An in-depth, easy to read explanation of the current state of AI, and the unethical nature of many popular AI tools

Antsstyle (@antsstyle)

antsstyle3@gmail.com

Version 1.1, May 30, 2023

Contents

Abstract

The training of AI models presents a unique technological problem. Much like money laundering, AI can enable the "laundering" of data, allowing the effective plagiarism of copyrighted work without the ability to easily prove such theft is taking place. This is particularly a problem for artists and other creatives whose work current algorithms can easily analyse, such as images and text writing, but presents a general problem in many fields.

Recent high-profile AI uses have been marketed to the public as examples of technical innovation. However, it is more accurate to say that they are the result of data theft; their sudden leap in performance is due to the vast amounts of data they now have access to, much of it not being used with the permission of its owners. Many AIs also depend on gargantuan pools of underpaid human labor to classify objects for them, a stark contrast to the supposed intelligence of the technologies themselves.

The aim of this article is to explain the current state of AI technology as simply as possible, without requiring technical understanding, and showing examples from real life and popular culture to explain some of the issues. It also goes into detail about the nature of current AI systems, public misconceptions about the actual abilities of said AI systems, and other relevant topics regarding the use of and functionality of AI. It also details the regulatory and technological hurdles of stopping unethical AI, and proposes possible solutions to those ends.

In addition, one section of this article contains an explanation of Glaze, a software tool for protecting images from being used by unethical AI. I am not a member of the Glaze team or involved in its development; however, its efficacy and mechanisms make it one of the best anti-AI tools for artworks in my view, to use while legal regulations catch up with the AI industry.

Acknowledgements

I would like to thank Professor Ben Y. Zhao for his invaluable feedback on the first version of this paper, and for his insight and clarifications on the subject of AI attack methods.

Changelog

Version 1.1 (May 30, 2023)

- Added changelog section to paper.
- Moved adversarial examples to Section 6, and expanded the section to include additional information about targeted vs untargeted adversarial attacks on AI models.
- Expanded Section 2 (Ethical problems) to detail fair use laws, exploitation of opt-out clauses, economic coercion, and the inability of AI to perform meaningful transformations.
- Added Section 10 (Non-solutions), detailing the risks of bad solutions to unethical AI.

Version 1.0 (May 14, 2023)

Initial version of paper published.

1 The current (primitive) state of AI and ML technology

AI and ML (Machine Learning) technology is still in very early stages. In the minds of the public, AI seems to be advancing very quickly, but in fact AI is still very much in its infancy compared to what public perception might suggest. The term "machine learning" is much more relevant to current AI algorithms; they are not intelligent, or even close to intelligent.

Most "AIs" that the public are currently familiar with are, at their core, little more than statistical models. Underlying such AIs as ChatGPT for example are LLMs (Large Language Models); art-generating AIs are underpinned by very similar models.

1.1 Large Language Models (LLMs)

A Large Language Model is, in simple terms, similar to a kind of database: a gigantic pile of information about words and sequences of words, and relationships between them. The model is not intelligent; it simply contains a huge amount of computed relationships between the words and sequences. For example, it might record that statistically, the word "blue" is sometimes followed by the phrase "da ba dee", or that use of the phrase "Scott Morrison" is correlated with use of the word "coal".

The model knows nothing of language syntax, grammar, semantics, or any other rules that govern how language is used. It functions solely as a statistical model; if for example it has been given a billion essays written by humans, what may "appear" to be an understanding of syntax and grammar may be seen in the model's output simply due to the probabilities it has inferred from analysing the data set it has, despite it having no concept of syntax or grammar.

For an exceptionally well written, easily readable explanation of LLMs in more detail, I highly recommend reading Murray Shanahan's paper on the topic, "Talking About Large Language Models" [1]. Here is a paragraph from it, which I think does an incredibly good job of clarifying exactly how an LLM works in practice:

> Suppose we give an LLM the prompt "The first person to walk on the Moon was ", and suppose it responds with "Neil Armstrong". What are we really asking here? In an important sense, we are not really asking who was the first person to walk on the Moon. What we are really asking the model is the following question: Given the statistical distribution of words in the vast public corpus of (English) text, what words are most likely to follow the sequence "The first person to walk on the Moon was "? A good reply to this question is "Neil Armstrong".
>
> - Murray Shanahan, *Talking About Large Language Models* [1]

Recognising that AIs are not "thinking", and are simply being asked to provide the 'most likely' answer from the data they are trained on, is perhaps the most important step in seeing just how primitive current AIs are.

1.2 Understanding the total dependence of AI on data

It is inevitably the case that the power of any AI is entirely dependent on the total data it has available to it. Consider a simple mathematical equation:

$$2a + b = 10$$

We can propose an infinite number of solutions that satisfy this equation. Some examples:

$$a = 5, b = 0$$

$$a = 3, b = 4$$

$$a = 1, b = 8$$

However, this list goes on forever; these solutions give us no useful information. In mathematics, this is known as an underdetermined system; we don't have enough information to figure out the answer.

The problem here is simple: our brainpower, algorithms and any other thought process or intelligence available does not allow us to find any solution to this problem. Without sufficient data, the intelligence or processing power of whatever is trying to solve this problem is completely irrelevant.

Consider a hypothetical scenario, where we possessed a system of unlimited power and intelligence, and asked it to predict the weather for planet Earth over the next hundred years given a dataset containing today's weather forecast for New York. The lack of data overpowers all other considerations: the system's power and intelligence do not even factor into the equation. The data it has been given is not sufficient for it to predict anything.

This problem lies at the heart of why recent AIs are given so much attention: their seeming leaps in ability are much less due to technological improvements and more due to the amount of data they have to use, as a result of opportunistic and unethical mass scraping of Internet data.

1.2.1 Differences with human thought

A common counter-argument to the above is that humans possess the same problem: we, too, are unable to solve any problem where we have insufficient data. This overlooks many critical factors that distinguish humans from artificial intelligence.

To take an example, the average 20-year-old human has absorbed a truly gargantuan amount of data - much of it, never actively thought about. Humans who live in houses and thus frequently use doors, have a wealth of data they have accumulated about the texture, thermal conductivity, material strength, inertia and other properties of various types of door handles and the materials they are made from. Humans who often drink from a watertight container, like a glass or cup, have data about the weight, viscosity and physical properties of water and other primarily aqueous solutions.

Such data is taken for granted, yet we depend on it to function daily. When we open a door, we use our knowledge to estimate the weight of the door and its inertia, so we can apply the appropriate amount of force to open the door without swinging it open violently and damaging something. When we drink from a glass of water, we use our knowledge of water viscosity and weight to pour the water at an acceptable speed so we can drink it.

Theoretically, it is possible for an AI to learn this. However, a lot of this mundane data is not recorded in a way that is useful to AIs, and the AI is unable to connect the various required bits of data it needs - i.e. to understand the context. In our example above for instance, drinking from a container does not only require knowledge of the weight of the container and its contents, but also the range of drinking speeds of the entity who is using the container, and inferring the weight of the container requires data of the container's construction. Drinking 20ml of beer from a metal tankard requires significantly more force than drinking 20ml of beer from a hollow shot glass, since the weights of each container are very different; simply looking at the container does not provide this information.

1.3 Imitating intelligence with huge data sets

One of the fundamental issues present in the public discussion about AI is a misunderstanding, or perhaps more accurately an under-estimation, of the ability of programs that utilise huge data to look like intelligence under specific circumstances.

Given enough data, an unintelligent program with no concept or understanding of a topic can convincingly appear, to a person with little or no knowledge of that topic, to understand that topic simply by having enough information to regurgitate. This is not limited to AI; human fraudsters, "psychics" and other types of deceptive behaviour rely upon this premise, as does the ability of someone who has read sufficiently on a topic to seem like an expert even if they are not.

The character Henry Dobson of the television series House MD is a good example; despite having no medical education or training, he manages to convince other doctors that he is a doctor from his long experience auditing medical classes, and is only discovered upon refusing to perform medical procedures which his reading did not give him the skills to do.

1.3.1 The Infinite Monkey Theorem

The Infinite Monkey Theorem is an excellent analogy for why AI can appear to be advanced when it is not. The theorem states that:

> "A monkey hitting keys at random on a typewriter keyboard for an infinite amount of time will almost surely type any given text, such as the complete works of William Shakespeare."

In very much the same way, given a huge amount of data and months of training time, an unintelligent 'AI' can appear to produce something coherent and intelligent, when in fact it is simply a product of computational brute force and sheer quantity of data; not unlike having a trillion monkeys, a trillion typewriters, and a time machine. The AI 'understands' absolutely no part of the output it has generated, nor the data it generated it from; it does not understand language syntax, grammar, meanings of words, or anything else.

1.3.2 The role of survivorship bias and confirmation bias in the perceived power of AI

Survivorship bias is the logical fallacy of concentrating on entities that "succeeded" in some process or in some field, whilst ignoring those that did not.

AI content generation (such as text-to-image generation) has a very high amount of this, because the only content typically shown is that which the person using the AI decided to show - the 'successful' results. The many other results, that generated nonsensical or clearly flawed content, are silently discarded and not shown. Since the public only sees the "successful" generated images, this gives a heavily biased view of the actual ability of an AI to generate plausible images.

This is important in the context of AI, because for something to be considered intelligent or competent, there is an underlying expectation that said thing has at least a better success rate than random chance. For instance, nobody would consider an electrician who only manages to perform a job correctly in 1% of cases, or a baker who could only bake bread correctly 1% of the time, 'competent' at their profession. An AI that requires extensive trial and error, "prompting", and comes up with a passable result only after producing many failed ones could hardly be considered intelligent by any definition.

Confirmation bias is the logical fallacy of interpreting something as confirming existing beliefs, even if there is no evidence to support it. A common example would be astrology horoscopes or fortune telling, where the predictions are so broad and generalised that they are bound to fit most scenarios; the reader then incorrectly interprets this as proof of the prediction being correct.

"Anthropomorphic" AI uses, like chatbots, exploit this bias: even though they are not intelligent or capable of carrying conversation, a chatbot using an LLM and utilising pre-prompting techniques can "appear" to be conversing or understanding language, giving the illusion that it is far more capable than it really is. The language used to describe such AIs, by the companies who make them, frequently adds to this bias.

1.4 The manual labor component of AI data

AIs are often marketed to the public as amazing innovations and pinnacles of technological achievement, as if it is magic machinery that just "knows" how to do things. The less marketable reality is that many uses of AI are also dependent on gigantic numbers of humans performing manual tagging and classification work. A prominent example of this would be Google's reCAPTCHA, which functions both as an anti-bot measure and also as a way to train AI to recognise features of images [2].

People who do this work for platforms that offer payment for tasks are known as "microworkers" or "crowdworkers", who perform huge numbers of small tasks for even smaller pay (the ethics of which are detailed further in section 2.3 of this paper). This pool of human labor is not small; Amazon Mechanical Turk claims to have over 500,000 microworkers [3], while Hive.ai claims to have 2.5 million microworkers [4]. Tasks are generally quite simple, such as selecting areas of an image that contain a given object or objects, or listening to audio speech and transcribing it into text.

This gives the AI the data it needs to better recognise and classify such objects in future, provided a huge enough amount of human classification has taken place. However, this isn't a one-time barrier: the huge pools of human labor needed for these AI uses to work will be constantly needed, as objects and other real-life data changes. To take an example, an AI for a car has to be trained to understand what various road signs look like in order to recognise them accurately in different conditions; this means a lot of human classification. You may well have had to do this yourself for traffic lights or other road signals as part of reCAPTCHA challenges on websites; what happens, then, if society changes the normal design of road traffic signs?

A hypothetical concept of how a stop sign for traffic could change in future society.

If in a hypothetical future, MC Hammer becomes the standard symbol for a stop sign, AIs would have to be retrained to be able to recognise them accurately. Many aspects of society are constantly changing; as a result, the manual labelling, tagging and classifying of data by humans will be in constant demand for them to work. This rather manual, laborious process is in stark contrast to the marketing of AIs as intelligent machines running of their own accord.

2 Ethical problems in popular AI

2.1 LAION-5B and Stability AI

The LAION-5B dataset, created by the LAION nonprofit and released in March 2022, is a huge dataset consisting of 5 billion image-caption pairs. It is notably used by the for-profit AI company, Stability AI, the makers of the Stable Diffusion text-to-image AI. LAION-5B is highly unethical, for several reasons.

Firstly, LAION-5B's data was sourced from Common Crawl, a nonprofit that scrapes masses of Internet data. This data includes copyrighted content, which is not supposed to be used for illegal purposes such as copyright infringement or theft, as per Common Crawl's Terms of Service.

Secondly, LAION is legally registered as a nonprofit entity. However, its own paper declares that LAION-5B's creation was funded by Stability AI and huggingface, both for-profit companies [5]; both companies attracted hundreds of millions of dollars in investment later that year [6] [7] [8]. This unethical behaviour is not unprecedented; Big Tobacco and Big Oil have been found to influence seemingly impartial peer-reviewed research by funding said research [9] [10], which has led journals such as the British Medical Journal to reject any paper regarding tobacco that is partially or fully funded by the tobacco industry [11].

Thirdly, LAION-5B disclaims liability for the source data by only storing URLs to the images, shifting responsibility to those who use it [12]. The published paper for LAION-5B also states the following in bold text on page 3:

> **"we strongly recommend that LAION-5B should only be used for academic research purposes in its current form. We advise against any applications in deployed systems without carefully investigating behavior and possible biases of models trained on LAION-5B."** [5]

It is hard to see how LAION could have expected that the for-profit companies funding LAION-5B were not going to use it for profit. It also seems highly questionable for LAION to have accepted funding from companies with evident conflicts of interest, especially in the light of the Big Tobacco and Big Oil research influence mentioned above.

2.2 OpenAI, DALL-E and ChatGPT

OpenAI, a non-profit with a for-profit subsidiary, is responsible for the creation of DALL-E and ChatGPT. Microsoft recently invested $10 billion USD into the for-profit part of OpenAI [13], which gives an idea of OpenAI's perceived profit potential.

ChatGPT is currently powered by the GPT-3.5 and GPT-4 models. GPT-3, a previous version, was trained on predominantly scraped internet data, including copyrighted data, according to page nine of OpenAI's paper on the subject [14]. There is a level of irony in the fact that despite its name, OpenAI did not release the weights or technical details for GPT-4, and ChatGPT Plus (which uses GPT-4) is predominantly only available to paying customers [15].

2.3 Exploitation of low-paid human labor

The microworkers who perform AI labelling, tagging and classifications tasks are generally hired as "independent contractors", making them legally self-employed. This allows their employers to avoid upholding many of the rights given to employed workers, such as paid holiday leave, pension rights, and the right to the minimum wage.

In a study of 2,676 microworkers doing jobs for Amazon Mechanical Turk, the microworkers were found to be earning a median wage of ∼$2 an hour, with only 4% earning above the US minimum wage of $7.25 an hour [16]. An extremely comprehensive report by the International Labour Organisation in 2018 [17] detailed many aspects of the microwork industry as a whole; while there is far too much in it to cover in this paper, here are a few of the important findings:

- The median hourly earnings of microworkers was $3/hour (p.74)
- A clear correlation demonstrating that microworkers in developed countries were paid more than those in developing countries, isolating for other factors (p.76)
- The most popular reason for doing microwork was to complement pay from other jobs (p.62)
- 40-50% of microworkers did not have enough savings to cover an emergency (p.83)

This commercial exploitation of workers, by using self-employment arrangements to avoid giving workers the rights they are entitled to, is not new; similar structures exist in other areas of the so-called "gig economy". Uber was forced to give its workers access to paid holiday leave, the minimum wage, pensions and other workers' rights after employment tribunals and the UK Supreme Court ruled against its argument that its workers were "independent self-employed partners" who were not entitled to such rights [18]. Deliveroo has won similar cases [19], but is being repeatedly challenged by unions [20].

