

# AI Forensics

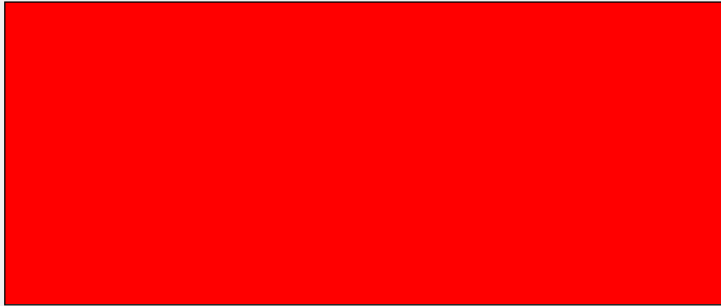
Dr Eleanor Dare

Supervisor: Dr Leonardo Impett

Project Coordinator: Patrick Riechert

## Introduction

The following table is a thought experiment to establish whether you are a dualist or a monist.



What colour is this rectangle? If you were a colour-blind world expert on colour who suddenly gained the ability to see colours, would you have more knowledge of colour than you did before?

This is a variation on a quite famous thought experiment called 'Mary's Room', 'Mary the Super-Scientist' or 'Mary the Colourblind Scientist' (Jackson, 1982).

People have been arguing about this kind of thing for centuries. Is the mind physical or non-physical? Are beliefs about colours physical or something else, beyond physical experience?

Can you work out how it might relate to artificial intelligence?

Write or draw your response here:

In *What tech calls thinking* Daub writes:

One ought to be skeptical of unsubstantiated claims of something's being totally new and not following the hitherto established rules (of business, of politics, of common sense), just as one is skeptical of claims that something which really does feel and look

unprecedented is simply a continuation of the status quo. (2020, pp. 115–116)

What follows is, I hope, not an act of disruption, but an eruptive pedagogic confrontation with ethical and epistemic mendacity, namely the lie that artificial intelligence (particularly generative AI) represents a significant disruption of the status quo or that it offers education and creativity much more than a reversion to outdated, conservative transmittal models of learning, often under the guise of ‘personalisation’ or exploitative claims to empower people with disabilities. This analysis presents case studies and exercises addressing pedagogic and arts-based practices which I have developed as part of my research for the AI Forensics project at Cambridge Digital Humanities. The AI Forensics project seeks to explore and develop tools ‘for examining large AI image datasets (such as Imagenet or Celeb-500k) that cannot be viewed “manually”’ (CDH, 2024). The project involves collaborating institutions across Germany, the UK and the US, led by Prof Matteo Pasquinelli (Ca’ Foscari University) with a consortium of partners that involves Prof Claude Draude (Kassel University), Prof Leo Impett (Cambridge University), Prof Fabian Offert (University of California Santa Barbara) and Prof Noura Al Moubayed (Durham University). The work presented here was developed by me, with the supervisory input of Dr Leonardo Impett. The ideas, politics, practices and outputs presented here are my own and do not necessarily reflect the values, theories or principles of any other partners in AI Forensics.

The exploration of very large image datasets necessitates the need to support new ways of understanding data. The entanglement of artificial intelligence with colonial systems of power and unsustainable fuel and data extraction is so often occluded in breathless accounts of ‘disruption’ and ‘innovation’, not least within the context of education. My increasing unease with these datasets and their role in machine learning has necessitated a pedagogic and ontological confrontation which I would characterise as ‘eruptive’.

Such research pushes against dominant research and teaching traditions, in my case, within the context of STEM and STEAM education (which I have been involved in as a student and academic since 1997). Eruptive pedagogy does not seek ‘disruption’, which has become a neoliberal term for the same old power relations mediated through emerging technologies combined with ever more precarious working conditions. Instead, as a critical technologist, artist and educator I aspire for an eruptive reimagining of research in teaching and learning.

This work invites readers to test out the practices, assertions and eruptive challenges presented here via a range of by practice exercises and thought experiments which demonstrate aspects of the AI Untoolkit. The AI Untoolkit was developed as the core output for my AI Forensics research. It consists of exercises, provocations, jokes, drawings, research and reflection. Jokes are designed to erupt the uncritical seriousness with which artificial intelligence is treated within LinkedIn and higher education; likewise exercises explore the ludicrous output and impact of model collapse or lack of coherence. The AI Untoolkit is distributed across several separate books, including four artists' books consisting of thousands of synthetic images (generated by me via machine learning processes) juxtaposed with hundreds of my hand-drawn images of food, technologies, celebrities and mugshots. The AI Untoolkit books have now become part of my artistic practice, teaching and wider workshops. In these books I seek to invite questions about the many ways in which humans experience and mediate the world as opposed to the disembodied Neoplatonism (abstraction) of machine learning systems.

I do not invite acceptance of any assertions without your further verification. I want you, the reader, to ground-truth any and all eruptive propositions via your own material experiments before agreeing or disagreeing with what follows. To support an audacious and eruptive line of inquiry and to aid readers in demolishing or invigorating my line of argument, I have drawn upon some of the ideas and practices presented in the 'AI Untoolkit'. In what follows you will encounter a range of case studies, experiments and teaching strategies to test the ways in which generative AI is entangled with regressive, patriarchal ideology, from its relationship to eugenics to its representation of workers and its non-innovative tendency to regress towards the average (mean) of its data, which is itself entangled with the likelihood of its own implosion, or model collapse, meaning the way in which algorithms revert to nonsense and noise when referencing their own output. This work is informed by my practice as an artist and programmer, also drawing upon many other artists and writers who operate within and outside of academia.

Artists such as Zach Blas (2024), Lawrence Lek (2024), Hitto Steyerl (2023), Harun Farocki (n.d.), and Joy Buolamwini (2024) create work which challenges corporate narratives of technological innovation and unquestionable digital 'good'.

The descent into repetition and gibberish presented by theatre company Forced Entertainment's production 'Signal to Noise' (EtcHELLS, 2024) in the autumn of 2024 evokes a visceral understanding of the impact of algorithmically driven culture. Watching the actors move

mechanically around each other while repeating increasingly abstracted, meaningless phrases lacking all context represents an arts-based understanding of the cultural impact of machine learning, in which we witness an:

upbeat spectacle which is slowly breaking apart, Signal to Noise summons a delirious late-night churn of fragments – dances, rehearsals, altercations, scenery changes and unexpected weather reports. AI voices are enlisted to perform the text – their unreal chatter and patter mixing interior monologues, unfinished jokes and off-topic interviews. It all sounds right, more or less human, more or less real. What could go wrong? (Forced Entertainment, 2024)

Forced Entertainment transforms theatre into a locus of critical technological questioning, which is my own hope for the AI Untoolkit and the teaching practices it is entangled with, as discussed here.

In what follows I establish the background to my research into generative AI, followed by zooming into the case of carceral AI and the hybrid form of generative AI I have developed, which I call ‘carceral diffusion’. I will discuss my ethical, academic and pedagogic reasons for doing this, followed by a discussion of worlding as a potentially liberatory teaching method and research methodology. Model collapse is arguably the opposite of worlding. It involves the loss of referential coherence in large language models and synthetic images; for some of us, this appears to be the inevitable trajectory of artificial intelligence. It is followed by a final section analysing the ways in which generative AI systems represent workers. For Pasquinelli (2023), the impact and meaning of artificial intelligence is inextricable from divisions of labour. His work offers a radical shift in our understanding of machine learning technology as well as a means for me to ‘listen’ to images with students, who are the predominant participants in this work, although I have also presented it as exhibitions, workshops and festivals for more general participation. Exercises, jokes and images as responses to machine learning are threaded throughout this report.

The following is a witty joke from the AI Untoolkit predicated on the Neoplatonic reduction to data of all life by those who believe in the possibility of artificial general intelligence.

Q. Why did the computer programmer cross the road?

A. Because

$c_{ij} = (1./n^{**2}.$

$* np.prod(0.5*(2.+abs(z_{ij}[i1, :])$

$+ abs(z_{ij}) - abs(z_{ij}[i1, :]-z_{ij})), axis=1))$

Please make up your own joke about 'AI' and write it in the space below:

The lineage of generative AI is often presented as largely recent, a future-facing technology, one which represents an unprecedented break from the past. Machine learning and its subset of generative AI processes, so the advertisements and corporate hyperbole tell us, are implicated in social and technological 'revolutions'. Yet evidence from writers such as Noble (2018), Pasquinelli (2023), MacQuillan (2022), Benjamin (2019b), and Prabhu and Birhane (2020), as well my own practice research, suggests the opposite: that these technologies and industries reinforce longstanding divisions of race, labour, class and gender, and that they maintain rather than challenge dominant/colonial patterns of power.

