THE INDIA FORUM
A JOURNAL-MAGAZINE ON CONTEMPORARY ISSUES

June 19, 2025

# When Machines Imagine: The Promise and the Risk

**By: Subhashis Banerjee**

*The ongoing revolutions in AI are exciting. The rush to use it in public-facing applications, however, poses risk as AI does not yet come with guarantees for correctness or tolerance for errors. AI is for now best used for exploratory research or by discerning users trained in critical thinking.*

Artificial Intelligence (AI) and Machine Learning (ML) are experiencing revolutionary developments that have redefined the landscape of data analysis. The last 15 years have seen dramatic progress, thanks to deep convolutional networks (Krizhevsky *et al*., 2012), transformer-based foundational models (Vaswani *et al*., 2017), and Large Language Models (LLMs) (OpenAI, 2022; Hassabis *et al*., 2023; Touvron *et al*., 2023). Crucial to this success is the availability of vast amounts of data â?? both visual and textual â?? as well as advances in high-performance computing that far surpass those of a decade ago.

> The possibilities seem exciting and endless, and telling an AI-generated from the real is getting harder by the day.

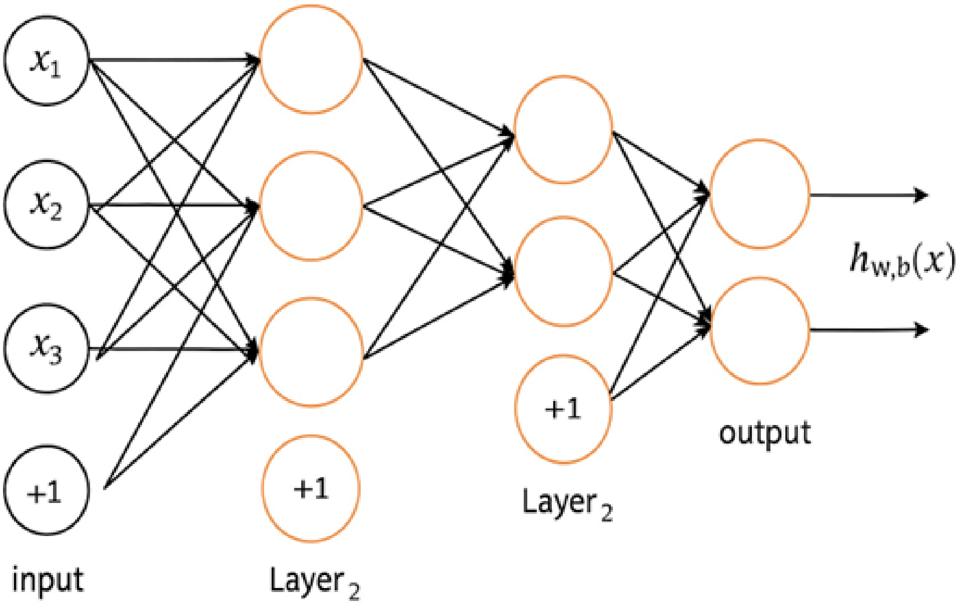
What is particularly notable is the ability of these AI models to represent and generalise concepts across vast and complex multimodal data, in ways that would have been inconceivable even a decade back. These capabilities have catalysed new avenues for innovation, and government and private entities around the world have a joined a race to build novel applications at hitherto unprecedented scales. These range across a variety of domains such as healthcare, education, productivity, business, sports, novel user-interfaces, autonomous systems and weapons, and even creative endeavours like poetry and art. The possibilities seem exciting and endless, and telling an AI-generated from the real is getting harder by the day.