2.4 The defiling of the term "Open Source"

In software, "open source" refers to a program or application for which the source code has been made publicly available. Open source software often has many benefits: it is generally free to use, allows others to build upon it, and having free open source tools improves industries by making all manner of tasks easier. Some examples of well-known open source software include Linux, Blender, OBS, the VLC media player, and Chromium (the core component of Google Chrome and several other browsers).

AI companies have a habit of marketing their "open source" credentials to give the impression of transparency. OpenAI implies it in the name, Stability AI's CEO claims Stable Diffusion is "democratizing image generation" [21], and so on. This is both misleading to the public and corrosive to the reputation of open source software.

A critical part of the trust given to open source software relies on the fact that no part of what the software does is "unknowable" to the user. Generally speaking, most open source software does not rely on private, undisclosed data; for instance, if you install VLC to watch movies, the only data it needs or uses is the data you provide it, i.e. the movie file you want to watch. Any other data it needs to download, like updates, is open source and thus inspectable. If for example VLC decided it had to download 5GB of mystery undisclosed data before it could run properly, its open source nature would no longer be relevant for trust purposes, as you wouldn't know what that data was for or what it was doing. (To be clear, that is only an example: VLC does no such thing that I know of).

If an open source project can't function without a huge amount of undisclosed data with no verifiable origins, the fact that it is open source is of no relevance when it comes to trust. This is exactly the problem with many current AIs: the code is open source, but the models (and more importantly, the data those models are trained on) is hidden from public view in most cases. AIs, unlike other programs, have no ability to function in any useful way without massive amounts of data; without the models trained on masses of data, they can't do a thing.

In other words, AI companies are marketing themselves as "transparent" and "open source" by releasing source code, despite the problem that this supposed transparency does not tell us anything of significance; the real "core" of a given AI project is the models it uses, and by extension, the data those models trained on. The reasons for doing so are clear: actually disclosing the source data would have huge legal implications, given said data is almost never used with the permission of its owners, and being seen as transparent is good for PR.

Stability AI and LAION might, on first glance, seem to be defying this rule; making their data set and models public. It is more of a carefully calculated maneuver: LAION delegates responsibility for usage of the dataset to users, and Stability AI disclaims liability for content generated by users. The likely aim of this is to "democratize" the legal consequences of unethical AI without democratizing the profits, rather than taking the risk of keeping it all private and potentially being held accountable if it was discovered later.

It is as yet unclear if this attempt will succeed. Lawsuits are currently in progress against Stability AI, including separate lawsuits by Getty Images [22] and a class action lawsuit by the three artists Karla Ortiz, Kelly McKernan and Sarah Andersen [23]. LAION is being sued by photographer Robert Kneschke after it refused to remove his work from their data set, and instead sent him a letter insisting he pay damages or face legal action [24].

2.5 The pretense of 'fair use'

'Content-generating' AIs often claim that their use of copyrighted works falls under fair use. In order to explain the utter absurdity of this position, first we must understand the concept of fair use.

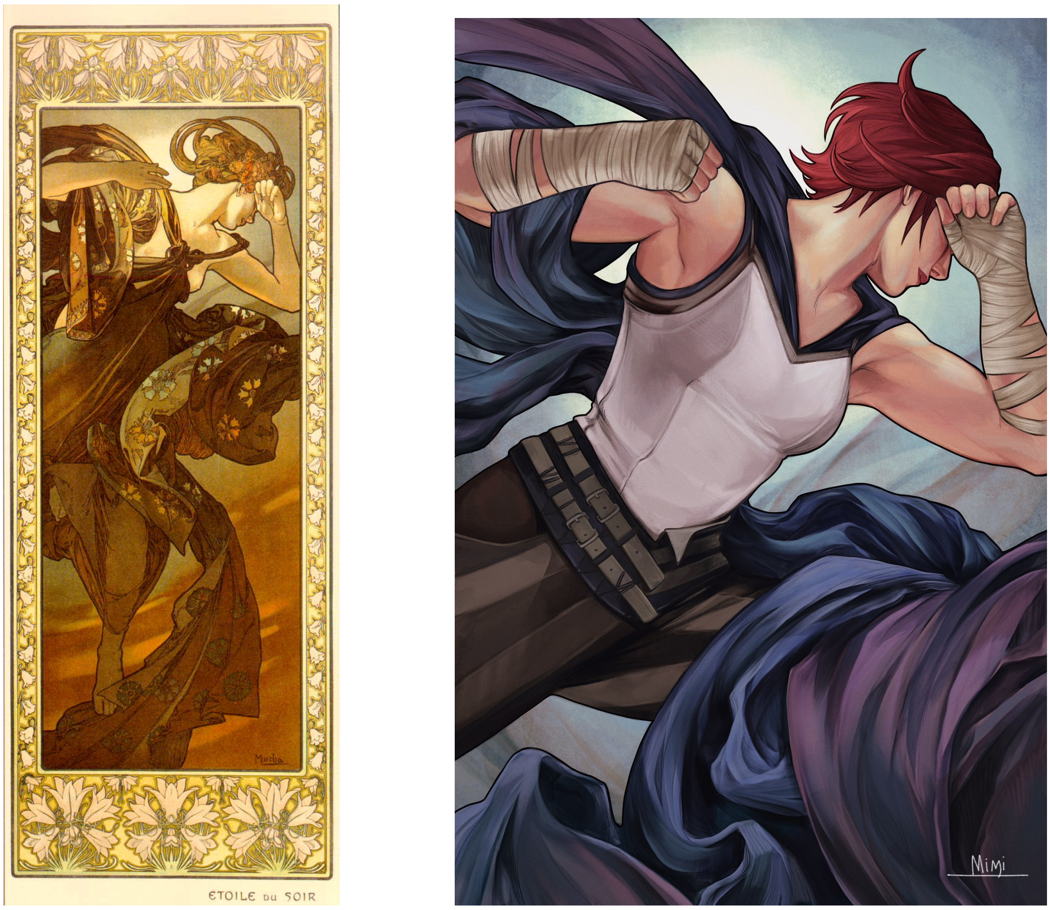
In US copyright law and many other jurisdictions, "Fair Use" exists to allow legitimate use of copyrighted works for certain purposes; the aim is to strike a balance between protecting the rights of creators, and allowing innovation, improvement and derivatives of existing works. Many factors are considered when determining if a given case counts as 'fair use': for instance, non-profit or charitable uses of a copyrighted work are more likely to be considered fair use, though that in itself does not automatically qualify it.

When it comes to commercial, for-profit works, the primary condition to satisfy for something to be 'fair use' is that it is **transformative**. For example, being inspired by a musician's song and incorporating elements of it into another track could count as fair use, depending on how much was incorporated and how much original contribution the new artist's track has. Changing 1% of it to a different tone and calling it your own music would count as **'transformative'**, but that wouldn't be a meaningful transformation: it would be plagiarism in all but name. As another example, suppose someone takes a prominent academic article, changes every other letter to be a capital letter and then publishes it as their own work. That would be a significant **'transformation'** of the original work in literal terms, but would not change the meaning of the text or in any way count as fair use.

2.5.1 'Meaningful' transformation

The concept I will call "**meaningful transformation**" is crucial. Taking an artwork and adding a slight blue filter on top of it changes every single pixel in the image; that doesn't count as a meaningful transformation. Splitting an artwork into 32 equally sized squares, and rearranging the squares randomly, would also be a significant literal 'transformation' yet provides no original contribution of its own; it is not **meaningful**.

A "**meaningful transformation**" requires a significant original contribution from the new artist, in addition to what they have taken or adapted from an existing work. In paintings, music and other forms of art, this original contribution comes from the artist; their personal experiences and knowledge are used to provide this original contribution. Let us consider an example.



On the left, *'The Evening Star'* by Alphonse Mucha (1902). On the right, *'Vi as The Evening Star (Alphonse Mucha, 1902)'* by artist @_mimimaru.

Here, we have a derivative of an existing work: the artist has been inspired by a pose from an existing artwork, and used it to create a new artwork. Of course, in this case copyright law would not apply due to the age of the existing artwork, but the conceptual question of fair use is still easy to think about.

This is a good example of a **meaningful transformation**; although the new artwork is clearly inspired by and takes elements from the old, it depicts an entirely different character with different anatomical features and clothing. Since the new character is not wearing a dress, the clothing folds and swirls that help to create the composition of the artwork have been implemented differently, informed by the artist's knowledge of the character and turning the clothing folds into a cape to depict her heroic qualities.

2.5.2 AI 'transformation'

Content-generating AIs, by definition and by the nature of the data available to them, are functionally incapable of **meaningful transformation**; they are solely capable of remixing or rehashing existing works. Unlike a musician or a painter, who can use their personal experiences and knowledge of the outside world and events to add their own contribution to a given work, an AI is only capable of mindlessly adding contributions made by other people. This is a simple and fundamental data problem, as explained in the earlier section 'Understanding the total dependence of AI on data'.

This is not to say that all AI uses are infringing on copyright; for example, AI and ML are used for such purposes as optical character recognition (OCR), which allows the easy digitisation of written works into digital form. Although such uses simply copy the existing text, they are not claiming to 'transform' the work, nor claiming to own it or be its author.

Recent legal developments have justifiably cast doubt on AI 'transformations' automatically being considered fair use. Jon Baumgarten, a former General Counsel of the US Copyright Office, recently wrote a letter to the US House Judiciary Committee's IP Subcommittee hearing, emphatically disagreeing with that position [25].

2.6 The use of opt-out consent systems

Opt-out consent, by definition, is not consent. European GDPR (general data protection) regulations explicitly detail this:

> "Consent should be given by a clear affirmative act... such as by a written statement, including by electronic means, or an oral statement. This could include ticking a box when visiting an internet website, choosing technical settings for information society services or another statement or conduct which clearly indicates in this context the data subject's acceptance of the proposed processing of his or her personal data. Silence, pre-ticked boxes or inactivity should not therefore constitute consent."
>
> - European GDPR law, recital 32 [26]

Tech companies have frequently attempted to gain dubious consent by pre-ticking boxes, having "you consent unless you tell us otherwise" popups, and other strategies to avoid actually making the user read what they are consenting to (which would reduce the number of users who would consent). AI companies, with a huge amount to gain by stealing data, have taken this a few steps further: assuming all data available on the internet can be used for AI purposes unless a creator or copyright holder specifically opts out, and sometimes not even then.

Much of the content useful for AI training available on the public internet is not held by large corporate entities, but by individual creators, who do not have the time to seek out every possible AI model to opt out from. AI companies know this; using an opt-in model would eliminate most of the data they train on, causing them to be unable to drive investor profits.

2.7 Economic coercion

It is a long-standing problem that tech giants - among other huge companies - have attempted, sometimes successfully, to force governments not to regulate certain technologies or ban specific practices. The aim of this is to ensure they can maximise their profits at the expense of consumers, hanging the threat of economic damage over the heads of governments by suggesting they might move their operations to other countries, losing the threatened country tax income and jobs. AI is heading down the same path due to the money being invested in it and the players involved, though it is easily arguable that should regulation fail, the consequences would be far bigger than in previous cases.

2.7.1 Recent examples

There is no shortage of examples of this behaviour. Meta Platforms Inc (owner of Facebook) threatened in 2022 to pull out from European markets if new data protection laws were enacted [27]. In 2021, Apple threatened to leave the UK market, after Opus Technologies sued it for using its patented technologies without fair compensation [28] [29].

Elon Musk, owner of Tesla, threatened to move a large Tesla carmaking factory out of California when the state government insisted he could not reopen it due to coronavirus restrictions [30] [31]; workers at the factory alleged they were fired for taking unpaid leave after Musk had previously stated workers who felt uneasy did not have to return to work [32]. Nine months later, data showed that COVID-19 cases spiked at the factory after its reopening [33].

2.7.2 AI examples

Sam Altman, CEO of OpenAI, recently suggested OpenAI would have to pull out of Europe if EU regulations regarding AI are passed [34], a veiled threat to try and coerce European regulators into allowing OpenAI to continue using stolen data unhindered. We should view his statements - both those against EU regulation and those in favour of US regulation - exactly as we would if the CEO of British American Tobacco warned EU regulators against regulating cigarettes.

This conflict of interest could not be larger; the very existence of OpenAI, and the extent of its profitability, depend entirely on which regulations end up being enacted and how waterproof and strong those regulations are. It is very telling that AI experts and researchers who do not have a financial conflict of interest overwhelmingly do not share the views of AI companies or their leaders.

3 The weakness of data in current AI systems: context

The data that current AIs are trained, despite its seemingly gargantuan nature, is actually rather limited. An AI which has been trained on 5 million essays about trees has no understanding of what trees are, what plants are, what nature is, what an essay is, what language syntax is... it knows nothing but relationships between words and data.

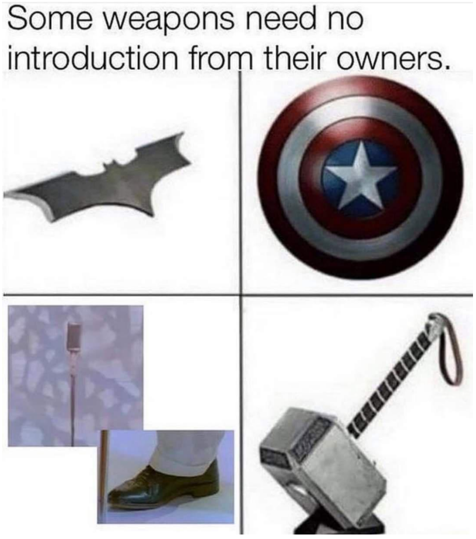
Similarly, AIs trained on images to 'generate' artwork have no understanding of what art is, or any of the components of images they have 'seen'. An AI that has seen millions of horses in pictures has no concept of what a horse is, only that a particular shape seems to appear frequently in the data it has been given. This is very different from humans, where the human has other contextual information about the topic that allows them to both avoid mistakes and infer attributes of the topic that the AI cannot.

To illustrate this more clearly, the next sections some art forms that demonstrate the importance of context.

3.1 Examples: Memes

Memes and many other forms of humour, by their very nature, depend upon an understanding of underlying context that is not contained in the meme itself. Below are a few examples.

An AI would have great difficulty understanding the meaning behind any of these images, as doing so requires vast amounts of contextual knowledge that is difficult to obtain by analysing isolated image data.

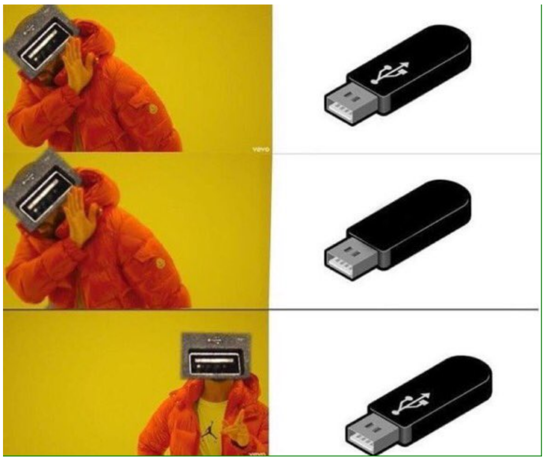


One of these is not like the others. But which one?

This meme relies on the viewer knowing of three fictional characters, and one real person. The fictional characters Batman, Captain America and Thor all use the weapons shown in their section of the picture; said weapons are iconic, recognisable aspects of those characters.

The humour comes from the fact that the bottom left image depicts Rick Astley's foot and a microphone from his music video Never Gonna Give You Up, which became a meme of its own due to its use to prank people. A foot and microphone, of course, are not "weapons" at all in a literal sense.