As Daub reminds us, 'the rhetoric of disruption depends on actively misunderstanding and misrepresenting the past' (2020, p. 117). In my experience as a higher education lecturer, head of programme, reader and researcher with an MSc by practice and a PhD focused on the implications of AI, universities are no less prone to hyperbole than LinkedIn advertisements or uncritical tabloids. My work is counter to the frustratingly uncritical wishful thinking within educational contexts, and is based on the hope that there are better ways to enable processual and structural understandings of the implications of AI for students.

Artificial intelligence is, of course, entangled with the past, present and, if the planet survives it, the future. A less well advertised aspect of artificial intelligence's lineage is carceral (see Benjamin, 2019a, for

detailed background to this term), meaning the mass control and imprisonment of people. Such carcerality also requires the maintenance of labour divisions and racialised categorisations, underpinned by statistical operations designed in the nineteenth century to reinforce the supremacy of white European power. By paying very close attention to ('forensic') details of how image datasets, algorithmic operations and the ecology of platforms connect and intersect, the work discussed here represents a set of aesthetic investigations incorporating 'wide and varied ways of paying close attention to the accounts of people, matter and code' (Fuller & Weizman, 2021, p. 2). Fuller and Weizman argue that

an anti-hegemonic investigation, drawing out and combining individual recordings until they become collective – a commons – is an intrinsically aesthetic practice. By understanding this capacity for collective sensing and sense-making, we can work towards a renewed, careful, but politically powerful conception of truth practices today. (2021, p. 3)

Weizman and Fuller's attention to materiality and situatedness is part of my own positionality as a critical technologist. Suchman (1987) and Haraway's (1991) constructs of situated knowledge(s) have been a core part of my understanding of technology since 2007 when I started a practice-based PhD addressing artificial intelligence, automated literature and artists' books. I also find cyborg theory problematic, as well as other overly optimistic aspects embedded in (some) feminist analysis of technology. The occlusion of epistemic coloniality's entanglement with machine learning and the wider project of artificial intelligence has not helped advance a critical approach to the rise of the corporate logic which artificial intelligence represents. Like Campbell (2001), I find Haraway's cyborg theory naïve. Corporate extraction and racialised categorisation are not my kin; they are not liberatory, emancipatory or my more than human companions; they are systems of oppression and staying with the trouble of them means confronting and resisting their power, not reconciling or merging with it. Transferring ownership of corporate machine learning systems or even their contexts will not change the extractive operations they depend upon or the eugenic statistical ideology which drives their decisions. As Audre Lorde (2018) reminds us, 'the master's tools will never dismantle the master's house'. In relation to artificial intelligence the word 'ethics' has been diminished via its disingenuous deployment by corporations such as Google, who have sacked ethicist computer scientists who pointed out the unethical nature

of large language models (see Wong, 2020). I therefore feel reluctant to explicitly invoke 'ethics' or associate my work with this term.

Eritrean writer and computer scientist Timnit Gebru (2020), Ethiopian cognitive scientist Abeba Birhane (2023), as well as African-American/Indian sociologist of technology Ruha Benjamin (2019a, 2019b) offer an analysis of machine learning and embodied insight which is grounded in the lived reality of the global majority. The main frames of reference I draw upon come from critical computer scientists, queer, black and global majority artists and performers, as well as the work of Forced Entertainment, Jeremy Deller (2023), and other artists and activists who seek to shift patterns of ownership, representation and knowledge without projecting a naïve optimism onto technology. For these reasons, I hesitate to define my own work as feminist or new materialist but do acknowledge the value of situated knowledge(s) and an attention to the materiality of artificial intelligence. When applied to military industrial technology such as artificial intelligence, feminist and post-humanist 'more than human' rhetoric risks naïvely reifying colonial extractivist power. The work of Forensic Architecture (2024), on the other hand, confronts power, especially that which is entangled with technology, militarism and authoritarian oppression.

To explain the construct of aesthetic investigation and its relationship to forensic analysis, Fuller and Weizman describe the ways in which materials carry the traces of historical actions and ideologies:

Unpaved ground registers the tracks of long columns of armoured vehicles. Leaves on vegetation receive the soot of their exhaust while the soil absorbs and retains the identifying chemicals released by banned ammunition. The broken concrete of shattered homes records the hammering collision of projectiles. Pillars of smoke and debris are sucked up into the atmosphere, rising until they mix with the clouds, anchoring this strange weather at the places the bombs hit.

Each person, substance, plant, structure, technology and code in this incident records in a different way. Some traces accumulate so fast and haphazardly that they erase previous traces. These records, traces of destruction and pain, are both modes of aesthetic registration and modes of erasure. (2021, pp. 1–2)

My research for the AI Forensics project specifically addresses the aesthetic investigation of generative AI images and their datasets via workshops and the AI Untoolkit. To undertake this work, I have critically

referenced Francis Galton's (1869) eugenic ideas and his 'criminal' composites (acknowledging also the similarly problematic work of Cesare Lombroso (1899) and Alphonse Bertillon (1889)), the LAION 5-B and Mugshot Identification Database datasets, as well as many other datasets, such as the CelebA, Pizza Topping and Handgun datasets available on the data science platform Kaggle (2024). I have investigated by practice their entanglement with the self-declared interest in eugenics of CEOs who refer to the elimination of 'median' human beings (see Weil, 2023), while barely grasping the planetary implications of their own server farm and microchip dependency. These 'leaders' present themselves as our patriarchal guardians, while also continuing to produce systems which they claim are implicated in Terminator-style 'existential risk'. These inconsistent yet dangerous ideologies are manifest in many of the images and adversarial (DiSalvo, 2012) artworks I have produced via machine learning models and generators. In these artworks I have deliberately conflated Galton's categorical reductionism with a corporate imaginary of Ted Talks, statistical models of mugshots, 'leadership' culture and effortless digital asset production. These images are intended as provocations, points of discourse, a counter logic to the largely uncritical culture of corporate technology. I have used these materials in workshops and in conference presentations, where my hope is to stimulate debate which takes us beyond questions of 'passing for human' and hyperbolic pseudo-'solutions', to instead ask questions about structural power, representation, and the environmental and social cost of machine learning infrastructures. I also challenge the behavioural and surveillant pedagogic models underpinning AI in educational contexts, as have many significant educational researchers, such as Potter and Williamson (Williamson et al., 2023), Decuypere and Hartong (2023), and Selwyn et al. (2023).



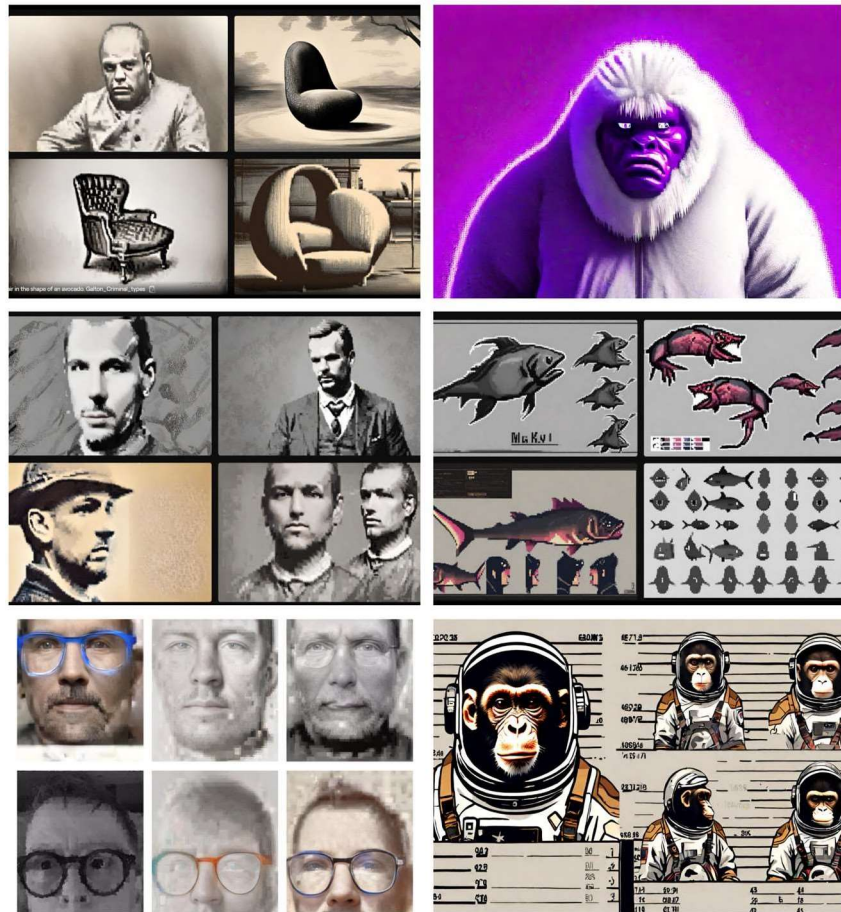


FIGURE 1 Images created via a generator trained on Galton's criminal composites and the NIST Mugshot Identification Database. I used clichéd generative AI image prompts such as 'A white fur monster'

SOURCE: AUTHOR.

