUATING THE AI HYPE

What is even more telling than the registers themselves, however, is the lack of critical engagement by AI proponents with the power structures and political ideologies that shape these governance-by-database solutions. Isolating governance mechanisms outside their social, political, and economic context allows for the perpetuation of a discourse that reaffirms the arbitrary notion of “AI

for good". This ignores the fact that most algorithms in use for urban management pre-date the idea of registers and are deeply rooted in a political ideology and organisational culture bent towards efficiency and cost reduction. Efforts to abstract and generalise AI accountability frameworks allow their proponents to move beyond the messy nature of reality and to further depoliticise AI by replacing the outdated idea that technology is "neutral" with the notion that the "great" potential of AI can be harnessed when harms are mitigated through voluntary procedural and ethical safeguards. Lauding the registers without understanding their context discounts concerns about the negative impact of AI on society, because it is this which aligns safeguards with the political environment and commercial interests that are enabling the AI hype.

The deployment of AI for public services, from the critical (like urban infrastructure, law enforcement, banking, healthcare, and humanitarian aid) to the more mundane (like parking

control and libraries), should be done with great caution. Not least as these systems are often implemented on top of, and in line with, existing neoliberal politics that are punitive and extractive. As such, AI systems cannot be seen as isolated from the context in which they are deployed. When looking at these registers, and other opt-in accountability frameworks, we need to consider what is missing and who is steering the conversation. Due to the complex nature of governance, registers are only a partial and, at times, overhyped answer to ensuring public accountability for the application of AI systems. Indeed, the *ex-post* model bypasses critical conversations about the root motivations for rolling out AI systems, and who is truly served by them.

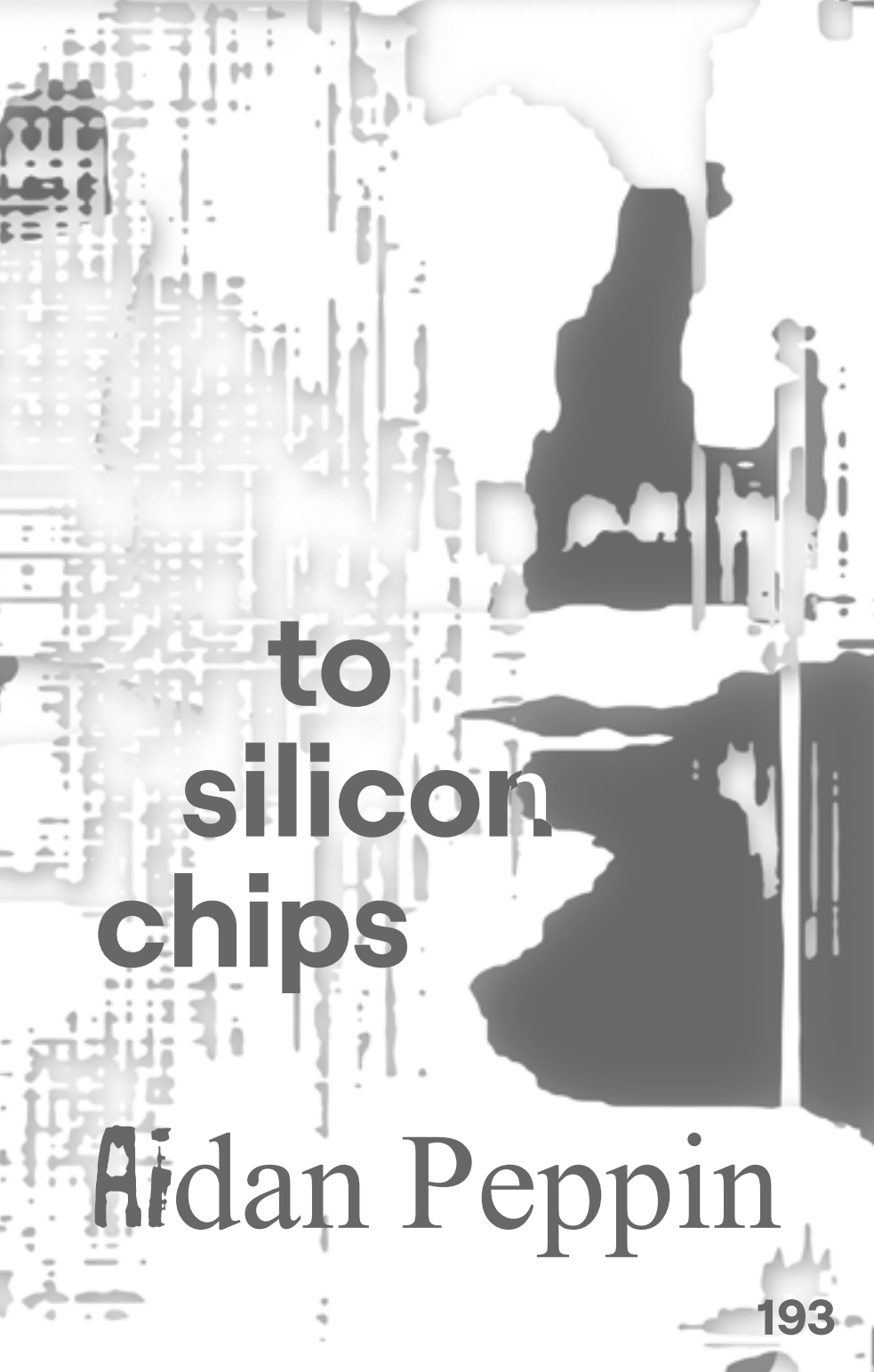
Fieke Jansen is a doctoral candidate at Cardiff University. Her research is part of the Data Justice project funded by ERC Starting Grant (no.759903).