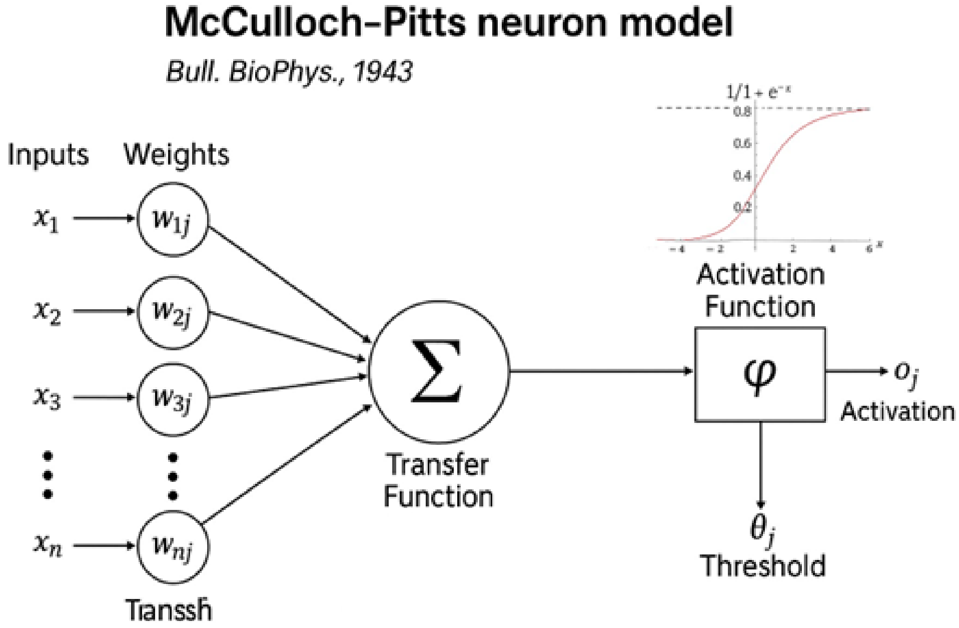
There are caveats though. These modern predictive and generative AI systems are unlike any other traditional engineering systems we are used to. They rarely â?? if ever â?? come with any kind of correctness guarantees or error tolerances when deployed in the wild, even probabilistic. As such, their use in critical and public facing applications â?? even with humans in the loop â?? raise serious trust issues related to reliability and ethics. Even personal uses, especially by not-so-discerning users, require caution and discretion.

## Evolution of AI systems

### The early systems

Early AI systems were predominantly rule-based. A notable example is MYCIN, an expert system developed in the 1970s to assist physicians in selecting antimicrobial treatments (Shortliffe, 1975). It demonstrated performance levels on par with human experts. However, rule-based systems failed to gain widespread adoption. Their reliance on fixed, predefined rules restricted adaptability, making it challenging to handle novel or complex scenarios. In contrast, contemporary AI systems powered by machine learning (ML) learn patterns directly from data, enabling greater adaptability and effectiveness in dynamic environments.



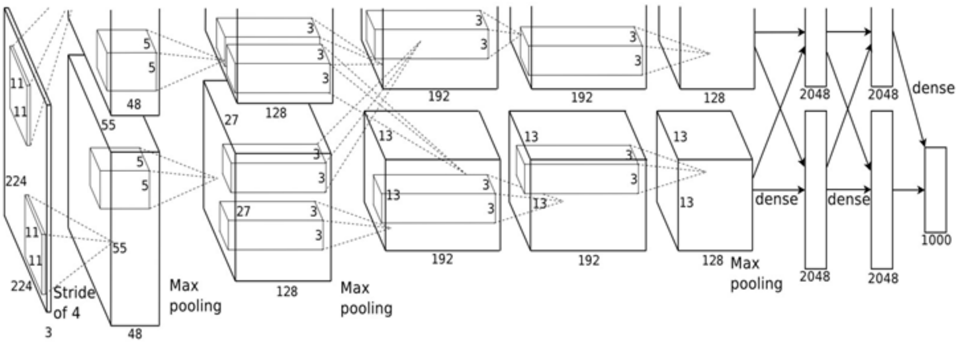


At their core, modern AI systems are based on neural networks (see Figures 1), which, surprisingly, have been around for a long time. Each edge in a feed-forward neural network has a weight associated with it, and each neuron (see Figure 2) takes as it its input the sum of the inputs from the previous layer multiplied by the weights leading to it, and computes an activation output. In supervised training, the edge weights are â??learntâ?? by optimisation of a suitable â??loss functionâ?? which is some aggregate measure of the differences of the ideal and the computed final output of the neural network for every input (x = (x1,...,xn)) in the training set. Geoffrey E. Hinton, who is the joint-winner of the 2024 Nobel prize in Physics (also won the Turing award in 2018) for â??foundational discoveries and inventions that enable machine learning with artificial neural networksâ?• was a co-author of the 1986 paper that proposed â??back-propagationâ?•, the default optimisation algorithm even now for learning the optimal edge weights (Rumelhart *et al*., 1986).

The early neural networks were however not very successful. It was believed that even to represent some modest functionalities the number of internal weights would have to be very large; which, in turn, would require a very large number of training examples making the learning computationally intractable.