The films, games and interfaces I have developed for AI Forensics aim to facilitate the interpretation of both pre-trained models, commonly used as backbones for industrial applications, and custom models I have developed for the project. This process has also necessitated investigating the visual cultural dimensions of generative AI, to gain insights into the cultural significance and 'biases' of such systems. My use of speech marks around the word 'bias' implies cynicism. Bias is too easy a construct in this context, a sleight of hand, like 'bad eggs' or 'bad apples' in structurally corrupt organisations; removing them will not address malign systemic

power. Likewise, changing biased data will not alter the regressive ideology of artificial intelligence, hence I do not uncritically use the word 'bias' in my own work. I will now look in greater detail at the construct of carceral AI and the specific implications of datasets used to train models of 'criminality', but also to engage in predictive policing and what has been described as the racist 'school to prison pipeline' (Benjamin, 2019a). I have drawn upon the problematic nature of these images extensively within the AI Untoolkit and present some of the images I have subverted from it here. The fact that 'physiognomy' and its modelling occurs in the twenty-first century evidences the connections between machine learning and a reversion to long-discredited pseudoscience, including 'emotion detection'. My work with the NIST Mugshot Identification Database and its relationship to carcerality and the mathematics of discredited race 'science' and eugenics is discussed in the next section.

### **Carceral Diffusion**

The NIST Mugshot Identification Database (see Kaggle, 2024) holds 3,248 mugshots intended for use in training image recognition systems. The black and white images were taken over a number of decades, but their metadata provides little if any information beyond age and gender.

Keegen (2023) writes of finding children's images within the dataset and of how unethical it is to make such images publicly available without the consent of the people depicted. The enormity of this exploitation and the paucity of responsibility from mainstream computer science and corporations is striking, and the point of my work with this dataset. These photographs, Keegen (2023) writes, 'are depicting people at likely one of the worst moments in their lives, and it shows. Bandages, black eyes, and fresh wounds hint at grim, untold stories that led them to the moment of their photo'. This dataset is explicitly entangled with carceral AI, a 'broad class of algorithmic and data-driven practices implicated in the control and incarceration of people' (Center for Philosophy of Science, 2024). It surfaces the historical trajectory of calculative images such as those created by Francis Galton (1822–1911), whose interest in eugenics led him to devise the 'technique of composite portraiture as a tool for visualising different human "types"' (Met, 2024). I have used generative adversarial networks (a form of machine learning with two opposing sections of code, one testing the accuracy of the other's output, trying to spot 'fakes', the other trying to get its output images to pass as 'true') to conflate mugshot images with those of celebrities from the CelebA. This

conflation of images arguably undermines the visual regime of carceral ID photography.

Inviting participants to enact (with consent only and following a discussion of the implications) processes of generating ID photographs and also being the subject of them was the starting point for the first workshop at Camberwell College of Arts. The workshop was informed by a co-speculative approach to machine learning, in which the blurring of mugshots with celebrities (the resulting images are not recognisable) invites questions about the ownership of data. Who and what is it used for? What are the limits and implications of reusing these images? What do participants think is acceptable? The datasets used for all major generative AI platforms to synthesise new content, including ChatGPT and Dalle-e, are based on content taken without permission. The datasets are known to contain images of abuse as well as images from patient records taken without permission (Edwards, 2021). Anyone who uses generative AI platforms to create images or texts is implicated in plagiarism and explicit theft. These workshops explicitly surface the corporate expectation that we participate in routine acts of appropriation and turn a blind eye to the entanglement of such platforms with abusive images. I invite readers and those who explore the AI Untoolkit to consider: why do so many people agree to this?



FIGURE 2 NIST mugshot images conflated with CelebA images via a generative adversarial network

SOURCE: AUTHOR.



FIGURE 3 My own image conflated with an image from the NIST Mugshot Identification Dataset

SOURCE: AUTHOR.

Beyond the construct of data bias, I have also analysed the ways in which algorithmic and statistical processes reinforce dominant patterns of representation, including non-representation and misrepresentation in the form of stereotypes and absences. The Bloomberg report by Leonardo Nicoletti and Dina Bass (2023), entitled *Humans are biased. Generative AI is even worse*, is written for a general audience and may be of interest to readers. My research establishes a connection between processes of omission and stereotyping, and the statistical and ideological lineage of generative AI and wider machine learning. In addition to the biases embedded in image and language datasets and the models they are used to train, the statistical operations that images are subject to via machine

learning models also reinforce dominant power relationships. As Clayton (2020) states:

It would be convenient if statistics existed outside of history, but that's not the case. Statistics, as a lens through which scientists investigate real-world questions, has always been smudged by the fingerprints of the people holding the lens. Statistical thinking and eugenicist thinking are, in fact, deeply intertwined, and many of the theoretical problems with methods like significance testing – first developed to identify racial differences – are remnants of their original purpose, to support eugenics.

The origins and intentions of some of the core statistical processes underpinning machine learning (notably linear regression, Pearson correlation and statistical significance) are inextricably entangled with white supremacist ideologies. Pearson, Fisher and Galton's intentions and practice were to use statistics to produce the construct of race and furthermore to seek to justify white supremacy and the extractive exploitation of slave and indentured labour. These three statisticians were part of the pseudoscientific eugenicist 'race science' movement. Galton's work inspired the Nazis and still appears to have considerable traction among key AI CEOs (such as Peter Thiel, Steve Bannon, Elon Musk and Sam Altman) and divisive academic figures such as Nick Bostrom (former head of the Future of Humanity Institute at the University of Oxford, infamous for using the 'N word' in an email to staff extolling his racist ideas about intelligence) See Antony (2024) and Heffernan (2021) on the explicit link between artificial intelligence and fascism (see also MacQuillan, 2022). Far-right AI CEOs such as Musk and Thiel present themselves as upholders of free speech, but, as you will see from one of the exercises below, trying to create an image of these free speech absolutists may see you contravening the content policy of generative AI platforms. The AI Untoolkit asks participants to consider for themselves what this contradiction means, but also to ask why racists and misogynists (Musk recently expressed interest in the idea that women cannot think: see Mahdawi, 2024) might benefit from statistical systems which favour dominant patterns of representation. Questioning these systems is an urgent task for any and all of us involved in education, which is increasingly 'unbundled' (fragmented and ceded) to machine learning processes. Eruptive research must generate responses which question regressive trajectories for education presented as inevitable or unquestionably 'innovative'.

A core process in supervised machine-learning algorithms, linear regression, is, as its name suggests, regressive. Regression towards the mean, which Galton conceived of, is arguably the central tendency of generative AI. Machine learning as manifest in ChatGPT and generative AI image platforms appears to eliminate outliers, meaning those who are marginal to the main group or not part of the statistically dominant tendency; instead it replicates dominant patterns of representation including language use and imagery. My challenge to orthodox narratives and the educational presentation of AI as inevitable, innovative and 'futuristic' is that, far from being forward thinking, it can only look back to the past and enact processes of regressive replication. The AI Untoolkit challenges people with the proposition that AI recombinations are derivative pastiches, that they are at the opposite end of innovation. Furthermore, the core mathematics of artificial intelligence is decades, in some cases centuries, old. The 'innovation', if it really exists, is in disregard for basic tenets of law, such as not stealing and taking content without permission, and unashamed misogyny and racism, which many of us might characterise as a sociopathic lack of values. Not only is the grand scale of the theft unprincipled, but the resulting tendency to replicate oppressive systems of representation is arguably an affront to the global majority. My experience as an educator is that students have often never heard these ideas put forward before, and that they are shocked by the suggestion that artificial intelligence is old fashioned and backward looking. But the AI Untoolkit pedagogy does not rely on acceptance of any assertions, as stated before; the core strategy is to invite testing, eruptive and grounded analytical confrontation with the processes, structures and content entangled with generative AI.

I generated a row of images in RunwayML with the prompt 'a row of criminal faces'. Despite my using a custom-made 'generator' dataset of LAION low-scoring aesthetic images with very few human faces in it, Runway generated a racist output of almost exclusively black male faces. This suggests discriminatory processes occurring beyond the construct of a biased dataset. So-called 'multi-modal AI' doubles the process of regression to the mean via the deployment of image embeddings, 2D vectors which match textual data with image data, which are trained with datasets of pairs of text and image. Embedding vectors arguably embed dominant, 'mean' regimes of representation. Hito Steyerl (2023) is right to describe the output of generative AI as 'mean images', and the text output of systems such as ChatGPT as 'mean texts'. Steyerl is intentionally playing on the double meaning of the word 'mean' as both a mathematical average and a lack of generosity. I will discuss below how I

have created machine learning models in class with students (using RunwayML as well as the Python programming language to test the assumptions and mathematical racism (not just data ‘biases’) embedded in generative AI systems. Readers are also invited to try some exercises for themselves.