Dr. Corinne Cath is a recent graduate of the Oxford Internet Institute's doctoral programme.





**The  
power  
of  
resistance:  
from  
plutonium  
rods**



**to  
silicon  
chips**

Aidan Peppin

LDF 0.036

LDR 0.1561

LG 4.5144

E 31420

Around the same time that Turing was devising his famous test, many countries were pouring concrete into the foundations of the world's first nuclear power stations. Nuclear power then was a technology that, like AI today, generated as much hope and hype as it did anxiety.

Today, the abundant hype around AI is being met with growing resistance towards harmful algorithms and irresponsible data practices. Researchers are building mountains of evidence on the harms that flawed AI tools may cause if left unchecked. Civil rights campaigners are winning cases against AI-driven systems. Teenagers scorned by biased grading software are carrying placards and shouting "F\*\*\* the Algorithm!". Public trust in AI is on rocky ground.

This puts AI's current success on a potential tipping point. Fuelled by misleading AI "snake oil", today's unrealised hype and very real harms are creating public resistance that could push AI into another stagnant winter that stifles both its harms and its benefits—much like those of the 1970s and 80s, when AI innovation was curbed and political will failed. But resistance also highlights what responsible, publicly acceptable technology practice could look like, and could help prioritise people and society over profit and power. To understand how, nuclear power's history of hype and resistance offers a useful guide. After all, the term "AI winter" was in part inspired by the idea of nuclear winter.<sup>1</sup>

1. Crevier, D. (1993)  
*AI: The Tumultuous  
Search for Artificial  
Intelligence*. New  
York, NY: Basic  
Books. p. 203

Despite the lethal dangers of nuclear technology being made horrifyingly evident at Nagasaki and Hiroshima in the closing days of the Second World War, the following years saw nuclear's military applications moulded to civil purposes. In the early 1950s, its potential for good was manifested in the world's very first nuclear power station at Calder Hall in Cumbria, UK, now known as Sellafield.



Since then, trends in public acceptance of nuclear power have been well-studied, and thousands of people across the world have been surveyed. These studies have shown how, in nuclear power's early years, the fearsome image of mushroom clouds was perceived as the distant past, while civil nuclear power was promised by governments and the industry as not only the future of electricity generation, but as the herald of a brave, new, technological society built on unlimited energy.<sup>2</sup> However, many hyped-up promises never materialised (such as Ford's nuclear-powered vehicle, the *Nucleon*) and the fear of nuclear war lingered.