This same underlying context makes it difficult an AI to 'understand' this image. It would have to know the cultural importance of Captain America, Batman, Thor and Rick Astley, all of which require knowledge of the franchises the first three appear in, the reasons why the symbols depicted are so well associated with those characters, and more.

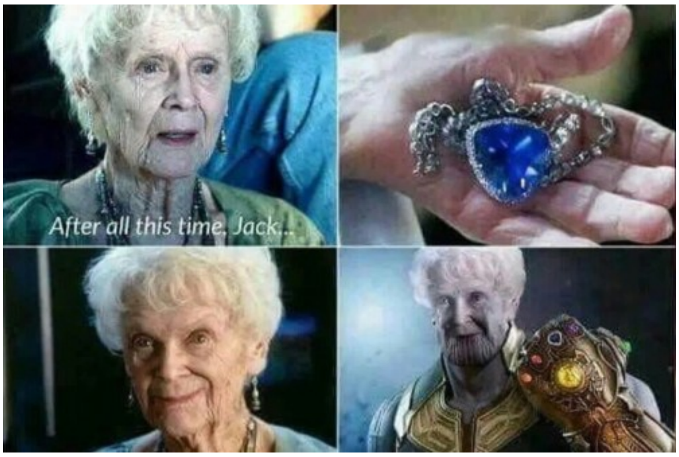


A universally frustrating experience for humans.

This meme depicts the musician Drake, with the humour being the often-frustrating difficulty of plugging in a USB device the correct way up.

Any human who has ever plugged in a USB device knows this pain, and yet for an AI this is very difficult to ‘understand’. Unlike an AI, a human relies on eyes for vision, and is generally trying to plug a USB device into a port located outside of our vision range, such as the back of a PC. This makes it difficult for us to figure out exactly where the USB port is and which way up we need to insert the USB device. Added to this problem is the lack of tactile or other sensory feedback on most USB devices, being smooth on both sides, making it hard to tell with our hands which side is which.

An AI, not using human hands or human eyes and lacking both experience and relevant data of both, does not possess the capability to comprehend this. It has not had to deal with this situation and cannot find any useful data about it, as humans do not generally have a motive to record such data outside of very specific circumstances.



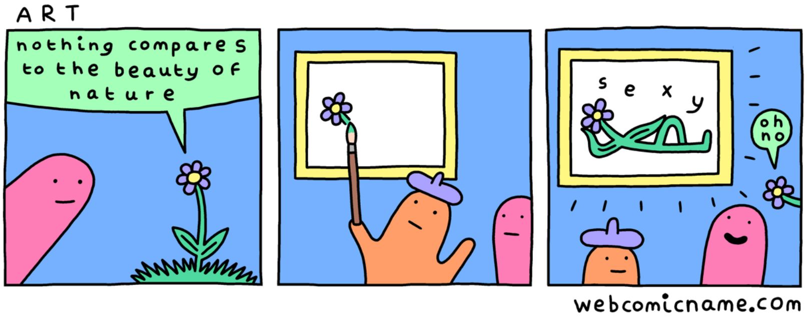
Legends say that half of all icebergs were never seen again.

In a similar fashion to the first meme shown here, this image relies on contextual knowledge of the movie Titanic and the fictional character Thanos.

If a human has not seen the movie Titanic, or any of the Marvel Cinematic Universe films, they would have absolutely no idea what this image is, who it depicts, or what its purpose is. An AI has a much bigger version of that problem: it does not know what a film is, what a ship or the Titanic is, it does not know whether the Heart of the Ocean gem depicted in the top left has any relation to the Infinity Gauntlet, what a gem is, or anything else. The contextual data it requires to obtain this is not readily obtainable.

3.2 Examples: Comics

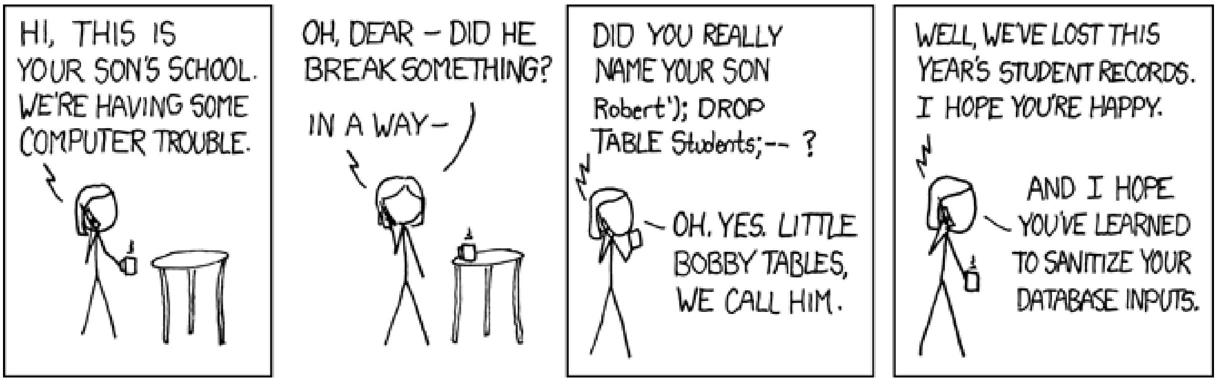
Another art form which AIs will greatly struggle to “understand” is comics. In the same manner as memes, comics generally rely on underlying context filled in by the viewer to elicit humour. Here are some examples.



That flower knows too much... (from webcomicname)

This comic relies on the viewer's experience and understanding of the sexualisation of many aspects of life, including art, and arguably also the lament of those who would like life to be less weighted in that direction. "Sex Sells" is an often-quoted phrase, alluding to the fact that that making advertisements, game characters and other things "sexy" frequently leads to higher sales and popularity, for instance.

An AI, not using chemical hormones to function and not having walked around observing billboards and TV adverts involving sexualised depictions, will be unable to determine the intended meaning of this comic.


HI, THIS IS YOUR SON'S SCHOOL. WE'RE HAVING SOME COMPUTER TROUBLE.
OH, DEAR – DID HE BREAK SOMETHING?
IN A WAY–
DID YOU REALLY NAME YOUR SON Robert'); DROP TABLE Students;-- ?
OH. YES. LITTLE BOBBY TABLES, WE CALL HIM.
WELL, WE'VE LOST THIS YEAR'S STUDENT RECORDS. I HOPE YOU'RE HAPPY.
AND I HOPE YOU'VE LEARNED TO SANITIZE YOUR DATABASE INPUTS.

Illidan Stormrage would not approve of those database statements. (from xkcd)

This specific comic will confuse most non-technical readers; its intended audience is software engineers. The joke is about a common vulnerability of some websites and databases, known as SQL injection, a type of code injection attack. SQL injection relies on exploiting the syntax of the language used to write database queries, and causing the database to do something it wasn't intended to do; in this case, dropping (deleting) the table of student records in the database.

For an AI to understand why this comic is humorous, or even what it is talking about, would require significant software engineering knowledge. A human without technical knowledge can understand a limited amount of this comic; it is clear that the school is not happy about losing the year's student records, and that the mother is being facetious and trying to teach them a lesson. However, without an understanding of the database language SQL, a non-technical reader will be unable to understand the significance of naming one's son "Robert'); DROP TABLE STUDENTS;–", or what it means to sanitize database inputs.

3.3 Examples: Art

While comics and memes rely on context, we can also see the weakness of AI by analysing an artwork, where physics, biology and other fields come into play. Let us examine a famous painting as an example.

"Napoleon Crossing the Alps", by Jacques-Louis David

The horse upon which Napoleon is riding is an animal, with hooves, commonly used for riding by humans and as a draft animal for pulling loads such as carts. Its skeletal and muscular structure allow it to move in certain ways; like other animals, its joints and muscles provide it with a limited range of possible motions, data that cannot be obtained from its surface appearance. Each of its muscle groups has varying amounts of power, much as in humans, where our legs are significantly stronger than our arms, for example.

An AI does not know the anatomy of a horse, nor its muscular structure or its maximum abilities in terms of endurance and load bearing. It does not know the horse's ability to gallop or canter in different terrains such as snow or mud, where the ground will give way more easily and thus be more difficult to maintain balance on. It does not know the significance of the horse's head; if you train the AI on images of horses where the head is hidden behind other objects, or fantastical depictions of headless horses, then as far as the AI knows, horses have no heads.

By using a sufficiently large pool of data, an AI can mimic this not by actually "understanding", but by statistics alone. Most artists will not draw, for instance, a horse with six legs. Thus, the AI - despite having no ability to understand this - will not "draw" that, simply because it does not seem statistically likely from the images it has analysed. However, when it comes to other examples, the AI may have less data. Without a truly gargantuan data set, the AI - unlike a human - may generate a picture of a horse galloping in mud, yet fail to have the mud flinging around everywhere.

A human, who has a large amount of data from observation, education and experience about mud and other types of terrain, knows that a heavy object impacting the mud and then applying a backwards and upwards force will cause the mud to fling upwards and backwards and out of the ground. They can recognise that this depends on the physical properties of the object hitting the mud (the surface area of impact, the material strength of the object, the weight of the object, etc), as opposed to the exact type of object. For instance, a car, a horse and a human all cause this effect to varying degrees, whilst a beetle is not big enough to fling large chunks of mud. The human, with this knowledge, knows that mud being flung into the air is not a unique property of horses, but is a general physical property of mud and other materials possessing similar characteristics.

An AI has no understanding of this; its only ability is to correlate the flinging of mud in specific pictures. If it has seen many pictures of humans running in mud, and mud flinging everywhere, it will associate these two activities without any underlying understanding of the mechanisms involved. Precisely because it has no understanding of the mechanisms, it will fail to associate this mud flinging with any other object that moves through mud, as its data gives it no reason to do so.

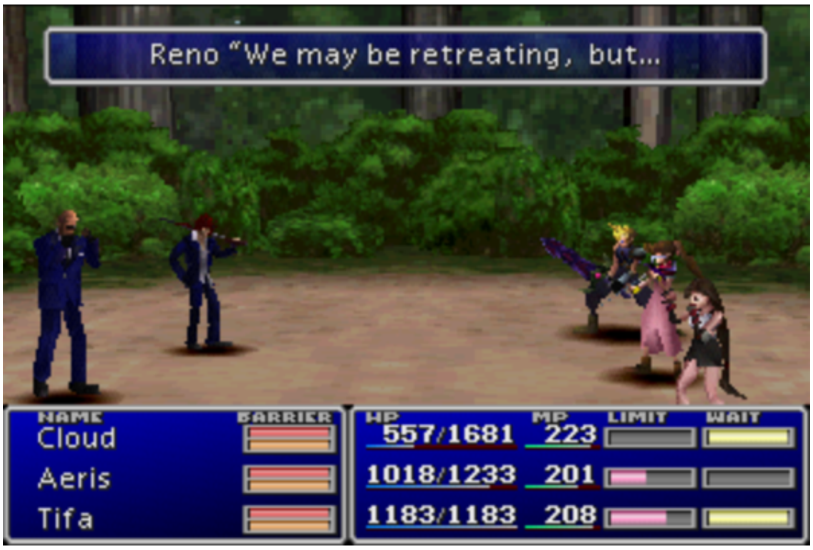
All of these shortcomings are important because in many situations, they can be very dangerous. An AI that is incapable of understanding different races of humans leads to racial bias; an AI that cannot understand different terrain types leads to major accidents, and so forth.

4 The primary strength of AI: limited, isolated environments

AI works best when it does not have to consider a potentially unlimited number of variables and factors. In order to explain this properly, we need to have a basic understanding of **heuristics**.

A **heuristic** algorithm or program is one that does not guarantee an exactly accurate answer. A good heuristic can provide an acceptably close approximation to the correct answer, whilst being much faster than a non-heuristic algorithm. This makes heuristics invaluable when an exact answer is not critical, or when time or processing constraints make finding the exact answer impractical or infeasible.

Heuristics are crucial to many programs, including video games, AI and more. To get an idea of their critical importance, let us consider a single frame (image) from a video game.



Reno “We may be retreating, but...

NAME	BARRIER	HP	MP	LIMIT	WAIT
Cloud		557/1681	223		
Aeris		1018/1233	201		
Tifa		1183/1183	208		

“...we’re still victorious” (Final Fantasy VII, 1997)

The processing power required to accurately show every visual, real-life detail is huge. A video game must finish fully rendering every detail in the entire scene 30 times per second (and in modern games, often 60 times per second) or it will experience frame lag, which causes the game to look ‘choppy’ and inconsistent.

Being able to calculate or ‘render’ these frames at a fast enough speed is a constant problem in video games. Producing fully accurate graphics, down to every last tiny detail, is unnecessary; objects only seen from long distances, for example, need not be fully detailed or accurately calculated. Strictly speaking, the term ‘heuristic’ refers to algorithms such as calculating the light on a given object, and not the details of the object itself, but the relevant point here is that a heuristic is an approximation.

AIs, because they share the same computationally expensive nature as video games, have the same problem. Trying to get an AI to fully understand every possible aspect of the outside world is not feasible for the long-term foreseeable future: there are too many variables and too many factors, and the amount of data required is far too high. However, if the AI only has to consider certain factors and does not have to be exactly accurate - if it can be treated as a **heuristic** and operate in a limited environment - then it can perform various useful tasks.

Refer to the appendices, Examples of good uses of AI and Examples of bad uses of AI, for examples and counter-examples.

5 Understanding the concept of data laundering

“Data laundering” refers to the problem of extracting technically useful data from an input, such as an image, that allows effectively plagiarising that data - but does not retain enough of the original data to prove that plagiarism has occurred.

The main difficulty in proving that data laundering has occurred is that much of the “original” data has been lost; a similar approach exists in money laundering, where the objective is to effectively “lose” as many records about the transfer of the money as possible, to obscure its origins.

To properly explain this in the context of data laundering, we must know the concept of data loss.

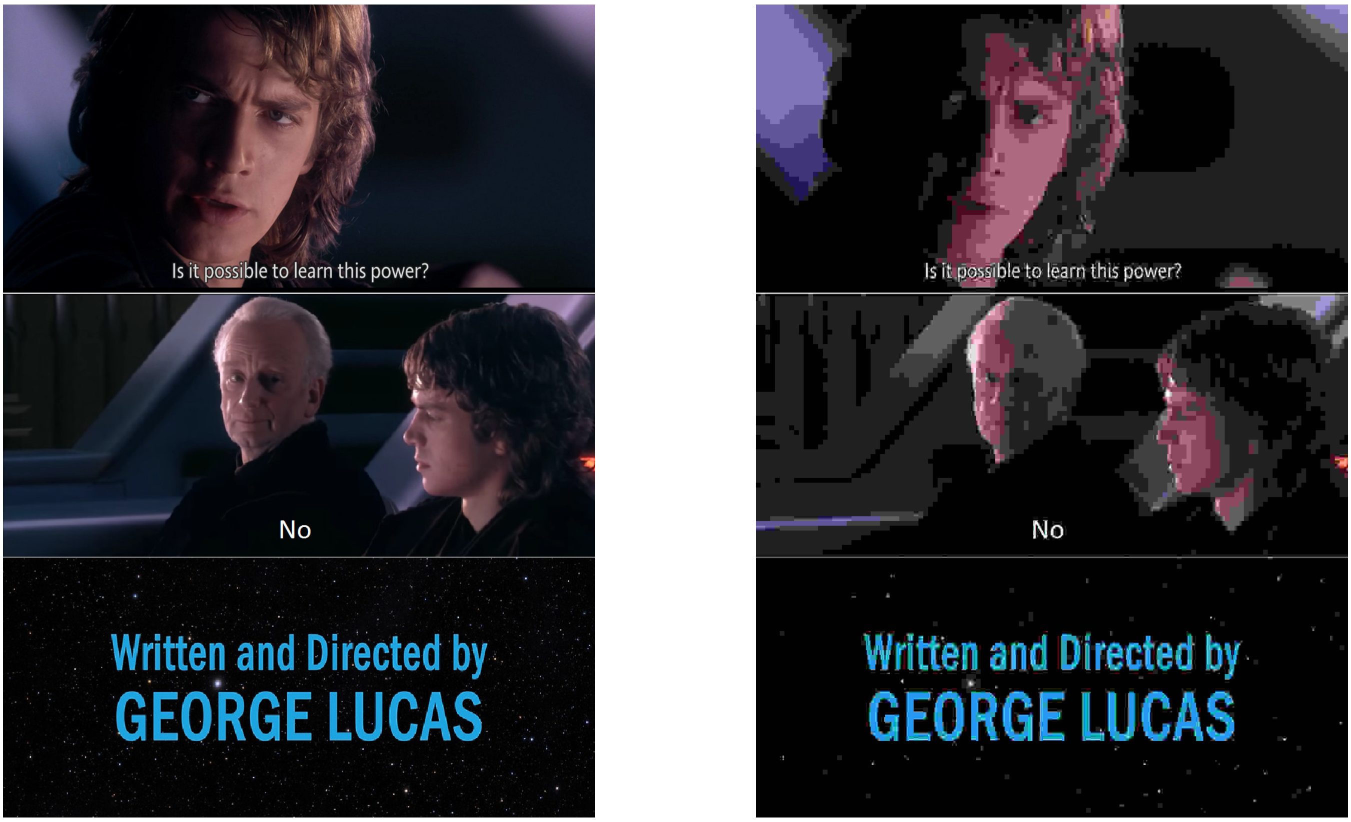
5.1 Data loss problem

The primary difficulty in identifying when an AI has used stolen data is, fundamentally, a problem of data loss. In order to illustrate what precisely data loss means, let us consider lossy image compression as an example.