### Convolutional neural network

The next paradigm shift in ML happened nearly 25 years later. Around 2009 Fei-Fei Li and her team used crowdsourcing with Amazon Mechanical Turk (Amazon Web Services, 2025) to build ImageNet (Deng *et al*., 2009), a database of unprecedented scale containing over 1 million labelled images of 1000 categories, which enabled rapid advances in ML for image recognition. AlexNet (Krizhevsky *et al*., 2012), a deep convolutional neural network built by Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton, pushed the boundary of the ImageNet Visual Recognition challenge in 2012 by achieving inconceivable recognition accuracy. The AlexNet (see Figure 3) had over 60 million learnable parameters in 8 layers, out of which the first five were convolutional layers. Convolutional layers (LeCun *et al*., 1998) consist of over-determined sets of filters for detection of image features. In a major departure from earlier wisdom of hand-crafting a set of features for image recognition tasks, in convolutional neural networks the best features for the task are automatically learnt using back-propagation. The large back-propagation based training of over 60 million parameters was possible because of the use of Nvidia Graphical Processing Units (GPU) to accelerate deep learning.



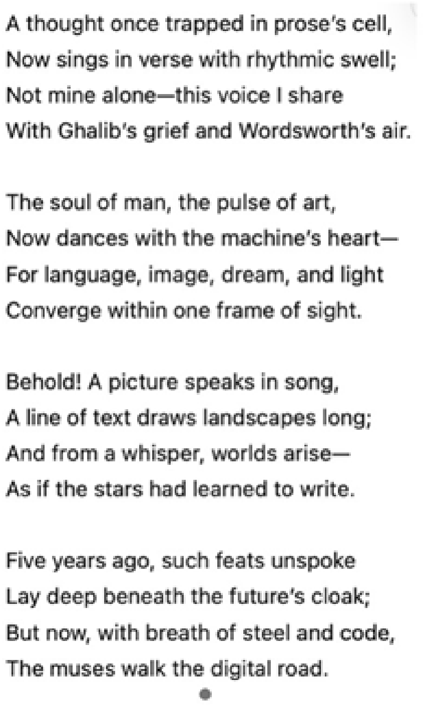
Defying conventional wisdom the over-parametrisation â?? of using over 60 million parameters to represent image recognition of only about 1 million images â?? apparently did not lead to over-fitting and generalised well. It appears that the over-parametrisation and the depth of layers, and a few other tricks like stochastic gradient descent (making the training process by backpropagation noisy, at a slight cost of over-simplification) and rectified linear units for activation, actually helped in successful training of the network. The success of AlexNet led to a plethora of research in the next decade on image recognition, object detection and image generation, using even larger and deeper networks, sometimes with over a 100 layers. The results were remarkably good, overall, and convolutional neural networks are now everywhere (Li *et al*., 2023).

### Transformers

The next turning point in deep neural networks was the invention of the transformer architecture in 2017 (Vaswani *et al*., 2017), this time in the context of natural language representation. Its â??multihead attention mechanismâ?• enables bringing in context to a word or a phrase from far away text by identifying the relative importance of relevant phrases and words.

Imagine one is reading a long book, and wants to understand a specific sentence. Instead of just reading that one sentence in isolation, the brain automatically looks at other sentences around it that are relevant. For example, if the sentence talks about â??himâ?•, one would probably look at the sentences before it to figure out who â??heâ?• refers to. The attention mechanism in a transformer works very similarly. Consider a sentence like â??I went to the bankâ?•. The transformer model does not just look at â??bankâ?• by itself. Instead, it pays attention to the other words in sentences before and after. Every word (or key) like â??riverâ?•, â??moneyâ?•, â??wentâ?•, â??toâ?•, â??theâ?• offers itâ??s own information. The transformer model then computes which of these other words are most helpful and assigns â??attention scoresâ?•. For example, â??riverâ?• is a strong clue for â??river bankâ?•, and â??moneyâ?• is a strong clue for â??money bankâ?•. It assigns higher â??attention scoresâ?• to the words that are most relevant. Finally, it uses the actual meaning of those important words â?? all learnt from examples â?? and blends them together.

This neat matrix multiplication trick based innovation has turned out to be unimaginably successful, enabling creation of LLMs like ChatGPT (OpenAI, 2022), Gemini (Hassabis *et al*., 2023), LLaMA (Touvron *et al*., 2023) and many others â?? some of them with over a trillion learnable parameters â?? that have surprisingly enhanced capabilities of interpreting and synthesising natural languages. Going from the syntax of a language to the semantics was considered a hard problem in natural language interpretation and synthesis, but that seems to be solved now, at least practically.