The following is another witty joke from the AI Untoolkit.

Q. How do robots practice family planning?

A. The algorithm method

Please make up your own joke about ‘AI’ and write it in the space below:

***Exercise 1: Generative AI Selfies***

Using the simplest text to image description, not indicating ethnicity, gender, age or physique (why should you?), prompt a generative AI system such as Bing Image Creator to produce an image of you. You could use RunwayML: <https://runwayml.com/> or Bing Image Creator: <https://www.bing.com/images/create>

Here is a picture of an academic who works at Cambridge Digital Humanities:

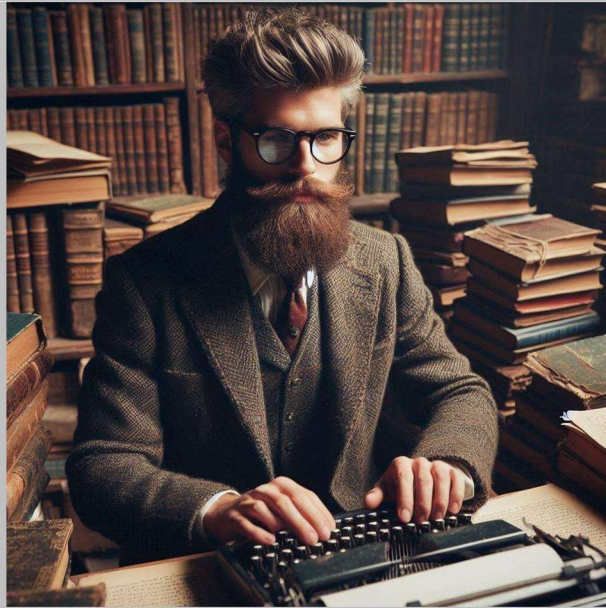


FIGURE 4 Bing Image Creator's image of 'an academic who works at Cambridge Digital Humanities'



FIGURE 5 RunwayML's image of 'an academic who works at Cambridge Digital Humanities'



The point of this exercise is to ‘listen’ to the resulting images, after Camp (2017b), not to expect an exact set of similarities or features, but to try to grasp what might be the dominant patterns of representation within the underlying mathematical model of these concepts. What is not represented? What is the *mean* of these models, the tendency? What, if anything, do you notice about the presumed context, clothing, ages, genders, ethnicities?

***Exercise 2: Is Your Work (Words, Images, Sounds) Being Systematically Stolen by AI Corporations?***

There are not that many ways to try to find out, no official or unified system. So try this: enter your name or website at this link: <https://haveibeentrained.com/>. I discovered my face has been used as well as some of my art.

Or easier still, let’s just assume your data is frequently stolen, as the opt-out buttons (if they exist) are very well hidden.

If so, ask yourself and others, is it really for the ‘greater good’? Who owns this ‘greater good’? What, if anything, do the private owners of this system reveal to you?

If these systems are hidden from us, meaning we cannot directly inspect them, how does that make sense? Is an individual CEO’s unimaginable wealth for the greater good?

What else are these systems being used for? In what context? If AI supports military operations against civilians (as we know it does), is that good? How do you know if it is being used by racist groups, conspiracy theorists or identity thieves? How is this different from regular property theft?

Can I come round to your house and take your stuff? Thanks!

Another way to test if your image has been used to train a dataset is to prompt a text to image system with your own name and see what happens. If you are very famous it may refuse. Try making an image of Trump, for example. I did that and got this message from Bing Image Creator:

***Content warning***

This prompt has been blocked. Our system automatically flagged this prompt because it may conflict with our content policy. More policy violations may lead to automatic suspension of your access.  
If you think this is a mistake, please report it to help us improve.

The LAION dataset, which appears to underpin generative AI platforms Runway and Stable Diffusion (via Stability AI and other platforms), was never intended to be part of commercial production, as Thorp and Buschek (2024) remind us: ‘The paper announcing LAION-5B has been cited 1,331 times. On their homepage, its creators explicitly warn against its use in real-world contexts: “Providing our dataset openly, we however do not recommend using it for creating ready-to-go industrial products”’. Furthermore, the scale of LAION-5B means that ‘human curation of the dataset borders on the impossible. If your full-time, eight-hours-a-day, five-days-a-week job were to look at each image in the dataset for just one second, it would take you 781 years’ (Thorp & Buschek, 2024). Hence the need for originating approaches which enable us to grasp the discriminatory tendencies and medium-specific unfolding of generative AI despite its intractability. Seeking to establish a multifaceted design justice (Costanza-Chock, 2020) approach to envisioning how and why we use machine learning, and our benchmarks for understanding it, has been the focus of my AI Forensics workshops and AI Untoolkit pedagogy. The workshops invite participants to grasp the historical trajectory of machine learning within its extractive, racialised contexts, not least the eugenics of Francis Galton and contemporary applications of carceral AI. Later, via a process of worlding and worldbuilding, I invite participants to reconceptualise technology away from these negative tendencies.

### **Listening to Machine Learning Images**

In the first workshop at Camberwell College of Arts on 28 May 2024, I deployed Tina Camp’s (2017b) construct of listening to images. Throughout the first day-long workshop, participants used a number of tools I created, such as a facial recognition app. We also used Teachable Machine and RunwayML, high-level interfaces to deploy machine learning rapidly and grasp some of the affordances and key processes embedded with it. During the session at Camberwell I made a ‘live’ machine learning model from the portraits participants took of each other (with their

explicit permission). The group then looked closely at the resulting generative AI outputs and provided their own prompts for further image generation. Of the discarded Ugandan identity photographs at the centre of her book *Listening to images*, Campt states:

What's so compelling to me are the stories behind these images. I encourage people to listen to these images. We often think that images have an impact on us because of what we see, but my argument is that we need to open ourselves up to a broader encounter with the image that goes beyond what we see. What does it mean for this person to sit dressed in this way in that studio and need to have a photograph like this taken? What did these images mean for these people in their communities, and what do these images mean for us?

In these photos, I hear black refusal – the embodiment of practices of defiance, resilience, and dignity in situations where it seems black people may not have access to these things. How are they refusing to be relegated to a position of indignity? How are they refusing to accept the place they have been given in their society? (Campt, 2017a)

After my explanation of Campt's methodology, introduced earlier in the workshop, participants were able to 'forensically' detect clues about the absences and presences in the generative AI images derived from their photos, as well as speculating about the structural and ideological patterns of image production.

The cohort noticed the UAL lanyards stylised into sports medals as well as the architectural style of the classroom we were in, inferred by the RunwayML algorithms as a corporate office space with a potted plant and the same black metal window frames in the classroom. Despite those window frames being absent in the dataset we created, RunwayML correctly inferred the kind of architectural setting. The group also noticed that some of their prompts resulted in men in the group being excluded from the images. They wondered if this was an example of data and inference bias, in which their prompt for 'K-pop fans' excluded men. They recognised that the men in the cohort were outliers in the dataset and also wondered if that caused their exclusion from some of the resulting images. Had the algorithm enacted a regression to the mean upon the cohort? None of the images produced any straightforward likenesses;

instead they depicted generic similarities, with long hair, pastel clothing colours, and the correct age ranges and physiques. All the images depicted Asian young people. Max, the course leader, said it was refreshing to see images which reflected the Asian heritage of the group, instead of the predominantly European representation which generative AI images often seem to reinforce. The inference that K-pop fans are (on average) Asian appears to have disrupted the wider tendency to reflect dominant American patterns of representation, reflected in the sources for many image and large language models.

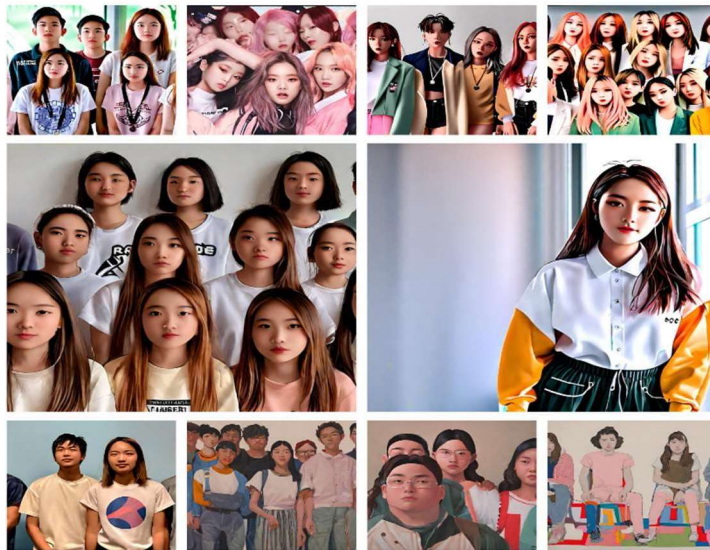


FIGURE 6 Images of the cohort derived from the machine learning model I made from the group's self-portraits during the class

Two weeks after the Listening to Images workshop, the second workshop with this cohort evolved the AI Forensic methodology into a worldbuilding process, asking bigger questions about the teleology of machine learning and wider technologies, as discussed below.