5.1.1 Lossy compression

As a general concept, compression algorithms optimise how a file is represented and stored on a computer, to reduce the amount of memory or disk space needed to store it. **Lossy** compression discards some of the less ‘important’ data in the file, in contrast to **lossless** compression where none of the data is discarded.

The advantage of lossy compression is that in many applications, the full data is not needed or the quality loss is acceptable, and lossy compression generally results in far greater file size reduction than lossless compression. JPG, MP3 and MP4 - used for images, audio and video respectively - are three examples of prominent formats that utilise lossy compression.

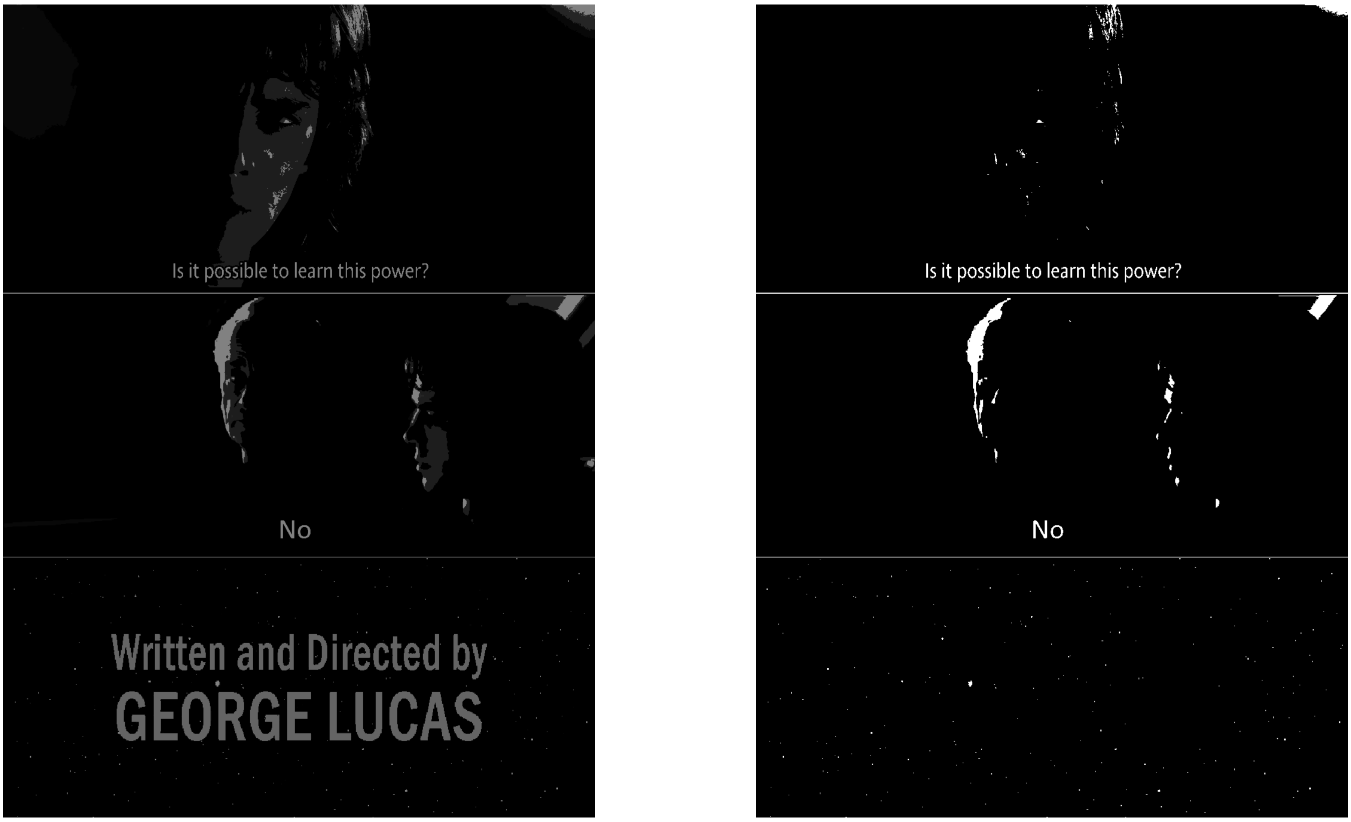


A JPG image at two different levels of compression. The image on the right is 6 times smaller in file size, but clearly a significant amount of detail has been lost.

Here is a Star Wars meme, showing Anakin Skywalker and Chancellor Palpatine in discussion. Even though the first JPG image is a fraction of the filesize of the full uncompressed image, it is readily viewable and understandable, and the quality loss is quite difficult to see.

The second image shows a more heavily compressed JPG image, where a significant amount of data has been discarded. At this point it is very noticeable that data loss has occurred, yet a human can still easily recognise this image.

One of the techniques used by lossy image compression algorithms involves reducing the range of colours in an image, known as chroma subsampling. This takes advantage of the fact that the human eye does not notice some colours as much, and in many scenarios the difference in the final image will be hard or impossible to notice. NTSC and PAL, the two main colour encoding systems for old analog televisions, relied on this technique among others.


Is it possible to learn this power?
No
Written and Directed by
GEORGE LUCAS
Is it possible to learn this power?
No

A more extreme example of data loss; the range of colours has been vastly reduced in the first image and even more in the second, leading to the image becoming much harder to identify.

The first image still contains enough data for humans to potentially recognise it, if we know the context. The second has lost so much data that it has become almost impossible to recognise; if we continued to compress this image or discard data from it, we would lose enough data that we could no longer be certain what this image was, even though it might still contain some useful information.

“Data laundering” works in a similar manner; an AI only needs to retain a tiny fraction of the original image data to do its work.

5.1.2 AI metadata

AI metadata refers to the data that an AI retains about a given image. Taking the image above (the unmodified version) for example purposes, a vastly simplified example of the metadata could look like this:

Image colours: 5% white, 10% blue, 60% black, 25% other colours
Text concentrated at: bottom of image
Image separated into: 3 sections
Human shapes in image: 3

An AI will analyse a huge number of attributes like this about an image, but not retain the image itself. On its own, this data is of very little value; we certainly would never be able to guess from the above characteristics that the original image involves Chancellor Palpatine and Anakin Skywalker, or that it was a meme, or much else about it. This ‘metadata’ only becomes useful if there is a large enough amount of it, such as if there are billions of images in the dataset the AI is trained on.

It may seem an almost alien concept to imagine that, simply by having enough data of this kind, an AI could appear to be ‘writing’ essays or ‘drawing’ images. It is doing neither; much like the Ridiculously Old Fraud, it has analysed and collected so much data that it can occasionally seem to mimic an understanding of these topics by accident, despite not having any clue about them.

6 Adversarial examples

One of the primary unsolved weaknesses of AI algorithms is known by the term "adversarial examples", which is one part of the field known as "adversarial machine learning". In simple terms, adversarial examples are similar to how fake news works: AIs, because they do not "understand" the data they are looking at, are highly vulnerable to being fed misleading information.

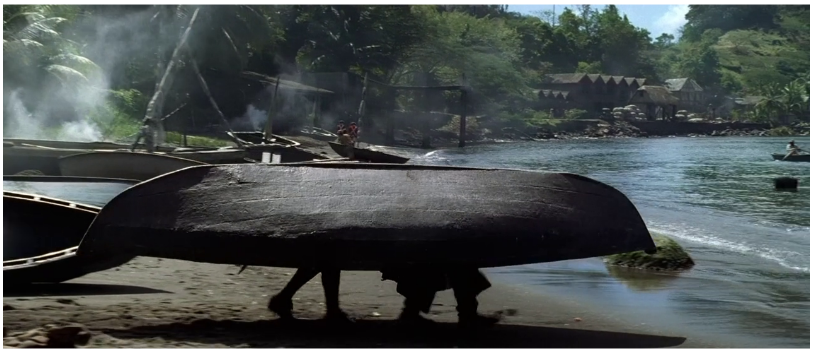
As such, it is possible to introduce information into an image or other data which is clearly insignificant to a human, but that the AI mistakes as being highly important. This allows for two possibilities: firstly, the disrupting of the AI's capability to be trained on a given set of data, by effectively poisoning the data well from which the AI acquires its information. Secondly, adversarial examples can be used to try and 'cloak' or 'shield' a specific piece of data or object from being properly recognised or classified by an AI.

6.1 Adversarial examples in non-AI scenarios

While the term "adversarial examples" refers specifically to their use against AIs, the concept itself is not specific to AI. The aim of any adversarial example is to provide misinformation: fake news is a common example, and in image and vision terms, camouflage is another. The intended result is that the viewer is given an incorrect understanding of a subject, misclassifies it, or fails to identify it.

Predator (1987). Major Dutch (Arnold Schwarzenegger) escapes the Predator hunting him after getting covered in cold mud; it disguises his body heat, preventing the Predator from detecting him.

The extraterrestrial "Predator", which detects its prey with infrared vision, is vulnerable to being fed false visual information. Even though Major Dutch still looks like a human (albeit a very muddy human) in normal vision, in infrared vision his body heat is being temporarily masked by the cold mud on the surface of his skin. In combination with Major Dutch staying still, this causes the Predator to misclassify him as a dead body.

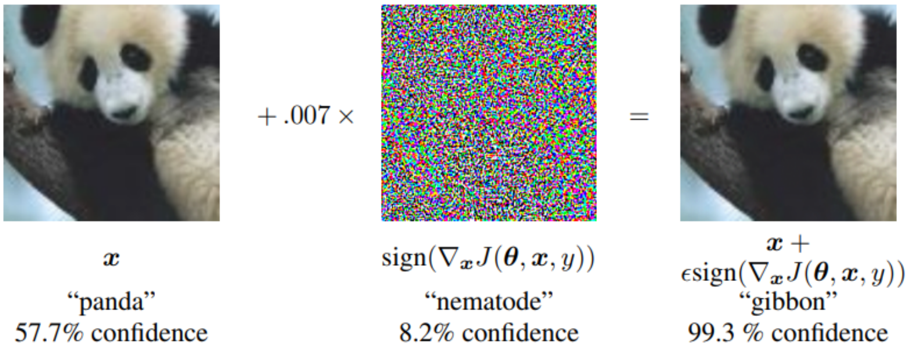


Pirates of the Caribbean, Curse of the Black Pearl (2003). Is that a boat-shaped creature with four legs, or Jack Sparrow and William Turner using a boat to disguise themselves?

At close range, this is very obviously two humans under an upside-down boat (assuming you are human, and know that boats do not have legs). However, if one looked from further away, or from above, at a glance it might simply look like a boat sliding along the shore and into the water; i.e. the viewer might misclassify this object as an empty, upside-down boat.

6.2 Adversarial examples in AI

In AI use, adversarial examples work with exactly the same methods, but are usually less obvious to the human eye. Often they can involve hiding “fuzz” or “noise” in part or all of an image, which to a human does not seem particularly significant or impactful, but that can cause an AI to misclassify an image entirely.


$+ .007 \times$
$=$
$\boldsymbol{x}$
“panda”
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“nematode”
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“gibbon”
99.3 % confidence

Applying noise to cause an AI to misclassify a panda as a gibbon (figure and caption from “Explaining and Harnessing Adversarial Examples” by Goodfellow et al. [35]

Here, the noise is applied to the entire image, confusing the AI’s ability to analyse patterns and causing it to misclassify the animal in the image. There are also other ways that adversarial examples can be used against AI, such as using adversarial clothing designs to fool person detectors [36].

6.3 Data poisoning

AIs all rely on analysing huge amounts of information provided to them by humans. By using adversarial examples to feed “poisoned” data to an AI model, its ability to function as intended can be severely crippled. For instance, feeding poisoned artwork to an art-generating AI can hamper its ability to generate artwork properly, by misleading the AI about the attributes of the image and causing it to gain an incorrect understanding of image attributes in general, poisoning the entire model.

While this approach can be very effective, it does have limitations. Most current AI models are trained on gargantuan data sets; to try and disrupt their ability to function with poisoned data requires either a sufficiently huge poisoned pool of data (and the AI encountering and training upon said data), or targeting specific attributes of the data to poison.

6.4 Targeted vs untargeted adversarial attacks

Most of the examples in this section so far have focused on misleading an AI in an **“untargeted”** way; this means that when we try to make the AI misclassify something, we don’t care exactly how it misclassifies it so long as it does. However, it is often more effective to perform a **“targeted”** attack: one where we attempt to mislead the AI into making a *specific* error.

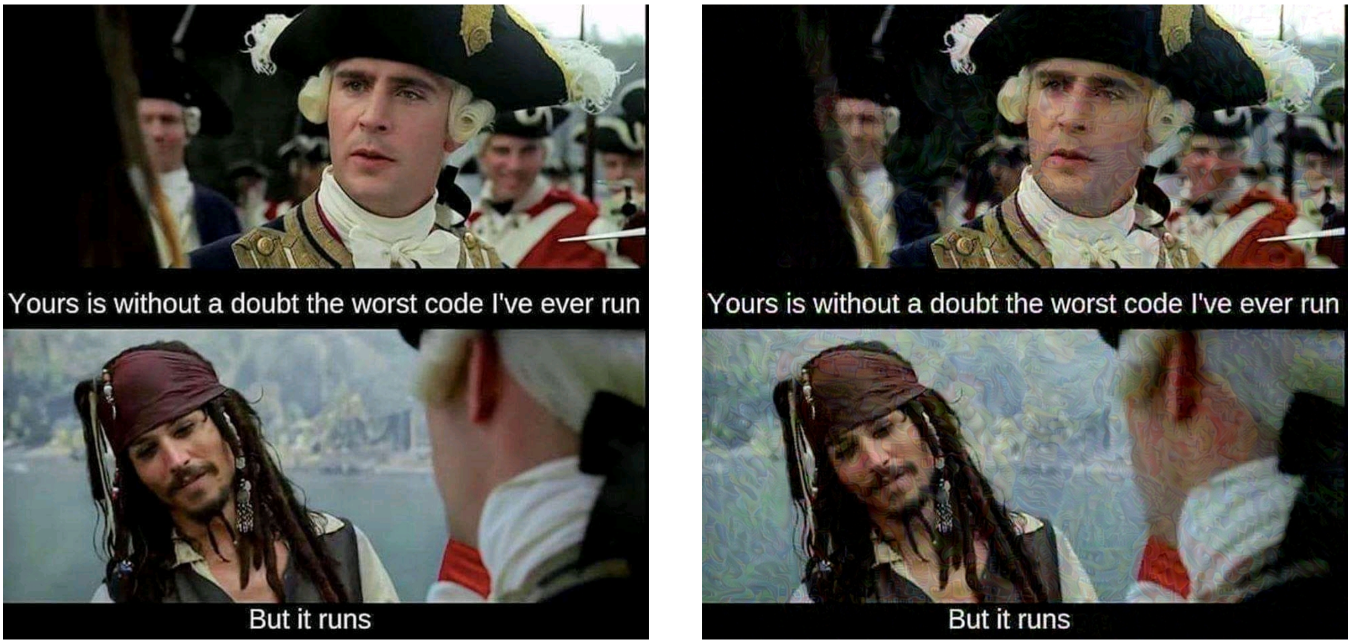
Narrator: he does not, in fact, love democracy.