Then, in 1986, the potent, dangerous reality of nuclear power was reproven when reactor No.4 at the Chernobyl nuclear power station failed catastrophically. The combined technical, chemical, institutional, and administrative break-down killed thousands according to official estimates, and spewed radiation across thousands of square miles (reaching as far as Sellafield in Cumbria, in a bitter irony).

In the wake of the Chernobyl disaster, support for—  
and investment

2. Gamson, W.A. & Modigliani, A. (1989) Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach. *American Journal of Sociology*, 95 (1), 1-37.

LDR 0.0251

in—civil nuclear power plummeted. Fears of nuclear’s destructive force were absorbed into the growing environmental movement, and the early enthusiasm for nuclear power swiftly gave way to a “dark age” for the technology, in which resistance grew and innovation stumbled. Nuclear power saw little new activity until the mid-2000s, when some nations, including the UK, France, US and China, began

(and continue)  
to reinvest in  
nuclear  
power,  
in

part thanks to public recognition that it generates low-carbon electricity.

LG 6.1414

Though the public is often dismissed as having little influence on the material shape of technology, the history of nuclear power shows that public opinion can have a profound impact on policy and practice—that is, how governments invest in and regulate technologies, how innovators develop them, and how people use them. This is seen with many other technologies too, from genetic engineering to cars, and now in AI.

580

The recent kindling of public resistance against AI must be recognised as the vital signal that it is. Previous AI winters of the late 20th Century

- followed seasons of over-promising: while

computing hardware was catching up with AI theory, excitement waned as the pace of innovation slowed and hyped-up claims about what AI could do were left unfulfilled. Today, however, superfast processors, networked cloud servers, and trillions of bytes of data allow many of those claims to be apparently realised. AI systems are helping detect and diagnose macular degeneration, powering GPS systems, and recommending the next best thing since *Tiger King*.

But success  
is a double-edged sword.

As AI's abilities are put to use in everyday applications, so too are its dangers. The dubious data practices of firms like Cambridge Analytica have disrupted politics, algorithms like COMPASS have exacerbated pre-existing bias and injustice, and facial recognition systems deployed by law enforcement are infringing on civil and data rights. As people realise they are largely

powerless to challenge and change these systems by themselves, public concern and resistance has been spreading to all corners of the AI landscape, led by civil rights campaigners, researchers, activists, and protesters. If ignored, this resistance won't just weed out harmful AI. It may stifle the social and political will that supports beneficial AI too.

Social scientists remind us that correlation is not causation. Simply because public support and resistance to nuclear power has neatly tracked levels of government investment and technological innovation does not mean one creates the other. But an established body of research shows that they *are* connected, and that public resistance is an important signal.

After Chernobyl, public resistance to nuclear power made clear that its risks and dangers were no longer considered acceptable. This had a chilling effect on political will and investment. It wasn't until industry and governments committed to more responsible, transparent, and accountable practices that investment and innovation in nuclear power began to reignite. Even today, nuclear power is not a widely accepted technology—public support is rocked still by incidents like the 2011 Fukushima disaster—and healthy public scepticism plays an active part in balancing hype against responsible practice.

## Silicon chips and plutonium

fuel rods do not share much in common technically, and the threat of nuclear meltdown differs from the threats posed by biased algorithms. But we cannot let the hope that AI may never have a disaster on the same scale as Chernobyl lure us into complacency.

The many small, individual disasters that are already occurring every day around us will continue to add up. As irresponsible AI practices lead to social, political, cultural, and economic harms, public acceptance will falter and resistance will grow.

However, resistance is not a force to fear: it is a powerful signal. It may threaten the current hype-fuelled AI summer, but it need not stifle responsible innovation. Harnessed well, public resistance

can help shine light on what must be improved, weed out, AI “snake oil”, define what is socially acceptable, and help a more responsible AI industry flourish. But if it is ignored, the current public mood around algorithms and big data could forecast more than just the winds of change. It could be the first cold breeze of another AI winter.

Aidan Peppin is a Senior Researcher at the Ada Lovelace Institute. He researches the relationship between society and technology, and brings public voices to ethical issues of data and AI.