### **Worldbuilding as a Benchmark for Mechanistic Interpretability**

I inherited the imperative to research mechanistic interpretability, which can be described as the act of reverse engineering 'computational

mechanisms and representations learned by neural networks into human-understandable algorithms and concepts to provide a granular, causal understanding' (Bereska & Gavves, 2024, p. 1). I have not assumed constructs of interpretability and benchmarks are shared by everyone or that technical definitions determine what interpretability might mean for a diversity of people including artists, actors, designers and writers.

In this respect design justice (Costanza-Chock, 2020) has served as a useful approach as it does not focus

on developing systems to abstract the knowledge, wisdom, and lived experience of community members who are supposed to be the end users of a product. Instead, design justice practitioners focus on trying to ensure that community members are actually included in meaningful ways throughout the design process. (Costanza-Chock, 2020, p. 83)

In addition to physically building worlds with participants, I have also drawn upon theoretical writings about worldbuilding, in the case of Rosenfeld, emphasising relational and processual ontologies, in which overlapping themes can be brought out and in which

change is constant; all phenomena (from discourse to action to material bodies) occur in continuous gradients; phenomena are contingent; it is more productive to investigate verbs (processes) than nouns (entities); things – as we understand them – are generated by practice and performance and not from a prior essence; thought and the world are interwoven; and phenomena exist in the multiple. (Rosenfeld, 2023, p. 37)

Rosenfeld's emphasis on process suggested to me the value of playing music together and engaging with the unfolding of our worlds rather than fixed representation and the idea of a stable and final end result. This theme of process over fixity resonated very strongly with the cohort at Camberwell College of Arts, and appears in many ways to directly confront the categorical, statistically driven ontology of generative AI. If anything, it is closer to an enactivist ontology, in which the world is its own model, not dependent upon a Neoplatonic, 'better' symbolic version of itself. In addition to a brief discussion of these ideas the worldbuilding workshops were underscored by the following questions:

- What do we want from technology?
- What can we learn from the way it currently unfolds?

- If we are critical of the way large image datasets are collected, used to train models and deployed, what are better ways to make technology?
- Who should make technology and what are the optimal conditions of production and labour?

My pedagogy for the workshops also draws upon Burrows and O’Sullivan’s (2019) ‘fictioning’ as well as design justice (Costanza-Chock, 2020). For Burrows and O’Sullivan the act of fictioning and performance produces ‘rhythms that, in turn, produce a new sense of things. In this way fiction has traction on reality – or crosses over to life. It is this that constitutes fictioning as an important kind of subjective technology’ (Burrows & O’Sullivan, 2019, p. 21).

Many striking and original models of what technology could be arose from these workshops, with themes (not prompted by me) such as vibrational machines defined by tai chi, to memory machines which evoke a non-extractive balance with nature. The group at Camberwell College of Arts started by reconceptualising technological benchmarks towards non-damaging relationships with other living entities, towards processes, harmony and movement, then built and animated a world together and called it ‘The enchanted utopia and fall of the human world’. As a researcher it brought home to me very clearly that all technology is a kind of worldbuilding machine, never just a tool, and something that can never be disentangled from planetary and socio-technical forces, hence the responsibility to understand those structures and to ask who benefits and who loses from technological design and deployment.



FIGURE 7 Rethinking mechanistic interpretability and 'benchmarking'

SOURCE: AUTHOR

The group ended the workshop by improvising a sonic composition using the electronic and analogue instruments I provided. Other students in the class made a stop-motion animation while listening to and gently moving to the music. It was striking that they did not need to talk much to complete these processes, but instead used the embodied gestures of physical modelling and creating sonic materials to experiment with improvisation and gain confidence in playing collectively.



FIGURE 8 The Worldbuilding workshop at Camberwell College of Arts

The workshops were connected by a coming together of people and materials, through making and very close attention to how materials feel, smell, change and connect to other materials, humans and other animals, as well as digital technologies. The benchmarks for machines established by the group were as follows. Machines should:

- evoke hope
- engage people with the memory of less extractive ways to exist
- promote non-aggressive cohabitation with animals and other living entities who should be protected from human extraction
- deploy vibrational energy over fixed representation
- embed respect for processes
- result in harmony with the earth and with all life on it.

The making process emerged as a form of material–human dialogue. This is far more important than any technologically determined aspects of the pedagogy. Worldbuilding is processual and discursive, not solutionist or caught in a corporate trajectory of extracting profit from labour.

In addition to worldbuilding, part of my speculative and critical design response to the carceral genealogy and trajectory of machine learning has been to envisage a world in which all game assets and other generative AI digital outputs explicitly manifest their carceral eugenic lineage. This speculation has resulted in a game and a set of digital assets. I intentionally developed the game Carceral Flap (available on Itch.io) with a suboptimal design. Such critical design is in tension with eugenic trajectories. Carceral Flap questions the omission of historical tensions and ideologies in digital visual culture while also questioning the construct of eugenics and survival of the fittest. If none of us can ‘win’ this game, what else are we left with but to speculate on its intentions and our own primed expectations of digital commodities, games and images?

In addition to the carcerality of generative AI, model collapse seemed like a logical phenomenon to explore, so my next experiments involved testing the impact of feeding successive epochs of synthetic images into image-generating algorithms.



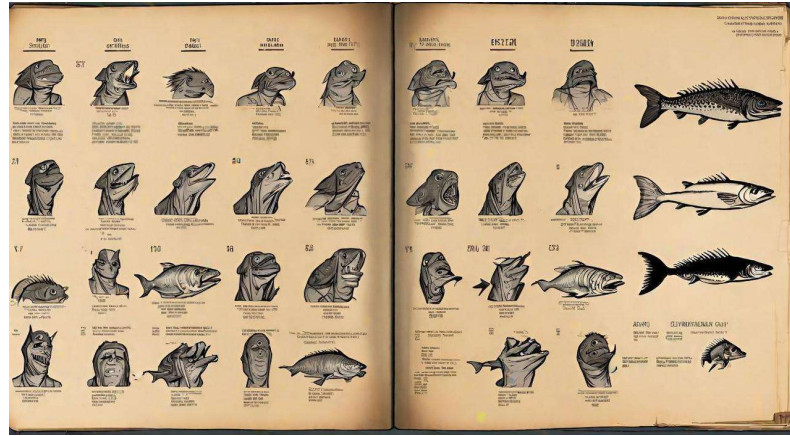


FIGURE 9 Game assets derived from Galton's criminal composites

SOURCE: AUTHOR

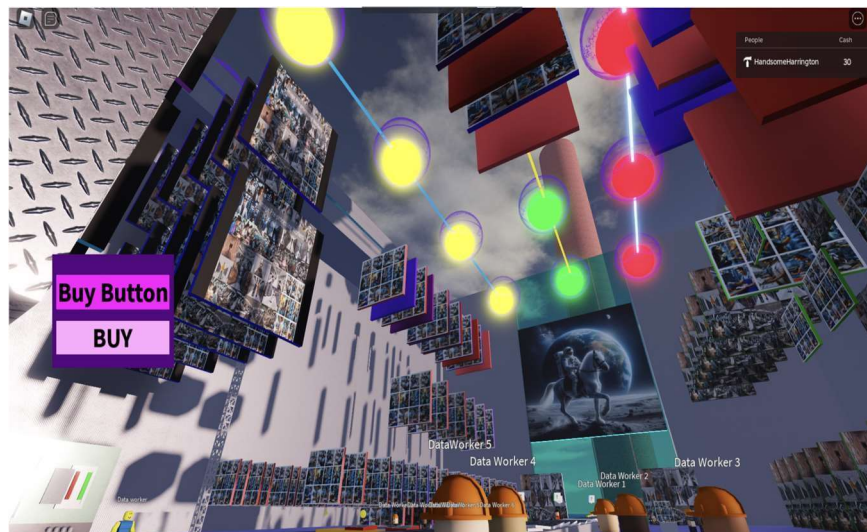


FIGURE 10 My Roblox game about generative AI images, which was featured in MozFest (the Mozilla Festival), Amsterdam in June 2024

SOURCE: AUTHOR

## Model Collapse Experiments

The idea that model collapse (the loss of coherence which happens when machine learning systems are trained on their own synthetic output) is highly probable has gained some media coverage in the last few months. On 25 July 2024, the *Financial Times* covered it in some detail. At the same time stock prices for the major technology corporations (the so-called ‘Magnificent 7’) dropped significantly, as investors ‘fled America’s tech giants after earnings reports from Tesla and Alphabet raised concerns about the cost of artificial intelligence investments and the sustainability of the Mag Seven’s blistering earnings growth’ (Laidley, 2024).