Sheev Palpatine of Star Wars achieved his goals, in large part, by deceiving his would-be opponents. Successfully disguising himself as a legitimate politician, to hide his true nature as a Sith Lord, gave him the ability to manipulate galactic politics to his advantage. We can consider this something akin to a **targeted** attack; Palpatine needed his enemies to see him exactly as a legitimate politician and not anything else, or he would be unable to conduct his plans. Being misclassified as a "hooded villain" instead of a Sith Lord would not have boded well for his political career.

In AI, a similar incentive to use targeted attacks exists; while they are harder to execute, they are more effective at disrupting AIs in specific ways. Glaze is one example of a targeted attack.

7 Glaze, a tool for protecting images from AI training via adversarial examples

Glaze is a tool for protecting artworks and other images from unethical AI usage, created by experts in the AI and ML field at the University of Chicago [37]. It works by utilising machine learning to calculate a unique watermark of sorts for each image, confusing the AI into misinterpreting the significant attributes of the image.


Yours is without a doubt the worst code I've ever run
But it runs
Yours is without a doubt the worst code I've ever run
But it runs

Un-glazed image on the left, glazed image on the right at the highest intensity (most visible) level; note the 'submerged-in-water' like markings on the right image.

Even at the high level of visible changes to the image shown in this example, using Glaze's highest settings, the meaning and clarity of the image has not changed in any significant way to a human eye. An AI, however, is thrown off by these markings. Glaze's aim is not to simply make the AI confuse the artwork as something else, but to confuse it as a **specific** “something else”; it poisons the artwork in such a way as to make the AI misclassify the artist’s style as a specific ‘target’ style.

7.1 Glaze’s tamper resistance

Creating a tool that can add a protective watermark to an image, and also be extremely difficult for anyone to remove, is an exceptionally difficult task; a static watermark can be easily detected and removed. AI models often embed an invisible watermark into their final result when e.g. ‘generating’ an image, to aid in determining their AI-generated nature, but it has been shown that using adversarial examples can render these watermarks undetectable [38].

In Glaze’s case, however, that approach would not work. The unique watermark Glaze creates for a given image is itself an adversarial addition to the original image; attempting to modify or tamper with that would not cause an AI to be able to process the image normally, as it has already been ‘poisoned’. Someone wanting to overcome Glaze would have to try and effectively reverse the process, which without the original image data is a very difficult proposition. Thus, while no tool is ever perfect in this situation, I think it is fair to conclude that Glaze’s tamper resistance is high.

7.2 Measuring the effectiveness of Glaze

While measuring the effectiveness of Glaze in its application to a given image is possible, measuring its overall effectiveness at preventing unethical AI from utilising stolen works is much more difficult. There are many factors that are not in its control; for example, how many non-Glazed versions of a given image exist online, the actual uptake of Glaze by artists, and so on.

The purpose of Glaze is not to disrupt existing datasets; those have already been used to train unethical AI models, which cannot be “un-trained”. Its purpose is to disrupt the training of future AI models that would use unethical datasets, by effectively poisoning their source data. In that, Glaze has been demonstrated to be very effective [39]. Page 2 of the Glaze paper reports that Glaze has >85% effectiveness at withstanding adaptive countermeasures; in surveys of artists using the tool, 93% rated the protection as successful when all images were glazed, while 87.2% rated it as successful even when only one quarter of images were glazed.

7.3 Making Glaze more accessible

A common question asked by creatives about Glaze is whether there could be a web interface or API to more easily use it, or integrating it with existing applications such as Adobe Photoshop.

Unfortunately, while this would be very convenient and make it easier for creatives to include Glaze in their workflow, this can’t be done without compromising trust in the application itself. This is because once a tool, e.g Glaze, has been integrated into another service such as Photoshop or made into a web API, there is no guarantee that it is, in fact, actually Glaze being used; it is an unsolvable trust problem.

Consider going into a shop, and handing your digital artwork over to have Glaze applied to it. The shop assistant goes into the back, then comes out with your glazed artwork. You have no guarantee that the shop assistant actually applied Glaze and not something that looks like it, and did not keep a copy of your original un-glazed artwork.

The same problem applies here. Once you have transmitted your original image file to an external server that provides an API, you cannot trust or be certain that it has not retained that file, or verify that it has run the software you wanted it to run (and not, for example, a maliciously modified version).

It is important to note that the Glaze team is (or was, at the time of writing) considering making a web API for Glaze funded by the University of Chicago. Such an API would represent the sole exception; it would be possible to trust it in this one scenario, as there would be no motive for misuse of data and the accountability of a public university, and since there would also therefore be no opportunity for the code to have been maliciously modified.

7.4 Limitations

Glaze, by the nature of any tool, has limitations. It cannot change the past; models trained on existing data will not be changed by applying Glaze to new or existing artworks.

In addition, poisoning data of any type generally involves some level of collateral damage: in this case, Glaze’s watermark is generally not a desirable visual addition to the artwork, reducing its appeal to a varying extent depending on the properties of the artwork itself. The fact that Glaze needs to be run on its own, for the reasons outlined above, makes it an extra step in an artist’s workflow.

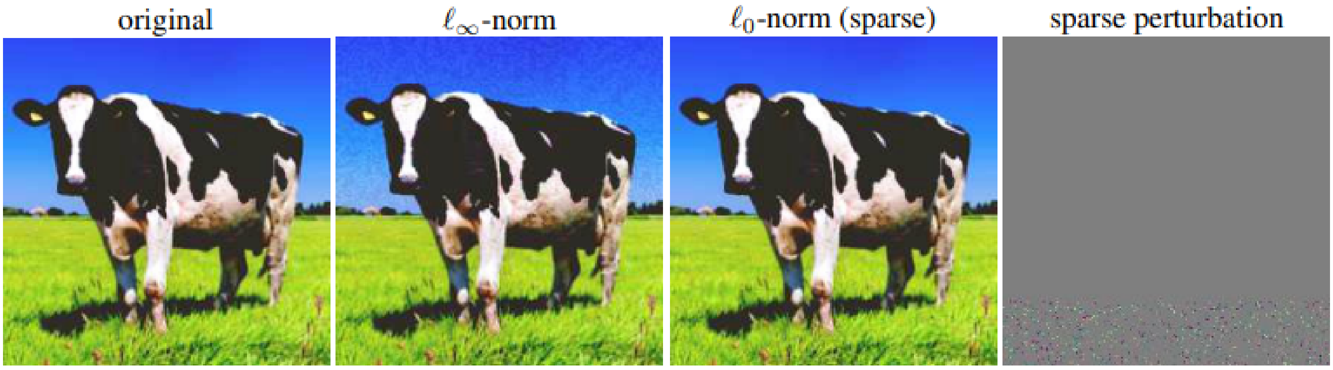
In addition, Glaze requires large amounts of memory, and a fairly decent graphics card, to run. This makes it inaccessible or impractical for those with particularly old computers, but will likely improve in future.

7.5 Future improvements

Glaze is still in a very early stage as an app; as I understand it, the Glaze team's initial priority was getting a working version out as soon as possible, rather than optimising everything. This means there is a lot that may likely improve in future versions, making it even better to use.

Firstly, it's possible the app could be optimised to use less memory and require a less powerful graphics card in future, allowing it to be more accessible to those with older hardware. With the core functions finished, it may also be possible for a more streamlined interface to be made, so that it becomes even easier to use and navigate.

Secondly, there is a possibility that Glaze could have a feature to only add adversarial noise to specific areas of images, instead of to the whole image; this could work by hiding the adversarial 'fuzz' within 'busy' areas of the image, where it will be less visually obvious to a human.



A figure and explanation from the paper *"Are adversarial examples inevitable?"* by Ali Shafahi et al, demonstrating adding adversarial noise in a busy area of an image, in this case in the grass. [40]

This would reduce both the processing time required to apply Glaze to an image, and also reduce its visible impact on the image itself, making it more appealing to use. The Glaze team have mentioned other improvements they are working on, such as preserving image metadata and improving its use on comic-style art, on their Twitter page.

8 Proposals for stopping unethical AI

Unethical AI is a very difficult problem to solve. This is due to the aforementioned problem of data laundering; much like money laundering, the trouble lies in proving that the data in question was in fact stolen or not permitted to be used. In this section, I propose what I believe will be necessary to prevent unethical AI from existing in future.

8.1 Legal regulation and compliance requirements

Many of the currently proposed legal solutions against unethical AI use focus on making the concept itself illegal. While that's fine in itself, this doesn't fix the problem, because the main draw of unethical AI - data laundering - makes it very difficult to prove said illegality is actually occurring.

One part of the solution for curbing unethical AI use is to implement similar data security compliance requirements as exist in some other areas. Finance, for example, has the PCI-DSS standard; while this standard is not law, it is in practice considered very much similar to one, as all of the major card providers require its use. Larger companies processing financial information, and those holding sensitive information (such as payment processors like PayPal and Stripe) are required to undergo independent, external auditing to prove that their data security and handling is secure. In a similar vein, the 1996 HIPAA US law sets requirements on how health data is handled, with severe penalties when it is not handled properly - including criminal penalties in some cases.

Companies intending to train AI models should be legally required to undergo regular, independent external auditing. This would need to be alongside legal requirements for said companies to hold full archive records of exactly what data they trained their models on for a minimum period, at least several years. This is not expensive with modern data solutions, where storing 1024 terabytes of data (1PB) in long-term archival storage costs $1,000 a month. This is small compared to the much higher cost of training an AI model [41] [42], which is set to become more expensive over time.

8.2 Heavy market regulation

AI should not be a free market under any circumstances. The reasoning for this is simple: if any company or startup can jump into the AI market, it will be impossible to enforce laws that are enacted to prevent data laundering. Whilst businesses often love to talk about what they perceive as the benefits of a free market, there are many markets in which a "free market" is objectively bad for consumers. To illustrate, let us examine the way governments regulate markets for controlled drugs.

In the US and many other countries, many drugs are controlled to differing degrees, restricting who is allowed to manufacture and supply them. Dangerous drugs without medical uses have the highest control, and in the UK for instance, such drugs require specific government approval [43]. Drugs with medical uses that have significant risks, such as the ADHD medications Ritalin (methylphenidate) and Adderall (amphetamines) have similar controls, allowing their prescription but involving lengthy documentation, audit trails and licensing.

These laws limit which companies can manufacture controlled drugs, the result being that the companies who produce them tend to be large, established entities. Due to their size and the limited number of them, it is far more plausible to audit those companies and enforce rules on them; this means less manufacturing defects, "leaked" drugs outside of the controlled system, and fewer rule breaches overall. Given the nature of controlled drugs, the reduced competition and market freedom is a price well worth paying for a safer manufacturing ecosystem.

AI is exactly the same. Limiting AI training and development to specifically licensed companies, that have to go through rigorous auditing and inspection processes, will allow governments to properly enforce laws that require AI companies to retain their training data; this would knock opportunistic, data laundering AI startups and private VC-backed unicorns out of the market. The result would likely be that the biggest 5-20 companies or so would dominate the sector, and yet this would be a desirable outcome: those 5-20 companies can be regulated, audited, and kept under intense political and public pressure, whereas trying to regulate an endless number of small AI companies becomes an endless game of whack-a-mole.

This is not to suggest that such a system is perfect. It can invite problems such as abuse of monopoly power and political difficulty in regulating large, economically important companies; these problems can be managed in a carefully regulated market. It is still a much better system than the alternative where little or no regulation is realistically enforceable.

8.3 Outlawing the use of models using non-compliant datasets

Some AIs, such as Stable Diffusion and ChatGPT, are trained on datasets scraped from the public internet. These datasets contain copyrighted data and other data that should not, by any sane definition of 'fair use', be allowed to be used for copyright infringement. Here is some of the research demonstrating this.

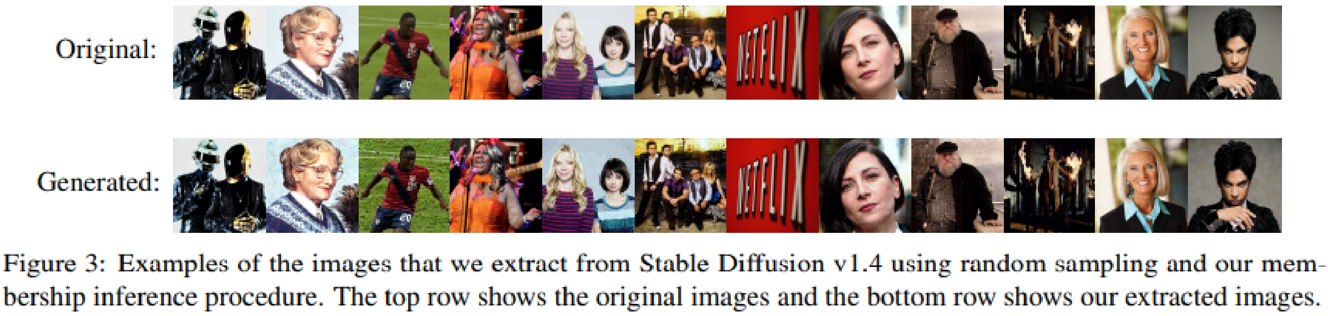


Figure 3: Examples of the images that we extract from Stable Diffusion v1.4 using random sampling and our membership inference procedure. The top row shows the original images and the bottom row shows our extracted images.

A figure and explanation from page 5 of the paper *"Extracting Training Data from Diffusion Models"* by Nicholas Carlini et al. [44]

In addition, recent research from MIT and Harvard has demonstrated that Stable Diffusion objectively succeeds at copying the styles of artists [45]. Not outlawing the use of such models would create a race against time: even if unethical data use was banned, sufficiently powerful models could still be used to enable all manner of copyright infringement.

9 Temporary solutions

Technological solutions and PR solutions are unfortunately only capable of being temporary stopgaps. Not all forms of data can be poisoned by anti-AI techniques, and PR solutions only work when a company is particularly well-known or directly ‘connected’ to its end consumers, something which is frequently not the case in big tech companies.

9.1 Technological solutions

In some scenarios, technological solutions to the problem of unethical AI are not feasible. There is no way to avoid an unethical bot scraping data from a website, other than not hosting the data on that website; the only effective technical solution is to poison the data in some way.

For some forms of data, this may not be practical; the “poison” must be sufficiently strong to substantially affect the AI trained upon the poisoned data, it must be undetectable by the AI (or else it would be capable of removing or avoiding the poison), and to be used enough to make a difference, it must not cause an unacceptable drop in the quality of the data.