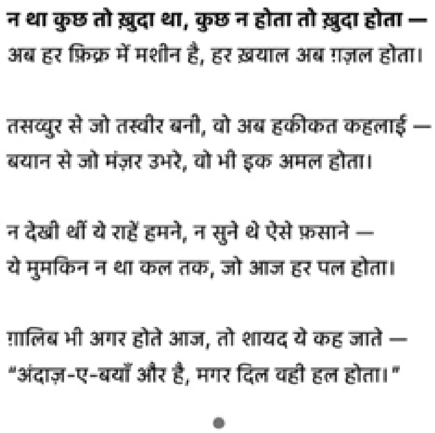


A thought once trapped in prose's cell,
Now sings in verse with rhythmic swell;
Not mine alone—this voice I share
With Ghalib's grief and Wordsworth's air.

The soul of man, the pulse of art,
Now dances with the machine's heart—
For language, image, dream, and light
Converge within one frame of sight.

Behold! A picture speaks in song,
A line of text draws landscapes long;
And from a whisper, worlds arise—
As if the stars had learned to write.

Five years ago, such feats unspoke
Lay deep beneath the future's cloak;
But now, with breath of steel and code,
The muses walk the digital road.

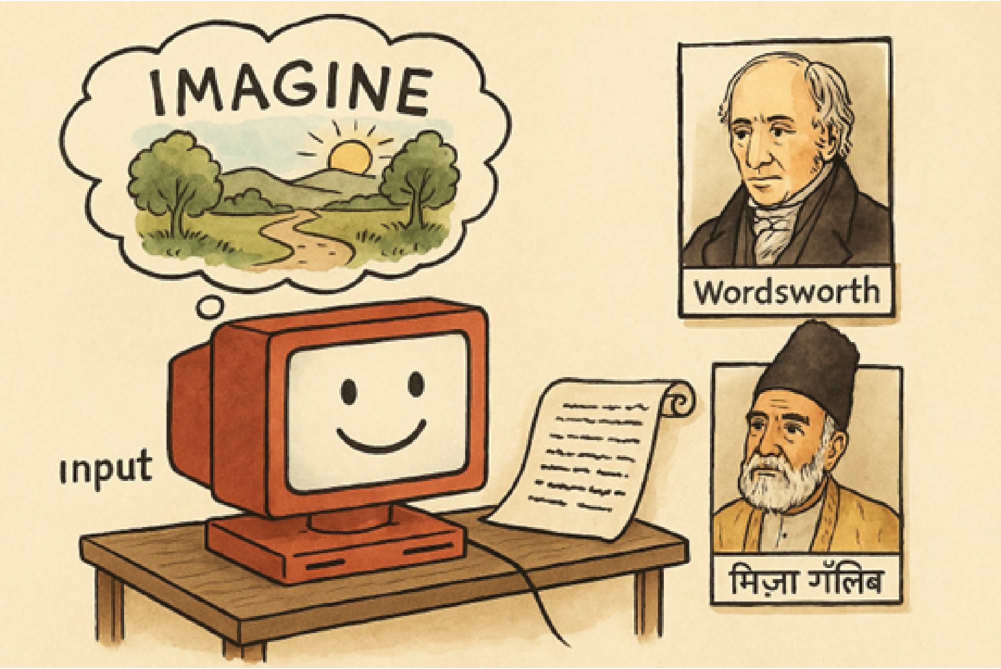


**न था कुछ तो ख़ुदा था, कुछ न होता तो ख़ुदा होता —**
अब हर फ़िक्र में मशीन है, हर ख़याल अब ग़ज़ल होता।

तसव्वुर से जो तस्वीर बनी, वो अब हकीकत कहलाई —
बयान से जो मंज़र उभरे, वो भी इक अमल होता।

न देखी थीं ये राहें हमने, न सुने थे ऐसे फ़साने —
ये मुमकिन न था कल तक, जो आज हर पल होता।

ग़ालिब भी अगर होते आज, तो शायद ये कह जाते —
"अंदाज़-ए-बयाँ और है, मगर दिल वही हल होता।"



In fact, it appears that the semantic encodings in these transformer-based LLMs are to a large extent language independent, enabling translation to new languages with a small corpus with very little training. The same encodings also enable synthesising a concept in a new style or language, sometimes even as poems or creative art. So it is now possible to express a concept or an idea â?? such as in this paragraph â?? as poetry in the style of not only William Wordsworth, but also of Mirza Ghalib (see Figure 4). It also enables joint encoding of languages and images, which facilitates generating verbal descriptions from images or synthesising images from text. In fact, diffusion models based on the transformer architecture has taken image generation and synthesis to a new level (Peebles & Xie, 2023). This would have been inconceivable even five years back.