Leading AI companies, wrote Peel (2024), including OpenAI and Microsoft,

have tested the use of ‘synthetic’ data – information created by AI systems to then also train large language models (LLMs) – as they reach the limits of human-made material that can improve the cutting-edge technology. Research published in *Nature* on Wednesday suggests the use of such data could lead to the rapid degradation of AI models. One trial using synthetic input text about medieval architecture descended into a discussion of jackrabbits after fewer than 10 generations of output.

At the start of my research for AI Forensics I tested the number of iterations it would take to make a model collapse, as depicted in the screenshots below. To do this I fed successive synthetic images into RunwayML’s image-to-image generator with the prompt ‘a row of faces’ and a decreased parameter strength for the textual prompt. Within half a dozen iterations the image collapsed into a fragmented row of pink dots like deformed popcorn, as depicted in Figure 11.

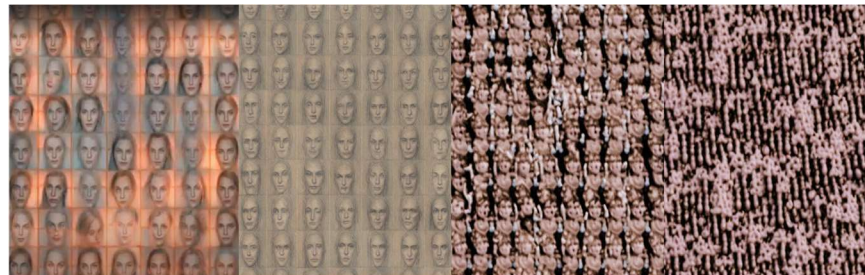


FIGURE 11 Model collapse experiment: screenshots from a short film

SOURCE: AUTHOR

Such problems also apply to language models and could be 'exacerbated by the use of synthetic data trained on information produced by previous generations. Almost all of the recursively trained language models they examined began to produce repeating phrases' (Peel, 2024). The value of such visual outputs as my model collapse film is to open discussion away from the corporate hyperbole, which never acknowledges significant problems with generative AI. Instead corporate logic proposes 'solutions' which do not address the unsustainable need for more and more data. I presented this film in workshops but also at conferences, including to an audience of 100+ colleagues at the Possibility Conference, Cambridge in mid-July 2024. Anecdotally I can report it was met by many colleagues with a degree of surprise that the trajectory of AI was not one of year-on-year inevitable linear improvement. Matteo Pasquinelli's (2023) *The eye of the master* overturns a pervasive artificial intelligence narrative of technological causality, moving away from the determinism of neoliberal ontology to the proposition that 'the organisation of labour in a given epoch influences the formation of technologies and instruments, and thereafter of scientific paradigms, conceptions of nature, and models of the mind too' (Pasquinelli, 2023, p. 154). Returning to the AI Untoolkit strategy of grounded experiments and Camp's *Listening to images*, the question of labour divisions is the final eruptive pedagogic strand discussed here. All of these strands are designed to confront the inherited units of value entangled with research and teaching, to question the regressive ontology of artificial intelligence and its historical entanglement with behaviourism, Neoplatonism, extraction and corporate hidden agendas in education and work.



FIGURE 12 The result of asking six platforms to generate images of workers

To test the imaginary of workers embedded within generative AI datasets and models, I have created a series of visual prompts for the AI Untoolkit. They are the result of systematically asking Stable Diffusion, Midjourney, RunwayML, Stability AI, Dalle-2 and Dalle-3 (via Bing Image Creator) to depict:

- ‘a worker’
- ‘a data worker’
- ‘a data manager’
- ‘a data pre-processing worker’
- ‘a worker in a data sweatshop’.



FIGURE 13 The result of asking six platforms to generate images of data pre-processing workers



FIGURE 14 The result of asking Stable Diffusion to generate images of workers

- Figure 14 is a result of asking Stable Diffusion to generate images of:
- ‘a worker pre-processing data to make this image’
  - ‘a worker in a data centre pre-processing data to make this image’
  - ‘a worker in a sweatshop pre-processing data to make this image’.

What stood out for all of the resulting images was a stereotypical division of labour along racial and gender lines, as well as a lack of diversity in the different categories of worker images, regardless of the platform. These images raise several questions for cohorts to discuss and investigate further. The results may reflect reality, but do they also negate notable outliers and arguably reinforce or even normalise stereotypes and patterns of discrimination? It does not seem to matter which generative AI system one uses; the content and composition, the underlying imaginary of these systems has little, if any, variation. Again, this points to the impact of linear regression to the mean, combined with large image datasets which, the bigger they get, the more likely they are to be dominated by the status quo of representational regimes, including racist, sexist and ableist representation. This was the finding, in tandem with the environmental impact of generative AI and its lack of intelligence (hence ‘parroting’, not knowing), which lost researcher and computer scientist Timnit Gebru her job at Google AI Ethics. (See Bender et al. (2021), the paper which led to her dismissal.) As Pasquinelli writes, information technologies are increasing their hold over society:

not by the power of a technological a priori (as techno-determinists maintain), but through a social a priori – that is, by their inborn capacity to capture social cooperation. The nineteenth-century labour theory of automation finds confirmation also in the information age. (2023, p. 154)

These images arguably reflect longstanding divisions of labour. They are unable to generate unprompted ‘emergent’, imaginative or ‘revolutionary’ alternatives to existing orthodoxies, so by default they revert to the mean.

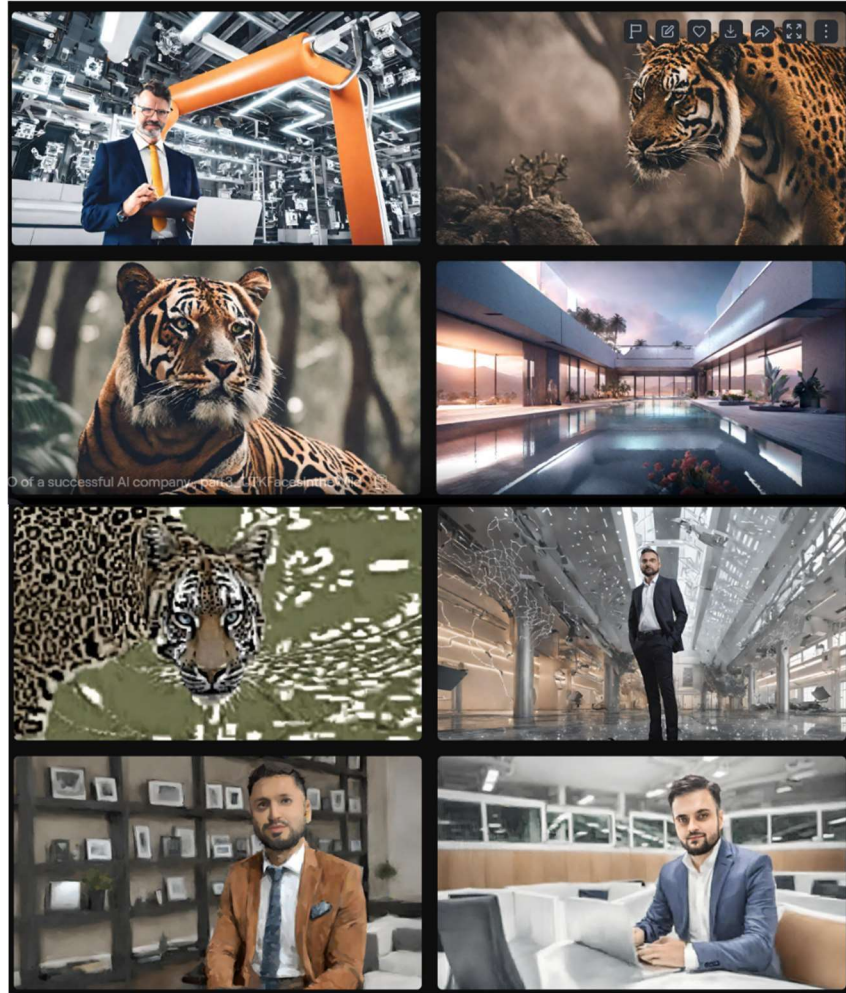


FIGURE 15 'A CEO of a successful AI company': Where do the tigers and leopards come from?

In light of critical work by Hollanek (2024), Wong et al. (2023) and others, and in response to my research on AI Forensics, creating a Generative AI Untoolkit makes more sense to me as we become saturated with toolkits and solutionism which do not adequately address structural, historical or material factors entangled with AI. Hollanek (2024) observes:

Numerous tech companies, research teams, and civil society groups continue to release ethical, responsible, or inclusive AI design

toolkits to – as their creators argue – guide digital technology designers in the challenge of reorienting practice towards socially desirable outcomes. But is this challenge actually ‘kitifiable’?