Adversarial attacks on text data are possible, but extremely challenging. Here is an example from an academic paper:

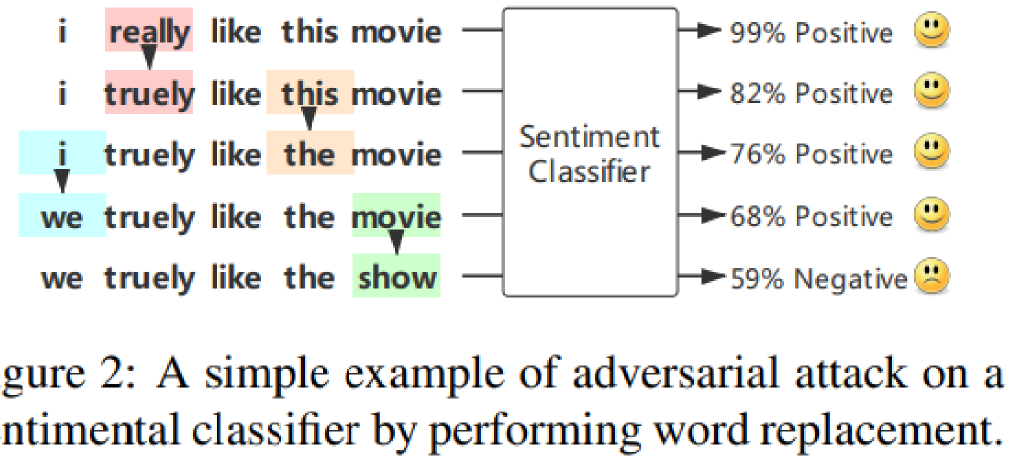


Figure 2: A simple example of adversarial attack on a sentimental classifier by performing word replacement.

A figure and explanation from the paper *“Generating Fluent Adversarial Examples for Natural Languages”* by Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. [46]

As we can see, the resulting “poisoned” data - and the meaning of the sentence itself - has been changed significantly, in a manner far more noticeable to a human than e.g. Glaze’s image markings. The sentence has shifted to being more abstract, using ‘the’ instead of the more specific and directed ‘this’, and changing from the specific format ‘movie’ to the more generic ‘show’. These issues in natural languages make it challenging to poison text data in a way that is still acceptable to read or publish, but is also sufficiently strong to affect an AI.

9.2 PR solutions

Companies - at least those not willing to risk the possibility of legal issues by hiding their AI use or otherwise obscuring it - can sometimes be forced to abandon unethical uses of AI if their customers and public opinion are against it. Celsys, the developers of Clip Studio Paint art software, were effectively forced to reverse their proposed use of Stable Diffusion as an added feature after artists using their software overwhelmingly gave negative responses [47] [48].

The ability of public opinion to influence this, on its own, is limited in many cases. Many companies do not directly sell to individuals or end users; for example, infrastructure companies like Google and Amazon often deal with the data of large companies, whose customers may not be thinking about or aware of what third parties do with their data.

10 Non-solutions: the risks of bad or ineffective solutions

One of the biggest threats to stopping unethical AI is not just making sure it is properly regulated, but as part of that, ensuring that non-solutions and ineffective solutions are not implemented. This tends to fall into two categories: malicious non-solutions proposed by those wishing to see AI poorly regulated, and unintentional non-solutions proposed by those wanting to see AI properly regulated, but not fully understanding the implications of the solution they are proposing.

To explain the damage both types can do (even though the unintentional sort is of course very different in morality from the intentional sort) and where the risks lie, let us examine an ever-relevant political concept.

10.1 The law of inverse relevance

In the fictional political sitcom Yes Minister, Sir Arnold Robinson describes a very relevant political concept: what he describes as the **"law of inverse relevance"**.

"It's the law of inverse relevance: the less you intend to do about something, the more you have to keep talking about it."

Yes Minister (1980). Sir Arnold explains the law of inverse relevance to Bernard.

This common political strategy can be seen everywhere; AI companies crying out for regulations are utilising it. In proposing weak regulations that will not effectively regulate their activities, AI companies hope to divert public and lawmaker attention from genuine, effective regulation, and weaken public and regulatory pressure on the ethics of their activities. While those unintentionally proposing weak solutions are not trying to do this, the result is the same: when a weak solution is implemented, public pressure to implement a solution disappears, because the public believes it is already implemented.

As Sir Arnold also astutely said in another episode of Yes Minister, *"doing the wrong thing is worse than doing nothing"*. While this statement is not universally true, it very frequently is; AI is an extremely good example of a space where this assertion holds. Ineffective solutions don't just fail to achieve any benefit; they divert attention and public pressure from effective solutions, and give the false sense of the original problem being solved.

10.2 AI detectors

AI detectors are shady on many levels, but in the context of non-solutions, the largest problem they present is that they are severely unreliable and prone to error, including an unacceptably high rate of false positives. Many claim that their scores are only to be used as an 'indication' of whether content is AI-generated and should not be relied upon as conclusive proof, yet that is inevitably what is happening. Their disclaimer is much more likely to be there to avoid liability than in any genuine belief that users will not rely on their detectors' outputs.

In April 2023, the Washington Post tested Turnitin, a company that sells plagiarism detection services including AI detection. On a test of sixteen written submissions containing AI material, original material and a mix of both, Turnitin got 8 out of the 16 texts at least partially wrong [49]. In addition, it flagged an entirely original essay written by one student as being written with the help of AI.

In January 2023, OpenAI launched a classifier to distinguish between human-written and AI-generated text. They admitted that the accuracy rate of detecting AI-generated content was 26%, and that human-written content was falsely flagged as AI-generated in 9% of cases [50].

Recent research has demonstrated the ease of fooling AI text detectors, demonstrating a major reduction in accuracy using simple paraphrasing techniques [51]. Most AI detectors boast high accuracy scores, without any meaningful evidence to back up their claims or make any verification of them, with tests not done by the providers of said tools almost invariably finding far lower accuracy rates.

10.3 Subjective human review

It is inevitably the case that when something is suspected of being generated by an AI, in a situation where that is prohibited, there needs to be some manner of evaluating whether or not that suspicion is true.

Suppose, for example, that a prize-winning artwork or published illustration is suspected of being generated by AI. It is perfectly proper for a human, or several humans, to review the artwork in question to establish the probability of AI generation being used, but this review must be **objective** and be made based on **demonstrable, published evidence**. The aim of these requirements is not to eliminate the risk of false positives, but to mitigate that risk to the maximum extent possible.

Compounding the risk of this problem is the inability to verify when it happens; false positives in these situations are unreported. If a group of expert artists conclude an artwork was likely generated by AI, there is no plausible way to verify the correctness of that conclusion: the rate of false positives will be implausible to measure. The same problem applies to other areas such as when essays are accused of being AI-generated. In addition, many human reviewers turn to AI detection tools either as additional proof or as sole proof of AI use, further increasing the rate of false positives.

It is therefore imperative that when reviewing a work that has commercial or other significance (as opposed to e.g. an ordinary submission to an art website) to establish the likelihood of it being made with AI tools, that the following principles are upheld:

- There is a heavy leaning towards the presumption of innocence
- The contents of the review, including the evidence used to make the determination, is made publicly available to enable proper oversight
- Full objectivity is required, with no reliance on “expert’s instinct” or similar subjective measures which are prone to Dunning-Kruger overconfidence
- AI detectors are not utilised, as their scores are inevitably subject to human confirmation bias irrespective of their accuracy, further reducing objectivity

For more typical content, like screening a fanfiction submission or a normal artwork upload, a more reactive approach can be taken. AI detectors should still be avoided, at the very least until they are subject to far heavier regulation than they currently are.

10.4 The damage inflicted by false positives

The mental, physical and sometimes economic damage done to individuals by being falsely accused cannot be overstated. In legal systems in various jurisdictions, Blackstone’s ratio, noted by William Blackstone in the 1760s, is a common way of expressing the legal view that the law should err on the side of finding people innocent of crimes to minimise the conviction of innocents. The ratio states:

> “It is better that ten guilty persons escape than that one innocent suffer.”
>
> - William Blackstone, in the *Commentaries on the Laws of England*

The dangers of improper review, or relying on AI detectors in any way, are not theoretical. In December 2022, the artist Ben Moran was banned from a Reddit community with 22 million members after being falsely accused of using AI to create a commissioned artwork [52] [53].

In March 2023, The University of California accused a student, William Quarterman, of generating his submitted work with AI. The university required him to speak before the university's honor court, causing him to suffer full-blown panic attacks [54]. The university later cleared him of the accusation, stating they did not believe him to have used AI and had no reliable evidence to the contrary.

A unique problem with AI detection is that while an artist, writer or other expert has expertise in their fields, they are not typically technical experts, and are thus less well placed to distinguish human content or processes from AI-generated ones than they might believe. This creates an illusion of overconfidence in the ability to detect when a work is AI-generated, and the lack of any ability to verify if their conclusion was correct reinforces this bias; if they were wrong, they will not know it, giving the illusion that they are much more accurate at this detection than they actually are.

10.5 Ineffective regulations

Where the regulation of AI is concerned, there are relatively few regulatory approaches that have any hope of being effective if certain conditions are not met. Specifically, if the issue of data laundering is not addressed and companies are not forced to record the data they train their models on, enforcing any other legal provisions will be effectively impossible.

One area of this that needs to be considered is proactive versus reactive laws. This isn't referring to the enacting of said laws, but who is responsible for enforcing them; requiring AI companies to be proactively audited is far more effective than only checking a company's legal obligations if an entity files a legal complaint. If a regulation requires that an individual makes a legal complaint, the AI company can swamp the individual in legal delays and costs, effectively forcing them to either drop their complaint or have significant financial backing. This is a common legal problem, and the lawsuits that categorise it are known as **SLAPPs**.

10.5.1 SLAPPs

SLAPPs, or intimidation lawsuits, are a type of lawsuit where the aim is to intimidate and silence the defendant. They work by forcing the defendant to either abandon their claims, or face the heavy costs of mounting a legal defence, which many individual defendants will not be able to afford. The plaintiff, or accuser, typically doesn't expect to actually win the lawsuit if it proceeds to court: this kind of lawsuit is a strategic abuse of money and power.

Libel law is a prominent example of this problem. Around the world, libel law is often viewed by the legal profession as the archetypical "rich man's law": it is easily used to suppress those without the financial means to mount a defence and is thus often settled out of court, and frequently does not come down to which side was actually in the right [55] [56] [57].

The same problem can and will occur in AI if weak regulations against unethical AI are passed; the ability to sue an AI company for data theft would be of very little consequence to individuals, who typically cannot hope to endure the stress and financial burden of defending themselves against vexacious retaliatory litigation made by the companies in question. LAION's recent legal threats against photographer Robert Kneschke are an example of this behaviour [24]. Only proactive legislation, where the AI company is required to prove it is in compliance rather than others being made to prove it is not, can be effective in this situation.

11 Determining which companies can be trusted with AI data

To properly explain this, we must first define two "types" of companies for clarity:

"Web Services Providers" are those who provide a massive array of different web services. Some of the major players in this field include Microsoft Azure, Google Cloud Platform, Amazon Web Services, DigitalOcean, and Oracle Cloud.

"AI Companies" are those which provide solely or primarily AI services; most are very new companies compared to other tech companies. Examples include OpenAI, Stability AI, Hive, and Scale AI.

The aim of this section is to provide guidelines on how to assess a company's likelihood of using AI to launder data. In general, there is a clear correlation: Web Services Providers can generally be trusted with data put into their paid non-AI web services, whilst primarily or solely AI companies cannot be trusted. This is not because of a difference in ethics per se, but due to their different business models and the risk/reward of laundering data.

It goes without saying that this is not an entirely black and white matter; that a company falls into one bracket or another does not inherently render it trustworthy or untrustworthy. While there is no way to "determine with certainty" to decide whether a given company can be trusted or not, there are ways to reasonably estimate it. There are several factors we can examine which give us a good insight into a given company's motives and likely intentions.

11.1 Company age and breadth of operations

In the digital world, cryptocurrencies and NFTs have presented an opportunity for scammers and fraudsters to perform "rugpulls": building up a product or financial store, then taking the money and running, effectively "pulling the rug out" from under those who invested in it [58] [59] [60].

A similar problem is likely to happen in AI: companies that started comparatively recently (e.g less than 5-10 years ago) and only have significant business in AI have a large incentive to misuse customer data. Under current legal rules they are not required to retain the data they trained their models on, making it impossible to ascertain if they stole user data, and since they do not have operations in other sectors, they have less to lose.

Companies that have been established for longer, and that have many different operations, are in the opposite situation: unlike smaller or more recently formed companies, they have a huge amount of reputational capital to protect and larger revenues to lose. Being discovered misusing customer data would bring them colossal damage in PR, legal and financial terms.

11.2 Size of company customers, and type of data said customers store

It is the business of most web services providers to host all manner of data and websites; many have dedicated services for storing government information, storing data in compliance with HIPAA, confidential information under PCI-DSS rules, and other requirements.

In addition, such companies often count huge, multinational corporate giants among their main customers. To take a few examples:

- Microsoft Azure is used by Shell, AT&T, Autodesk and eBay [61]
- Amazon Web Services is used by Adobe, Netflix, Samsung and Unilever [62]
- Google Cloud Platform is used by PayPal, Twitter, Procter & Gamble and Etsy [63]

This is significant because companies of this size have the legal and financial muscle to give even huge web services providers a whole host of problems if their sensitive or commercial data was to be misused, used without consent or leaked in any sense.

Web services providers have nailed their colours to the mast in this regard; the slightest suspicion of their customers' data, that they have paid to store securely, would have gargantuan consequences. In addition, the larger providers are so large that being caught in such actions would inevitably cause political disquiet and calls for more stringent regulation or the breakup of the larger players, which is the last thing they want to happen.

A company training AI models on its own file storage service would also not be an appealing proposition for that company. They will have little specific information about what those files contain; an image file could be a tiny company logo, a human patient's HIPAA-protected medical results, or any number of other things. The risk of using legally protected data, and the likely result that the resulting model would be contaminated with huge amounts of irrelevant and unhelpful data, makes it an unlikely scenario. In addition, the agreements for these file storage services generally make it clear said data will not be used for AI or other non-customer-requested purposes, and take responsibility for handling said data (see the Shared Responsibility Model subsection below).

Note that in the context of web services providers, this refers specifically to data stored for non-AI purposes, such as in file storage services like Amazon S3 or Microsoft Azure Files. Data given to AI services within web services providers, such as Amazon Rekognition or Microsoft Azure AI, should generally be assumed to be used for training company models. Amazon Rekognition explicitly says so and has an opt-out option [64]; Azure Face API says images are "automatically deleted after processing" [65], a statement that does not give any useful information where AI training is concerned. If it were to turn out that it is training its models on customer data, the image file itself is not needed after processing; only the resulting training data is used by an AI.

11.3 Number of company employees and size of operations

This is directly to do with the likelihood of whistleblowing occurring. In smaller, newer companies with fewer employees, this is less likely: it is harder for a whistleblower to remain anonymous if less employees had access to the information they share.