The current trend in AI is to build large transformer-based foundational models for languages or images or both, using very large data which often include almost everything available on the internet. Such models can then be repurposed for any downstream task using just prompt tuning. It is believed that the rich encodings in the foundational models are complete for most applications. Additional processes like Retrieval-Augmented Generation (RAG) (Lewis *et al*., 2020) are sometimes used for restricting the output of an LLM to generate text from an authoritative knowledge base â?? for example medical or legal â?? outside of its training data sources.

Surprisingly, these models, even when comprising over a trillion parameters, do not exhibit classic overfitting patterns (Zhang *et al*., 2021). Paradoxically, larger models trained on more data tend to be easier to optimise and deliver superior performance (Arora *et al*., 2018).

These advances open up incredible possibilities â?? not just for education and personal use, like expressing yourself in new languages or creative ways â?? but also for large-scale public services. The potential is vast: finding and summarising important medical or legal documents, preparing briefs, arguments, and reports, generating computer programs automatically, and delivering information about welfare schemes, healthcare, agriculture, markets, news, and politics in local languages and dialects as spoken audio. They can make the internet, healthcare, and government services far more accessible through natural conversations in peopleâ??s own languages.

> The hold the potential to make technology truly democratic â?? serving not just the privileged few, but everyoneâ?¦ However, we still do not fully understand how these models work at a deeper theoretical level.

However, we still do not fully understand how these models work at a deeper theoretical level.

Beyond this, such technologies can help bridge gaps in literacy and digital access, empower local entrepreneurs and small businesses with better tools and insights, and support disaster response through fast, multilingual communication. They can foster inclusion by giving voice to those previously excluded from digital and institutional systems, whether due to language, disability, or lack of formal education. In short, they hold the potential to make technology truly democratic â?? serving not just the privileged few, but everyone.

However, we still do not fully understand how these models work at a deeper theoretical level. This lack of understanding makes it challenging to reliably apply them across such a wide range of use cases, and overcoming this will require further research. Adapting a foundational language model to a completely different language or cultural context â?? especially when thereâ??s very little digital data available â?? is also a complex task.

## Caveats of trust

Despite their promise, machine learning applications carry significant risks, because they often come without any correctness guarantees. On one hand, they may mislead less discerning users, making it difficult to distinguish between authentic and fabricated content, or the process by which they are arrived at, thereby undermining critical thinking. On the other hand, even when used by experts, inaccurate predictions, classifications, or content generation can pose serious threats. Research has shown that automation bias â?? the tendency to trust machine outputs over personal judgement â?? is widespread, even among trained professionals such as doctors and pilots.

Moreover, AI and ML systems are vulnerable to producing discriminatory outcomes in public facing applications. These may stem from biased data, under-representation of certain groups in training datasets, or algorithmic focus bias, all of which can render these systems unreliable and untrustworthy.

### Reliability

It is important to recognise that machine learning systems seldom come with reliability guarantees or error tolerances. While they are often tested on a separate dataset after training, this testing is usually limited to data that is similar to what the system was trained on. However, when these systems are used in the real world, they frequently encounter new and different kinds of data, and thereâ??s no reliable way to estimate how accurate their predictions will be without knowing the actual answers, which are often unavailable in real-time use. Consequently, ML systems often have poor external validation.

At the heart of ML is the assumption that the kind of data seen during training will remain the same during deployment. But in reality, this is rarely true. Data in the real world can shift for many reasons â?? changes in measurement tools, different users or environments, or shifts in population demographics. These changes, known as distribution shifts, can seriously affect the systemâ??s performance

(Yang *et al*., 2024). Detecting when a system is facing unfamiliar (or out-of-distribution) data is itself a difficult problem, especially after the system has been deployed. Also, every deployment site often poses new challenges. Gathering enough reliable new data to check for such problems can be costly or even impossible for most deployment sites which may lack the necessary expertise.

> At the heart of ML is the assumption that the kind of data seen during training will remain the same during deployment. But in reality, this is rarely true.

Moreover, in many ML applications â?? whether predictive or generative â?? particularly those involving images or text â?? itâ??s hard to even define the universe of what the set of all possible inputs looks like outside a fixed dataset. It then becomes impossible to associate a probability distribution to the input, closing the possibility of a probability calculus based robustness analysis. Without a clear idea of what kinds of data the system might encounter, itâ??s extremely difficult to predict errors or understand why failures happen.