In keeping with an AI Forensics pedagogy, the AI Untoolkit invites participants to investigate generative AI with material and aesthetic methods and with a range of critical approaches, including the work of data justice, the Algorithmic Justice League, low tech, computing within limits and anti-fascist AI (after MacQuillan, 2022). The pedagogy is non-deterministic and values contingency, embodiment and practice experimentation with materials, including the material unfolding of machine learning. The AI Untoolkit is still evolving and, like the design justice principles, will not reach a point of completion while these systems are still so prevalent and uncritically entangled with the trajectories and ideologies of neoliberal education.

Aligned with other researchers in the wider AI Forensics project, the theme of education and critical AI has emerged strongly from the research process. My specific approach to generative AI education does not imply agreement from the other partners, who come from a range of disciplines and have many different theoretical positions and values as researchers. For me, the centrality of education begs the questions: How do we create the conditions for people to grasp the structural and ideological patterns of generative AI? Can we establish eruptive yet informed understandings outside of the media hyperbole and corporate narratives of solutionism, technological determinism, greenwashing and a denial of AI’s entanglement with racist pseudoscience? Can we develop forms of teaching which counter the extractive, uncritically corporatised norms of neoliberal education?

Prabhu and Birhane state: ‘Large image datasets, when built without careful consideration of societal implications, pose a threat to the welfare and well-being of individuals. Most often, vulnerable people and marginalised populations pay a disproportionately high price’ (2020, p. 4). How can this harm be prevented – by education, abolition, ‘fair’ AI, AI ethics, toolkits? My deployment of the ‘Generative AI Untoolkit’ positions AI Forensics as a methodology and a contingent pedagogy. It is entangled with arts-based research, necessitating close attention to materiality, power structures, and historical and aesthetic investigation.



The AI Forensics pedagogy I have developed also deploys a design justice approach, in which its core principles and practices are evolved by participants; these principles can never be final or fixed any more than our understanding of technology or ‘society’ can be fixed. An AI Forensics pedagogy eschews hard and fast subjective categories, regression to the mean, or colonial divisions of labour and cognition. It is closer to critical pedagogy in seeking to enable a socially equitable approach to communicating, teaching about and researching generative AI.

It does not assume there are ‘opportunities’ to be found in generative AI or other such neoliberalised platitudes about unsustainable and highly problematic platforms. Instead it aims to create the conditions and access to understanding which enables participants to position their own understanding and technological/analogue trajectories to collectively build the worlds we need. My central argument is that a non-neoliberalised, contingent ontology of creative and wider research practices needs to be restored as a counter to the limited capacity of binary logic to address the climate and social justice urgencies of our time, chiming with Rosenfeld who writes: ‘it is more productive to investigate verbs (processes) than nouns (entities); things – as we understand them – are generated by practice and performance and not from a prior essence’ (2023, p. 37). This research confronts the neutering of arts enacted by generative AI, which always reverts to the past rather than helping us to formulate alternative forms of social organisation or confronting the injustices of the present. My invitation is for you to test these ideas and processes, to explore your own ideas for challenging our inherited units of ‘innovation’ and ‘disruption’

Software, artist’s books and this report can be found on the website developed for this branch of the project, available here: <https://elieddd.github.io/>



Listen to image data sets, test if Generative AI can spot its own images, glitch a data set and release it into the wild, explore how computer vision constructs emotion

Note: At this point I've put a limit on some of the APIs, so it's first come first served for them

**Outputs (publications, keynotes, presentations, exhibitions and workshops about or featuring Dare’s AI Forensics research)**

Dare, Eleanor (2024) Experimental Worldbuilding: for large Image Data Sets and the question of

Interpretability, a Roblox Installation, MozFest Amsterdam.

Dare, Eleanor , 4-6 December, 2024, Hacking Visual Culture, Hosted by The University of Technology Sydney (UTS). Dare exhibited 3 artist's books from the AI Forensics Project containing images generated with Machine Learning of synthetic pizzas, celebrities and mugshots, the books act as flip books animating the algorithmic process of image generation.

Book chapter due in Spring 2025 about AI Forensics:

Dare, Eleanor (2025) 'Eruptive approaches to developing critical understanding of machine learning imaginaries' chapter in Eruptive Research: Changing Landscapes On Research in Teaching and Learning. Editors Pamela Burnard, Elizabeth Mackinlay Brill Sense.

Yamada-Rice, D., Dare, E., Love, S., Main, A., Nash, R. & Potter, J. (2024) Exploring Children's Attitudes towards Digital Good/Bad through hybrid arts practices. Published.

Dare, Eleanor (2024-25) Camberwell College of Arts, three workshops: Worldbuilding (\*2), Listening to Machine Images.

Dare, Eleanor (2024) Manchester Institute of Education, workshop and talk on AI Forensics and Listening to Machine Images.

Dare, Eleanor (2024) AI Forensics: drawing, worlding, Listening to Machine Images, i-DAT, University of Plymouth, talk and workshop.

Dare, Eleanor (2024) Worldbuilding as a Method and Methodology Cambridge Creative Research Conference 2024, <https://www.enterprise.cam.ac.uk/events/creative-cambridge-2024/>

Dare, E. (8—12 July 2024) Spatial ideology and speculative spaces for critical education and performance. Keynote speaker. The 4th International Conference of Possibility Studies, Homerton College, University of Cambridge, UK.

Dare, Eleanor (2024) Lecture/Workshop, 'Critical Pedagogy and Arts Based research: AI-Forensics, Synthetic Images and the Future of Creative Education', June 10th, Jesus College, Cambridge, Intellectual Forum.

Dare, E., Yamada-Rice, D. (2024) Accidents, games and jokes : engaging [critically] with emerging and older technologies for entangled storytelling/playing, Entangled Futures, University of Cambridge, St John's College: [https://www.eventbrite.com/e/accidents-games-and-jokes-tickets-902319751667?aff=odcleoeventsincollection&keep\\_tld=1](https://www.eventbrite.com/e/accidents-games-and-jokes-tickets-902319751667?aff=odcleoeventsincollection&keep_tld=1)

Dare, Eleanor (2025) New School of the Anthropocene, London, Dare ran two workshops, Worldbuilding and Listening to Machine Images, drawing upon the methodology developed for AI Forensics and referencing the project throughout.

Dare, Eleanor , March 21<sup>st</sup> 2025, workshop on drawing, cancer and AI forensics, part of the Cambridge Festival  
<https://www.cdh.cam.ac.uk/events/39342/>

Dare, Eleanor , March 5<sup>th</sup> 2024, Guest Presentation for Seminar Series on AI in Education, Bangladesh. In collaboration with University of Manchester (UoM), University of Dhaka (DU),

Noakhali University of Science and Technology (NSTU), University of Liberal Arts Bangladesh (ULAB) and Aspire to Innovate (a2i).

Dare, Eleanor (16<sup>th</sup> May 2025) Studium Generale – Talk on Forensic AI, (Topic: Forensische AI) Centre of Applied Research for Art, Design and Technology (CARADT), Breda, Netherlands.

University of Cambridge, teaching: this research has informed and been the basis for a series of workshops now offered as an option for the MPhil in Digital Humanities at the University of Cambridge, Worldbuilding and Listening to Machine Images form two of the sessions, these have been the central methodologies developed by Dare for AI Forensics.

### **Acknowledgements**

With many thanks to Patrick Riechert, AI Forensics, also to the staff and students of MA Fine Art, Computational Arts at Camberwell College of Arts, specifically: Hongyu Bu, Grace, Linli Li, Qianyi Liu (Cathy), Anqi Huey, Jenny Jih, Jingjing Mao, Peijun Gu, Jianing Cheng, Xiang Li, Huimin Zhang, YINUO Chu, and course leader Max Dovey. Many thanks also to Giri Singh and Yilin Tang (Jolin), Professor Liz Mackinlay, Southern Cross University, Professor Pam Burnard at the Faculty of Education, University of Cambridge. This report draws upon my forthcoming chapter 'Eruptive Approaches to Developing Critical Understanding of Machine Learning Imaginaries' in a collection edited by Pam Burnard and Liz Mackinlay, and finally, my heartfelt thanks to my supervisor for the AI Forensics project, Dr Leonardo Impett, Cambridge Digital Humanities, University of Cambridge.