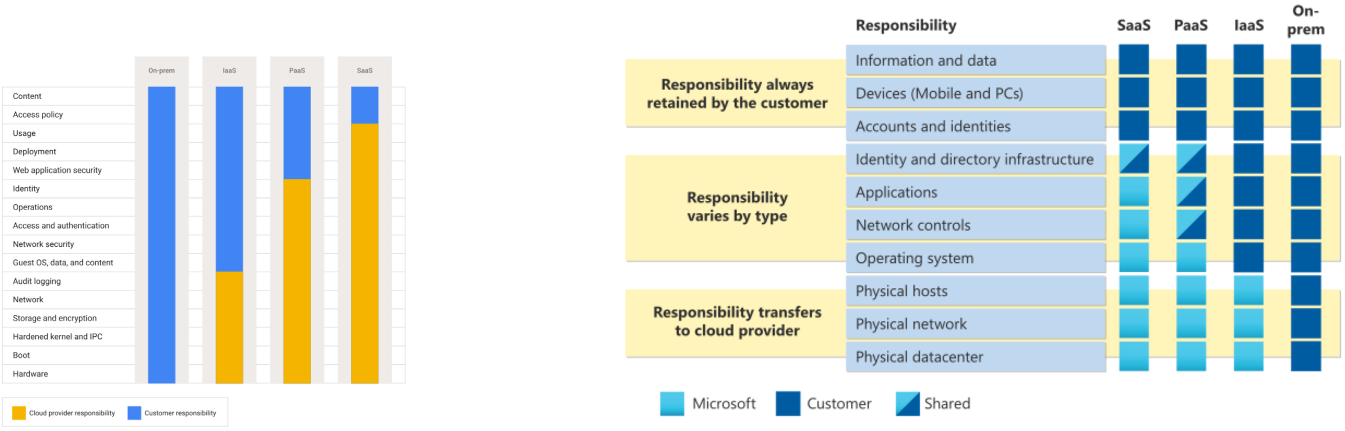
In a large company, it is harder to discern exactly who blew the whistle on a given subject: in addition, some large companies recognise that being caught suppressing whistleblowing can be a bigger business risk than allowing it to happen and having proper policies for it in place. Trying to keep unethical use of customer data for AI a secret would be far harder in a large company than a small one.

11.4 Terms of Service and other legal liabilities declared by the company

Probably the most obvious and telling difference between trustworthy companies and untrustworthy ones, in the context of data, is the TOS (Terms of Service) by which they declare their responsibilities for said data.

Web services providers, because their main business is holding customer data and depends on customers trusting them to do so, generally take responsibility for their part of the job that is keeping customer data secure by way of the "shared responsibility model" [66] [67] [68].

11.4.1 Shared responsibility model



	Responsibility	SaaS	PaaS	IaaS	On-prem
Responsibility always retained by the customer	Information and data	■	■	■	■
	Devices (Mobile and PCs)	■	■	■	■
	Accounts and identities	■	■	■	■
Responsibility varies by type	Identity and directory infrastructure	◩	◩	■	■
	Applications	▢	◩	■	■
	Network controls	▢	◩	■	■
	Operating system	▢	▢	■	■
Responsibility transfers to cloud provider	Physical hosts	▢	▢	▢	■
	Physical network	▢	▢	▢	■
	Physical datacenter	▢	▢	▢	■

▢ Microsoft ■ Customer ◩ Shared

The "shared responsibility" model used by major web services providers. Examples here are from Google and Microsoft respectively.

The basic premise of the "shared responsibility" model is that the provider is responsible for making sure their tools and systems are secure, and the customer is responsible for using the tools and systems properly so their data remains secure.

To explain a little more, web services are very complex products to use - they are not aimed at beginners or non-tech people, but rather at experienced engineers. Tasks like 'managing access to data for your web resources" are extremely complicated, and mistakes like accidentally giving the wrong permissions for data access is a common pitfall. Many organisations have entire teams or specialist employees dedicated to such tasks for this exact reason - mistakes can be very problematic, to say the least.

The shared responsibility model, then, means that the customer's job is to know how to use the services properly; i.e. not exposing their own data to the wrong people by accident. The provider takes responsibility for making sure the services do exactly what they're supposed to do; if a customer had a correct data policy that only allowed access to specific people, and that data was still accessible by people it wasn't meant to be accessible to, that would be the provider's fault.

This is in contrast to AI companies, who often take no responsibility for data, using legal workarounds often used in similar ways by businesses in other sectors.

11.5 Legal workarounds used by AI companies

11.5.1 Declaring legal responsibility on the user

The first workaround sometimes seen is one where the user is declared responsible for all output generated by a given system. An AI company can use this to avoid being held responsible for copyright theft enabled by their systems, for instance.

Such a clause, if it only disclaimed specific things such as offensive or illegal output, would have a legitimate purpose - it would be easy for malicious actor to use a company's AI system to deliberately generate offensive or illegal content, and it is difficult to reasonably prevent that from happening.

11.5.2 Declaring use of data for improvement purposes

Various companies state in their Terms of Service that your data may be used for purposes such as "improving their services" or "improving existing products". While in many non-AI cases this is a perfectly innocent statement and isn't anything to worry about, an AI company is effectively telling you that any data you provide may be used to train its models.

There is an important note about solely AI companies, that do not have other services, in this context. When such a company promises not to use customer data (or any other data given to it) for AI training, such a statement has no real value; there is no reliable way to monitor or enforce it, due to the data laundering problem explained earlier in this article. As a result, it should be assumed that any data given to such companies will be used for any purpose the company chooses, such as training AI models.

11.5.3 Delegating legal responsibility to huge numbers of subsidiaries

A workaround utilised by some companies, particularly those offering AI microwork services, is to declare themselves a 'passive' third party (ie they are not actually performing the task you pay them for), and declare their third party subcontractors or workers who perform the task legally responsible for the handling of data. This approach is most common where they are huge numbers of workers or subcontracted entities; such small entities are unlikely to have the resources to insist on alternative arrangements, and the company benefits by being effectively impossible to sue when data is mishandled, leading to a large incentive to misuse it.

This effectively allows a company to state in its Terms of Service that your data will not be used for X or Y purposes; since said terms declare the third party subsidiaries responsible for data anyway, such a statement has no real significance, but acts to lull those who do not read the fine print extremely carefully into a false sense of security. A company offering both microwork services and other AI services, such as Hive.ai, has an incentive to exploit this; by shifting the legal liability onto the microworkers, the company can use customer data to acquire training data for its models via its microworkers, and simply throw a given microworker under the bus if a breach is discovered.

12 Appendix A: Examples of good uses of AI

It is important to note that many applications of AI are entirely ethical and beneficial. It is very unfortunate that, despite this, the reputation of AI and ML is being badly damaged by unethical developers seeking to profit from large-scale data laundering.

This section lists some examples of good uses of AI. All of these uses have a common theme: AI can be useful for finding patterns in data that humans may struggle to find.

12.1 Agriculture

The use of AI in agriculture can improve crop yields and the ability of farmers to manage farmland efficiently [69] [70]. This can include better detection of pests and diseases, crop maturity, soil conditions, weed management, and many other variables that can be difficult to measure accurately.

12.2 Some areas of medicine

AI has been found to rival radiologists in screening X-rays in some situations [71]. It can be used to aid clinicians in making more accurate diagnoses by providing a second opinion, and by reducing the burden on specialist clinicians in less complicated cases.

It is also used in the discovery and development of new drugs [72] [73], and to assist humans in predicting the toxicity and bioactivity of new drugs.

Analysing X-ray images or other test results for patterns is another area AI is good at; it can find very subtle patterns in data that humans may struggle to find. Recent research has shown encouraging results for using AI to detect lung cancer where it outperformed radiologists [74]; while it could not replace radiologists, its use could help reduce the number of incorrect diagnoses and allow radiologists to treat more patients overall.

LLMs have been utilised to translate brain activity from fMRI scans into text, which could be used for helping patients otherwise unable to communicate [75]. Understandably, technologies that enable the “reading of minds”, so to speak, often concern the public in terms of the risks of their misuse. To be clear, the authors of the research are clearly aware of and alert to this risk, and the last paragraph of the ‘Discussion’ section of their paper is quoted below.

> “Finally, our privacy analysis suggests that subject cooperation is currently required both to train and to apply the decoder. However, future developments might enable decoders to bypass these requirements. Moreover, even if decoder predictions are inaccurate without subject cooperation, they could be intentionally misinterpreted for malicious purposes. For these and other unforeseen reasons, it is critical to raise awareness of the risks of brain decoding technology and enact policies that protect each person’s mental privacy.”
>
> - Jerry Tang, Amanda LeBel, Shailee Jain & Alexander G. Huth, *Semantic reconstruction of continuous language from non-invasive brain recordings* [75]

12.3 Financial fraud detection and prevention

The primary use of AI in finance is for fraud detection [76]. Much like the use of AI to find patterns in medical images, AI and ML algorithms are excellent at finding patterns in financial data that humans might miss, allowing it to improve detection and prevention rates for fraud and other bad activity.

12.4 Cybersecurity

AI can be used to help mitigate the effects of Distibuted-Denial-of-Service (DDoS) attacks, by detecting patterns in incoming network traffic [77] [78].

12.5 Specific image recognition applications

Modern social media platforms, by their size and scale, have absolutely massive amounts of new user data being uploaded to them. Given that various countries have been proposing rules limiting how long a social media provider should have to remove problematic material, and the need to make safer websites online, doing this purely via human moderation is not plausible both in monetary and time terms.

Using AI to identify problematic material, such as adult content where it’s not supposed to be or hate speech, can reduce the workload on human moderators [79]. It is not designed to, and should not, replace human moderators - its purpose is to reduce the amount of content that human moderators need to review. This has an additional benefit to human moderators, who often suffer mental health issues as a result of being exposed to extreme and disturbing online content as part of their role [80].

13 Appendix B: Examples of bad uses of AI

Bad uses of AI have one fundamental flaw in common: they attempt to solve problems in which the data required to do so is so gargantuan, or so difficult to obtain, as to be effectively unsolvable.

In addition, any use of AI - or in fact any other technology at all - which attempts to be “decentralised” and not require human supervision, comes into this category. Cryptocurrencies and NFTs are examples of technology that were never viable in any form for this reason, among others.

13.1 Image generation and content generation

Generating images - or any other form of content - in a manner that doesn’t simply steal from existing works, requires a level of data that cannot reasonably be obtained for the foreseeable future. This is explained in the subsection Illustrating the weakness of AI: art earlier.

The current “popular” AI applications are not intelligent; they are statistical amalgamations of relationships computed from their training data. Their sudden leap into the public eye is almost entirely the result of their use of much larger data sets than in previous times due to unethical data scraping, as explained in earlier sections of this paper.

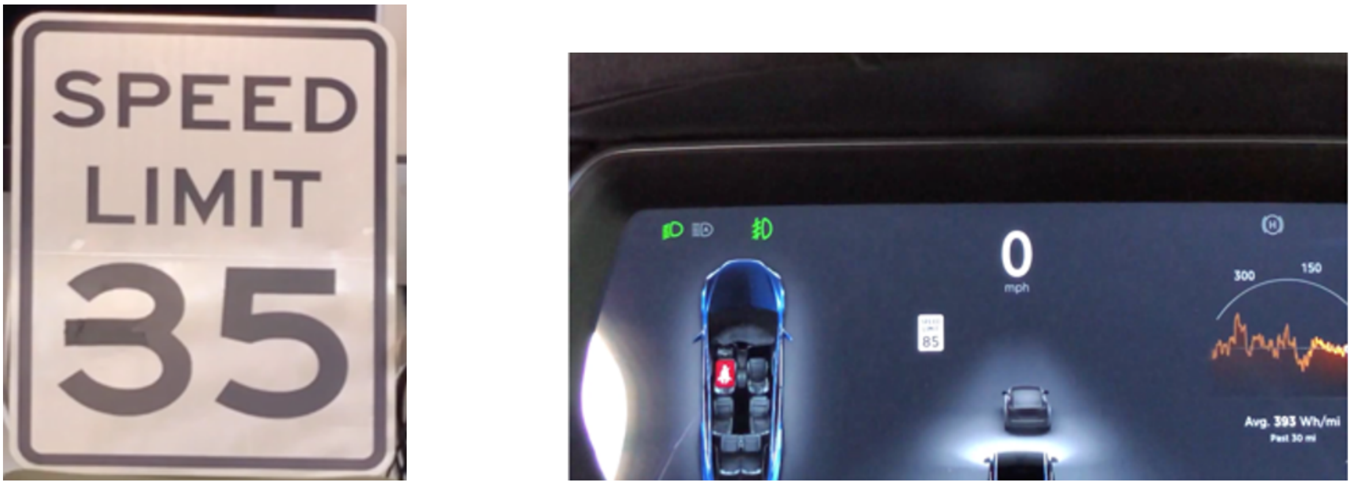
13.2 Self-driving cars

Unlike more restricted forms of transportation, like trains that run on tracks, there is no physical restriction on where a car can move in 2D space. In addition, cars are frequently in very close promixity to pedestrians, making the error margin for incorrect driving very small. This makes the dangers of a car that drives incorrectly a lot more dangerous.

AI, by the limited nature of the data it has available to it, is effectively incapable of making error-free safe decisions on roads, making self-driving cars a dangerous concept in many real life use cases. Humans, who are infamously very far from perfect when it comes to driving, must adjust to a massive variety of different circumstances that may be unreliable or unpredictable. Visibility, weather, traffic, local events, terrain, road maintenance condition, area type and population density (urban, suburban, rural, etc), and many other factors are constantly in consideration.

Unlike some other applications of AI, where the AI can be used to aid human decisions, this is not possible in driving. In situations where an AI fails to recognise a threat, the human will likely have a fraction of a second to react. A human who is assuming the AI is driving will not only have to react to the danger, but also the fact that the AI is unable to deal with the danger. This is next to impossible for any human to do; the same issue applies in any situation where the maximum safe “reaction time” is sufficiently small.

In research performed by McAfee in 2019, it was found that adding a black sticker to a speed limit sign was sufficient to cause a Tesla Model X to misinterpret the speed limit on a road.



Picture of a Tesla Model X’s HUD, showing that the simple black sticker, clearly readable by a human as 35 mph, caused the car to misclassify the speed limit as 85 mph - a much higher speed. [81]

While such research is often more focused on academic possibilities than current dangers, the illustrated problem is simple: if this were to happen in real life, there would not be enough time for a human to react in many cases. Additionally, while a human can mistake a number on a road sign, this is often specific to that human and their current state of health; if a hundred humans drive past the sign, a few may misclassify it in a bad situation, but a hundred AI-driven cars will have a far higher number misclassifying it as they do not differ in the same way the humans do.

13.3 Many applications of anthropomorphic (human-like) robotics

It is something of a running gag, not only in the technology field but in popular culture, that highly-hyped human-like robots end up failing in spectacular fashion every time they are presented, no matter how much showmanship and fanfare is given at their presentation. Such robots are generally made more for research or marketing purposes than anything else; to bring hype and attention to a company, rather than having an expectation of any practical use or commercial success.

This is because properly understanding, and reacting to, an external environment is an extremely difficult task. Humans have decades of experience of foreign objects, materials, and estimations of distance and speed such as figuring out how far away an approaching car is that an AI does not, and even humans frequently fail at this task. As Murray Shanahan recently pointed out, dogs are capable of navigating the external world, but current LLMs and robotics are not even close to being capable of this [82].

Most robots that function in a useful manner today, and will be capable of doing so in the foreseeable future, are those that work in limited, sandboxed environments, where reacting to a fully fledged external environment is not required. Anthropomorphic (human-like) forms are not frequently helpful, as human posture and center of gravity considerations make such designs unnecessary for many applications. Examples of useful non-anthropomorphic robotics applications can be seen in medicine and manufacturing [83], among other fields. While some useful anthropomorphic robots do exist, they are not often well known to the public, as they tend to only be useful in limited, sandboxed environments.

Dr. No (1962). Metal hands, though strong and durable, aren't always ideal.

It is important to note that this is only a bad use of AI at the current time, and only in the context of fully-fledged anthropomorphic robotics. Anthropomorphic robotics and the use of AI for them may be more feasible and useful at some point in the future when available data and other mechanisms are improved enough, though it is difficult to estimate when.

13.4 Facial recognition

Facial recognition, as an AI field, has been plagued with problems; in particular, the racial bias exhibited by most such systems [84]. It is no coincidence that the data such systems are trained on tends to be predominantly of specific ethnic groups, and that the AI - with little data of other groups - inevitably ends up unable to accurately recognise people in those groups and is prone to bias against them.