## DESIGN NOTES: AI-constrained Design

The design of this book explores the use of artificial intelligence to conceptualise its main visual elements. From an optimistic standpoint of AI abundance, the book design is a semi-fiction, a staged and micromanaged use of a GAN (Generative Adversarial Network), an unsupervised machine learning framework, where two neural networks contest with each other in order to generate visual output. Stretching the narrative, this book could be framed as *at/the (first) book designed by an AI*. In this scenario, the collaborating AI (more like the *AI-as-head-of-design-that-doesn't-know-how-to-design*), has informed, but also constrained the possibilities to work visually with the pages.

The design strategy adopts the *Wizard of Oz Technique*, a method originated from interaction design where what is seemingly autonomous, is in reality disguising the work of humans 'as a proxy for the system behind the scenes'<sup>1</sup>. The use of the GAN, which a reader could expect as a simplification, a symbol of technological ergonomics, has instead complicated the process. As a result, the contents contort around the spaces that the AI *imagination* left them to exist, revealing an apparently spontaneous visual language.

The book features results from two separate datasets, addressing the overall layout composition, and a (overly sensitive) recognition algorithm which targets all instances of 'AI, ai, Ai', regardless of their position or meaning.

### MetaGAN v.3 Layouts

The dataset used to produce the compositions above is a collection of book scans. The purpose of an image GAN is to create new instances by detecting, deconstructing and subsequently reconstructing existing patterns to create speculations about continuations. Reusing existing layout materials, conceived by human creativity, opens up the discussion of AI creativity. The outcomes, which could be perceived as surprising, original and novel, are however subject to human selection and valuation. In training the MetaGAN, the dissimilarity of the data points, in combination with the small size of the dataset (200 images), led to the idiosyncrasy of overfitting. An overfitted model generates outcomes 'that correspond too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably'<sup>2</sup>.





zi	ai	ai	fi	is	ai	ai	ai
ai	ia	ai	ia	is	ai	ia	fi
is	ai	ia	ai	ia	ai	ai	ai
ai	is	ai	ai	ia	ia	ai	ai
ai	ai	ia	ia	is	ai	is	ai
is	ai	ai	ia	ia	ai	ai	ai
ai	is	ai	ai	is	ia	ia	ia
ai	ia	ai	ai	ai	ia	ai	ia
ai	ia	is	ia	ai	ai	is	is
is	ia	ai	ai	ai	is	ai	is
is	ia	is	ia	ia	is	ai	ai
ia	ia	ai	ia	ai	ai	is	ai
ai	ai	ia	ai	ai	ia	ia	ia
ai	ai	ia	ia	fi	is	is	ia

## AI type results

These AI letterings are results of a GAN using a dataset containing logos from various AI related brands (or belonging to Anguilla, whose country code top-level domain is '.ai'). The use of these characters is indeed automated in the design of the book, but it is done using GREP styles.

## References:

1. Bella, M. & Hanington, B., 2012. Universal Methods of Design, Beverly, MA: Rockport Publishers. p204
2. <https://www.lexico.com/definition/overfitting>

Other publications by Meatspace Press include:

Taylor, L., Sharma, G., Martin, A., and Jameson, S. (eds) 2020. *Data Justice and COVID-19: Global Perspectives*.

Graham, M., Kitchin, R., Mattern, S., and Shaw, J. (eds). 2019. *How to Run a City Like Amazon, and Other Fables*.

Graham, M. and Shaw, J. (eds). 2017. *Towards a Fairer Gig Economy*.

Shaw, J. and Graham, M. (eds). 2017. *Our Digital Rights to the City*.

All Meatspace Press publications listed above are free to download, or can be ordered in print from [meatspacepress.com](http://meatspacepress.com)

From predicting criminality to sexual orientation, fake and deeply flawed Artificial Intelligence (AI) is rampant. Amidst this feverishly hyped atmosphere, this book interrogates the rise and fall of AI hype, pseudoscience and snake oil. Bringing together different perspectives and voices from across disciplines and countries, it draws connections between injustices inflicted by inappropriate AI. Each chapter unpacks lazy and harmful assumptions made by developers when designing AI tools and systems, and examines the existential underpinnings of the technology itself to ask: why are there so many useless, and even dangerously flawed, AI systems?

ISBN 978-1-913824-02-0



9 781913 824020 >



[www.meatspacepress.com](http://www.meatspacepress.com)