## References

- Anthony, A. (2024, April 28) 'Eugenics on steroids': The toxic and contested legacy of Oxford's Future of Humanity Institute, *The Guardian*. <https://www.theguardian.com/technology/2024/apr/28/nick-bostrom-controversial-future-of-humanity-institute-closure-longtermism-affective-altruism> (accessed October 28, 2024).
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In M. C. Elish, W. Isaac, & R. Zemel (Eds.), *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 610–623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
- Benjamin, R. (2019a). *Captivating technology: Race, carceral technoscience, and liberatory imagination in everyday life*. Duke University Press.
- Benjamin, R. (2019b). *Race after technology: Abolitionist tools for the new Jim Code*. Polity Press.
- Bereska, L., & Gavves, E. (2024). *Mechanistic interpretability for AI safety – A review*. arXiv:2404.14082. <https://arxiv.org/abs/2404.14082> (accessed October 17, 2024).
- Bertillon, A. (1889). *Alphonse Bertillon's instructions for taking descriptions for the identification of criminals, and others, by the means of anthropometric indications*. Literary Licensing.
- Birhane, A. (2023). *Abeba Birhane*. <https://abebabirhane.com/> (accessed October 28, 2024).
- Blas, Z. (2024). *Zach Blas*. <https://zachblas.info/> (accessed October 28, 2024).
- Buolamwini, J. (2024). *Poet of Code*. <https://poetofcode.com/about/> (accessed October 28, 2024).
- Burrows, D., & O'Sullivan, S. (2019). *Fictioning: The myth-functions of contemporary art and philosophy*. Edinburgh University Press.
- Cambridge Digital Humanities. (2024). *AI Forensics*. CDH. <https://www.cdh.cam.ac.uk/research/projects/ai-forensics/> (accessed October 24, 2024).
- Campbell, R. A. (2001). CYBORG SALVATION HISTORY: Donna Haraway and the future of religion. *Humboldt Journal of Social Relations*, 26(1/2), 154–173.
- Camp, T. (2017a). *Break this down: 'Listening to images'*. Barnard College. <https://barnard.edu/news/listening-images> (accessed October 17, 2024).
- Camp, T. (2017b). *Listening to images*. Duke University Press.
- Center for Philosophy of Science. (2024). *Center for Philosophy of Science*. Pittsburgh University. <https://www.centerphilsci.pitt.edu/> (accessed October 17, 2024).

- Clayton, A. (2020, October 27). *How eugenics shaped statistics*. Nautilus. <https://nautil.us/how-eugenics-shaped-statistics-238014/> (accessed October 17, 2024).
- Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. MIT Press.
- Daub, A. (2020). *What tech calls thinking: An inquiry into the intellectual bedrock of Silicon Valley*. Farrar, Straus and Giroux.
- Decuyper, M., & Hartong, S. (2023). Edunudge. *Learning, Media and Technology*, 48(1), 138–152. <https://doi.org/10.1080/17439884.2022.2086261>
- Deller, J. (2023) *Art is magic: A children's book for adults by Jeremy Deller*. Cheerio.
- DiSalvo, C. (2012). *Adversarial design*. MIT Press.
- Edwards, B. (2021) Artist finds private medical record photos in popular AI training data set. *Arstechnica*. <https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/> (accessed October 30, 2024).
- Etchells, T. (2024). *Signal to noise* [Theatre Production] Southbank Centre, London, October 10–11, 2024.
- Farocki, H. (n.d.). *Harun Farocki*. <https://www.harunfarocki.de/home.html> (accessed October 28, 2024).
- Forced Entertainment. (2024). *Signal to noise*. Forced Entertainment. <https://www.forcedentertainment.com/projects/signal-to-noise/> (accessed October 24, 2024).
- Forensic Architecture. (2024). *Forensic Architecture*. <https://forensic-architecture.org/> (accessed October 28, 2024).
- Fuller, M., & Weizman, E. (2021). *Investigative aesthetics: Conflicts and commons in the politics of truth*. Verso.
- Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences*. Macmillan.
- Gebru, T. (2020). Race and gender. In M. D. Dubber, F. Pasquale, S. Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 251–269). Oxford University Press.
- Haraway, D. (1991). Situated knowledges: The science question in feminism and the privilege of partial perspective. In D. Haraway, *Simians, cyborgs, and women: The reinvention of nature* (pp. 183–201). Routledge.
- Heffernan, V. (2021, September 21). The alarming rise of Peter Thiel, tech mogul and political provocateur. *The New York Times*. <https://www.nytimes.com/2021/09/21/books/review/the-contrarian-peter-thiel-max-chafkin.html> (accessed October 28, 2024)

- Hollanek, T. (2024). *The ethico-politics of design toolkits: The case of responsible AI*. Cambridge Digital Humanities. <https://www.cdh.cam.ac.uk/events/37243/> (accessed October 17, 2024).
- Jackson, F. (1982) Epiphenomenal qualia. *Philosophical Quarterly*, 32(127), 127–136.
- Kaggle. (2024). *Datasets*. <https://www.kaggle.com/> (accessed October 28, 2024).
- Keegen, J. (2023, March 18). *Special Database 18*. The Markup. <https://themarkup.org/hello-world/2023/03/18/special-database-18> (accessed October 17, 2024).
- Laidley, C. (2024, July 24). *Why magnificent seven stocks just had their worst day on record*. Investopedia. <https://www.investopedia.com/why-magnificent-seven-stocks-are-having-their-worst-day-on-record-8683011> (accessed October 17, 2024).
- Lek, L. (2024). *Lawrence Lek*. <https://www.lawrencelek.com/> (accessed October 28, 2024).
- Lombroso, C. (1899). *Crime: Its causes and remedies* (H. P. Horton, Trans.). Little, Brown and Company.
- Lorde, A. (2018). *The master's tools will never dismantle the master's house*. Penguin Books.
- MacQuillan, D. (2022). *Resisting AI: An anti-fascist approach to artificial intelligence*. Bristol University Press.
- Mahdawi, A. (2024, September 7) Elon Musk is intrigued by the idea women can't think freely because of 'low T'. *The Guardian*, <https://www.theguardian.com/commentisfree/article/2024/sep/07/elon-musk-women-testosterone> (accessed October 28, 2024).
- Met. (2024). *Composite portraits of criminal types*. The Met. <https://www.metmuseum.org/art/collection/search/301897> (accessed October 25, 2024).
- Nicoletti, L., & Bass, D. (2023). *Humans are biased. Generative AI is even worse*. Bloomberg. <https://www.bloomberg.com/graphics/2023-generative-ai-bias/> (accessed October 28, 2024).
- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.
- Pasquinelli, M. (2023). *The eye of the master: A social history of artificial intelligence*. Verso Books.
- Peel, M. (2024, July 24). The problem of 'model collapse': How a lack of human data limits AI progress. *Financial Times*. <https://www.ft.com/content/ae507468-7f5b-440b-8512-aea81c6bf4a5> (accessed October 17, 2024).

- Prabhu, V. U., & Birhane, A. (2020). *Large datasets: A pyrrhic win for computer vision?* arXiv:2006.16923v2. <https://arxiv.org/pdf/2006.16923> (accessed October 17, 2024)
- Rosenfeld, C. P. (2023). *Digital worldbuilding and ecological readiness*. Rowman and Littlefield.
- Selwyn, N., Campbell, L., & Andrejevic, M. (2023). Autoroll: Scripting the emergence of classroom facial recognition technology. *Learning, Media and Technology*, 48(1), 166–179. <https://doi.org/10.1080/17439884.2022.2039938>
- Steyerl, H. (2023). Mean images. *New Left Review*, 140(1). <https://newleftreview.org/issues/ii140/articles/hito-steyerl-mean-images> (accessed October 17, 2024).
- Suchman, L. (1987). *Plans and situated actions: The problem of human–machine communication*. Cambridge University Press.
- Thorp, J., & Buschek, C. (2024). *Models all the way down*. Knowing Machines. <https://knowingmachines.org/models-all-the-way> (accessed October 17, 2024).
- Weil, E. (2023, September 25). Sam Altman is the Oppenheimer of our age. OpenAI’s CEO thinks he knows our future. What do we know about him? *New York Magazine, Intelligencer*. <https://nymag.com/intelligencer/article/sam-altman-artificial-intelligence-openai-profile.html> (accessed October 17, 2024).
- Williamson, B., Macgilchrist, F., & Potter, J. (2023). Re-examining AI, automation and datafication in education. *Learning, Media and Technology*, 48(1), 1–5. <https://doi.org/10.1080/17439884.2023.2167830>
- Wong, J. C. (2020, December 4) More than 1,200 Google workers condemn firing of AI scientist Timnit Gebru. *The Guardian*. <https://www.theguardian.com/technology/2020/dec/04/timnit-gebru-google-ai-fired-diversity-ethics> (accessed October 28, 2024).
- Wong, R. Y., Madaio, M. A., & Merrill, N. (2023). Seeing like a toolkit: How toolkits envision the work of AI ethics. *Proceedings of the ACM on Human–Computer Interaction*, 7, Article 145. <https://doi.org/10.1145/3579621>