This bias problem is not solely about data. It is also a known problem that developers of AI algorithms are prone to infusing their algorithms with their own biases, effectively transferring responsibility for that bias to a machine that can be falsely claimed 'not to be biased'. An exceptionally good explanation of this phenomenon can be read in the paper "Conceptualizing Algorithmic Stigmatization" by Nazanin Andalibi et al [85].

13.5 Therapy

In January 2023, the mental health nonprofit Koko performed an experiment [86] involving the use of GPT-3 to write messages sent to patients who were seeking help. Putting aside for the sake of this article the numerous ethical concerns of said experiment, it was clear this would not work.

As the experiment noted, users rated AI responses highly until discovering they were written by AI, at which point they did not like them. Robert Morris, Koko's co-founder, noted that:

> "Once people learned the messages were co-created by a machine, it didn't work. Simulated empathy feels weird, empty." [87].

This should not come as a surprise: it feels weird not because it is simulated, but because it is not there at all. The basic concept of therapy relies on the patient being confident that the therapist does, in fact, have knowledge of a wide variety of different human conditions and the ability to empathise.

Suppose a patient goes to see a doctor complaining of an illness, and is prescribed a medicine. If the patient goes home and searches Google for their symptoms, and it comes up with the same medicine, that does not mean that the doctor and Google used the same information to decide which medicine was correct. The doctor has consulted their vast knowledge of medicine; Google has found the most "relevant and popular" search result. For this reason, patients self-medicating via Google is not known for being a medically wise decision; thus most medicines must be prescribed to the patient by a qualified medical professional, instead of the patient being allowed to buy it themselves.

An AI dispensing therapy is no different; that it came up with an answer does not mean it understood the answer, or that it understood whether the answer was correct in the context of the patient's medical history. It is therefore of no surprise that patients would immediately lose confidence in therapy when learning the messages were written by an AI, because they can no longer trust that medical understanding of their condition is guiding the treatment.

14 Declaration of interests

I did not receive any funding or other compensation to write this paper.

I have no known conflicts of interest or potential conflicts of interest.

15 License

16 Bibliography

16.1 Web Articles

[2] James O'Malley, TechRadar. Captcha if you can: how you've been training AI for years without realising it. [Accessed 08-May-2023]. Jan. 2018.

[3] Amazon Web Services. Welcome to the Amazon Mechanical Turk Requester User Interface Guide. [Accessed 11-May-2023].

[4] Hive. Why We Worked with Parler to Implement Effective Content Moderation. [Accessed 11-May-2023]. May 2021.

[6] Kyle Wiggers, TechCrunch. Stability AI, the startup behind Stable Diffusion, raises $101M. [Accessed 11-May-2023]. Oct. 2022.

[7] Katie Roof & Mark Bergen & Hannah Miller, Bloomberg. OpenAI Rival Stable Diffusion Maker Seeks to Raise Funds at $4 Billion Valuation. [Accessed 11-May-2023]. Mar. 2023.

[8] Kenrick Cai, Forbes. The $2 Billion Emoji: Hugging Face Wants To Be Launchpad For A Machine Learning Revolution. [Accessed 13-May-2023]. 2022.

[12] LAION. FAQ. [Accessed 11-May-2023].

[13] Dina Bass, Bloomberg. Microsoft Invests $10 Billion in ChatGPT Maker OpenAI. [Accessed 11-May-2023]. Jan. 2023.

[15] Reece Rogers, Wired. Is GPT-4 Worth the Subscription? Here's What You Should Know. [Accessed 12-May-2023]. Mar. 2023.

[17] International Labour Organisation. Digital labour platforms and the future of work: Towards decent work in the online world. [Accessed 11-May-2023]. Sept. 2018.

[18] Sarah Butler, The Guardian. Uber drivers entitled to workers' rights, UK supreme court rules. [Accessed 09-May-2023]. Feb. 2021.

[19] Paul Sandle, Reuters. UK Court of Appeal confirms Deliveroo riders are self employed. [Accessed 09-May-2023]. June 2021.

[20] Umbrella. Union to take on Deliveroo over riders' employment status. [Accessed 09-May-2023]. Sept. 2022.

[21] Kyle Wiggers, TechCrunch. This startup is setting a DALL-E 2-like AI free, consequences be damned. [Accessed 13-May-2023]. Aug. 2022.

[22] Blake Brittain, Reuters. Getty Images lawsuit says Stability AI misused photos to train AI. [Accessed 26-April-2023]. Feb. 2023.

[23] Matthew Butterick. We've filed a lawsuit challenging Stable Diffusion, a 21st-century collage tool that violates the rights of artists. [Accessed 13-May-2023]. Jan. 2023.

[24] Chloe Xiang, VICE Magazine. A Photographer Tried to Get His Photos Removed from an AI Dataset. He Got an Invoice Instead. [Accessed 13-May-2023]. Apr. 2023.

[25] Jon Baumgarten, Copyright Alliance. Former Copyright Office GC Warns Against Blanket Assertions That AI Ingestion of Copyrighted Works 'Is Fair Use'. [Accessed 24-May-2023]. May 2023.

[26] EUR-Lex. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). [Accessed 29-May-2023].

[27] Stephanie Bodoni, Bloomberg. Meta Repeats Why It May Be Forced to Pull Facebook From EU. [Accessed 26-May-2023]. July 2022.

[28] Upmanyu Trivedi, Bloomberg. Apple Loses UK Patent Appeal Over Essential Technologies. [Accessed 26-May-2023]. Oct. 2022.

[29] Jeff Parsons, Metro. Apple threatens to leave UK market over £5 billion lawsuit. [Accessed 26-May-2023]. July 2021.

[30] The Guardian. Elon Musk threatens to move Tesla HQ out of California over Covid-19 restrictions. [Accessed 26-May-2023]. May 2020.

[31] Elon Musk. Twitter. [Accessed 26-May-2023]. May 2020.

[32] Faiz Siddiqui, The Washington Post. Tesla gave workers permission to stay home rather than risk getting covid-19. Then it sent termination notices. [Accessed 26-May-2023]. June 2020.

[33] Mike Murphy, MarketWatch. Tesla factory had more than 400 COVID-19 cases after Elon Musk's defiant reopening: report. [Accessed 26-May-2023]. Mar. 2021.

[34] Siladitya Ray, Forbes. ChatGPT Could Leave Europe, OpenAI CEO Warns, Days After Urging U.S. Congress For AI Regulations. [Accessed 26-May-2023]. May 2023.

[37] Glaze Project, University of Chicago. About. [Accessed 21-April-2023].

[41] Jonathan Vanian & Kif Leswing, CNBC. ChatGPT and generative AI are booming but at a very expensive price. [Accessed 21-April-2023]. Mar. 2023.

[42] The Economist. The cost of training machines is becoming a problem. [Accessed 21-April-2023]. June 2020.

[43] National Institute for Health and Care Excellence. Controlled drugs and drug dependence. [Accessed 12-May-2023].

[47] Celsys. Twitter. [Accessed 24-April-2023]. Nov. 2022.

[48] Celsys. Twitter. [Accessed 24-April-2023]. Dec. 2022.

[49] Geoffrey A. Fowler, Washington Post. We tested a new ChatGPT-detector for teachers. It flagged an innocent student. [Accessed 30-May-2023]. Apr. 2023.

[50] OpenAI. New AI classifier for indicating AI-written text. [Accessed 30-May-2023]. Jan. 2023.

[52] Chris Stokel-Walker, BuzzFeed. A Professional Artist Spent 100 Hours Working On This Book Cover Image, Only To Be Accused Of Using AI. [Accessed 30-May-2023]. Jan. 2023.

[53] Ben Moran. Twitter. [Accessed 30-May-2023]. Dec. 2022.

[54] Kayla Jimenez, USA Today. Professors are using ChatGPT detector tools to accuse students of cheating. But what if the software is wrong? [Accessed 30-May-2023]. Apr. 2023.

[55] Brian Martin, University of Wollongong. Defamation law and free speech. [Accessed 30-May-2023].

[56] Ministry of Justice, UK Government. Government clampdown on the abuse of British courts to protect free speech. [Accessed 30-May-2023]. Mar. 2022.

[57] Equality Now. Weaponizing Defamation Lawsuits Against Survivors Violates International Human Rights. [Accessed 30-May-2023]. Nov. 2021.

[58] Molly White, Web3 is Going Just Great. Africrypt investors disappear with $3.6 billion of investor funds. [Accessed 29-April-2023]. Apr. 2021.

[59] Molly White, Web3 is Going Just Great. $60 million disappears in AnubisDAO project within a day of its launch. [Accessed 29-April-2023]. Oct. 2021.

[60] Eric James Beyer, NFT Now. The Biggest Rug Pulls in NFT History. [Accessed 29-April-2023]. July 2022.

[61] Dave Tracey, Contino. Who's Using Microsoft Azure? [Accessed 27-April-2023]. Feb. 2020.

[62] Nikhil Suryawanshi. The Biggest AWS Users. [Accessed 27-April-2023]. Sept. 2020.

[63] Google. Google Cloud customers. [Accessed 27-April-2023].

[64] Amazon Web Services. Amazon Rekognition FAQs. [Accessed 13-May-2023].

[65] Microsoft Azure. Face API FAQ. [Accessed 13-May-2023].

[66] Microsoft Azure. Shared responsibility in the cloud. [Accessed 29-April-2023]. May 2022.

[67] Amazon Web Services. Shared Responsibility Model. [Accessed 29-April-2023].

[68] Google Cloud Platform. Shared responsibilities and shared fate on Google Cloud. [Accessed 29-April-2023]. July 2022.

[69] V7 Labs. 8 Practical Applications of AI in Agriculture. [Accessed 21-April-2023]. Oct. 2021.

[71] Hanae Armitage, Stanford Medicine. Artificial intelligence rivals radiologists in screening X-rays for certain diseases. [Accessed 21-April-2023]. Nov. 2018.

[76] Stripe. A primer on machine learning for fraud detection. [Accessed 02-May-2023]. Dec. 2021.

[79] Madhurjya Chowdhury, Analytics Insight. What is AI Image Recognition? How Does It Work in the Digital World? [Accessed 21-April-2023]. Feb. 2022.

[81] Steve Povolny, McAfee. Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles. [Accessed 24-April-2023]. Feb. 2020.

[82] Murray Shanahan. Twitter. [Accessed 23-April-2023]. Apr. 2023.

[83] Mark Fairchild. Top 12 Industrial Robot Applications and Uses. [Accessed 29-April-2023]. Aug. 2021.

[84] Alex Najibi, Harvard University. Racial Discrimination in Face Recognition Technology. [Accessed 21-April-2023]. Oct. 2020.

[86] Chloe Xiang, VICE Magazine. Startup Uses AI Chatbot to Provide Mental Health Counseling and Then Realizes It ‘Feels Weird’. [Accessed 23-April-2023]. Jan. 2023.

[87] Robert Morris. Twitter. [Accessed 09-May-2023]. Jan. 2023.

16.2 Academic Papers

[1] Murray Shanahan. *Talking About Large Language Models*. 2023. arXiv: 2212.03551 [cs.CL].

[5] Christoph Schuhmann et al. *LAION-5B: An open large-scale dataset for training next generation image-text models*. 2022. arXiv: 2210.08402 [cs.CV].

[9] Lisa A. Bero. “Tobacco industry manipulation of research”. In: *Public Health Rep.* 120.2 (2005), pp. 200–208. DOI: 10.1126/science.abk0063.

[10] G. Supran, S. Rahmstorf, and N. Oreskes. “Assessing ExxonMobil’s global warming projections”. In: *Science* 379.6628 (2023), eabk0063. DOI: 10.1126/science.abk0063.

[11] Fiona Godlee et al. “Journal policy on research funded by the tobacco industry”. In: *Thorax* 68.12 (2013), pp. 1090–1091. ISSN: 0040-6376. DOI: 10.1136/thoraxjnl-2013-204531.

[14] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[16] Kotaro Hara et al. “A Data-Driven Analysis of Workers’ Earnings on Amazon Mechanical Turk”. In: (2017). arXiv: 1712.05796 [cs.CY].

[35] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. 2015. arXiv: 1412.6572 [stat.ML].

[36] Kaidi Xu et al. *Adversarial T-shirt! Evading Person Detectors in A Physical World*. 2020. arXiv: 1910.11099 [cs.CV].

[38] Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. *Evading Watermark based Detection of AI-Generated Content*. 2023. arXiv: 2305.03807 [cs.LG].

[39] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, Ben Y. Zhao. *GLAZE: Protecting Artists from Style Mimicry by Text-to-Image Models*. 2023. arXiv: 2302.04222 [cs.CR].

[40] Ali Shafahi et al. *Are adversarial examples inevitable?* 2020. arXiv: 1809.02104 [cs.LG].

[44] Nicholas Carlini et al. *Extracting Training Data from Diffusion Models*. 2023. arXiv: 2301.13188 [cs.CR].

[45] Carl Guo et al. Stable Diffusion Objectively Succeeds at Copycatting Specific Artists’ Styles. May 2023.

[46] Huangzhao Zhang, Hao Zhou, Ning Miao, Lei Li. *Generating Fluent Adversarial Examples for Natural Languages*. Generating Fluent Adversarial Examples for Natural Languages. 2020. arXiv: 2007.06174 [cs.CL].

[51] Vinu Sankar Sadasivan et al. *Can AI-Generated Text be Reliably Detected?* 2023. arXiv: 2303.11156 [cs.CL].

[70] Eli-Chukwu, N. C. “Applications of Artificial Intelligence in Agriculture: A Review”. In: *Engineering, Technology & Applied Science Research* (Aug. 2019), pp. 4377–4383. DOI: 10.48084/etasr.2756.

[72] Debleena Paul, Gaurav Sanap, Snehal Shenoy, Dnyaneshwar Kalyane, Kiran Kalia, Rakesh K. Tekade. “Artificial intelligence in drug discovery and development”. In: *Drug Discov Today* 26.1 (2020), pp. 80–93. DOI: 10.1016/j.drudis.2020.10.010.

[73] Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, Madhu Babu, Mohamed Jawed Ahsan. “Machine Learning in Drug Discovery: A Review”. In: *Artif Intell Rev* 55.3 (2022), pp. 1947–1999. DOI: 10.1007/s10462-021-10058-4.

[74] Benjamin Hunter, Mitchell Chen, Prashanthi Ratnakumar, Esubalew Alemu, Andrew Logan, Kristofer Linton-Reid et al. “A radiomics-based decision support tool improves lung cancer diagnosis in combination with the Herder score in large lung nodules”. In: *The Lancet* (2022). DOI: 10.1016/j.ebiom.2022.104344.

[75] Jerry Tang, Amanda LeBel, Shailee Jain, Alexander G. Huth. “Semantic reconstruction of continuous language from non-invasive brain recordings”. In: *Nature Neuroscience* (2023). [Accessed via The Guardian on 02-May-2023]. DOI: 10.1038/s41593-023-01304-9.

[77] Bashar Ahmed Khalaf et al. “Comprehensive Review of Artificial Intelligence and Statistical Approaches in Distributed Denial of Service Attack and Defense Methods”. In: *IEEE Access* 7 (2019), pp. 51691–51713. DOI: 10.1109/ACCESS.2019.2908998.

[78] Martin Zadnik, Elena Carasec. “AI infers DoS mitigation rules”. In: *Journal of Intelligent Information Systems* 60 (2023), pp. 305–324. DOI: 10.1007/s10844-022-00728-2.

[80] Miriah Steiger, Timir J. Bharucha, Sukrit Venkatagiri, Martin J. Riedl, Matthew Lease. “The Psychological Well-Being of Content Moderators”. In: *CHI Conference on Human Factors in Computing Systems* (2021). DOI: 10.1145/3411764.3445092.

[85] Nazanin Andalibi et al. “Conceptualizing Algorithmic Stigmatization”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3580970.