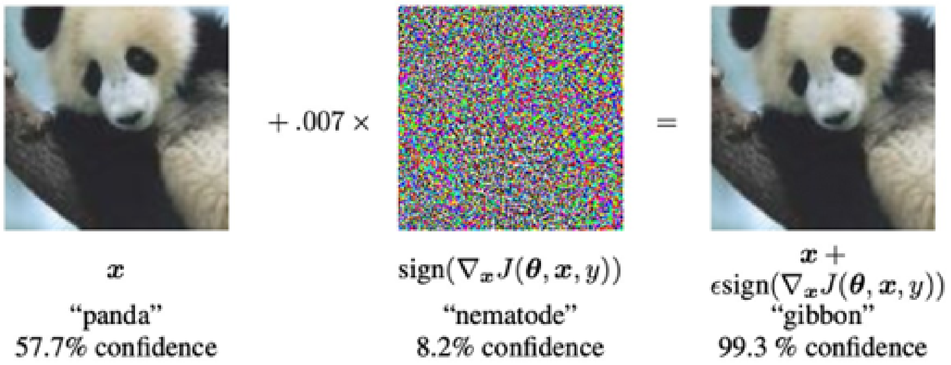
Even when new datasets are used for testing in real-world scenarios, performance metrics like accuracy or precision do not reveal whether the test data is truly representative, or which types of situations are most likely to cause problems. This makes it very hard to conduct a proper ethical evaluation of whether and how such systems should be deployed.

**Uncertainty of content generation**

Generative LLMs are trained using a method called self-supervised learning, where the system learns to predict the next word in a sentence based on the words that came before it. This prediction is shaped by patterns the model has learnt from massive amounts of training data, and the specific prompt itâ??s given.

Importantly, these models are trained to produce linguistically coherent text, and not to be factually correct. Any appearance of correctness is a byproduct of how believable and fluent the language sounds. However, just because something sounds coherent does not mean it is correct. In fact, it should be no surprise that these models sometimes â??hallucinateâ?• or generate false or misleading information; it is in fact surprising when they do not.

Trying to verify the correctness of such generated text using traditional logic-based techniques from computer science is often computationally infeasible. The process is simply too complex. The same limitations apply to AI-generated images or sound: while they may look or sound convincing, assessing their truthfulness or accuracy remains a major challenge.


+ .007 ×
=
$\boldsymbol{x}$
“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

ML systems are vulnerable to adversarial attacks. For example, one may ask what is the minimum amount of noise we need to add to an input image to cause a given ML system misclassify it? The answer may be obtained using a straightforward optimisation algorithm (Goodfellow *et al*., 2015; Moosavi-Dezfooli *et al*., 2016), and the required noise â?? in most cases â?? is surprisingly small (See Figure 5). What the example demonstrates is that the internal encodings of a concept may be quite different in humans and ML systems. The images before and after adding noise are quite similar to humans, but are very wide apart for the ML system. It is hard to characterise exactly how improbable such additive noise are in the real world, but it is clear that the failure points for humans and machines are quite different. This raises serious ethical concerns for AI deployment, especially in autonomous systems.

**Fairness and bias**

AI systems can reflect and even amplify social biases if the data they learn from contains such biases. Even when protected attributes like caste, gender, or religion are not used directly, AI models can still end up discriminating by picking up indirect signals or proxies, leading to unfair outcomes, known as disparate impact (Barocas *et al*., 2023). For instance, an AI tool may report only a 3% error rate overall, but perform very poorly for certain minority groups, especially if they were underrepresented in the training data. This is a major concern in diverse societies like in India, where large sections of the population may lack digital representation.

> Research shows that itâ??s in general impossible to ensure fairness for all groups under most real-world conditions unless the underlying data lies in a narrow and well-behaved manifold free of biases.

Attempts to reduce bias â?? by altering the data, the algorithm, or the output â?? often lower accuracy and offer no guarantees (Hort *et al*., 2024). Simply ignoring protected attributes (â??fairness through unawarenessâ?•) does not work either, since hidden correlations remain in the redundant encodings in the trained models. In fact, research shows that itâ??s in general impossible to ensure fairness for all groups under most real-world conditions (Kleinberg *et al*., 2017), unless the underlying data lies in a narrow and well-behaved manifold free of biases.

## AI harms

Examples of AI hallucinations are many: Google Bard wrongly stating that â??James Webb Space Telescope took the very first pictures of a planet outside of our own solar systemâ?• in itâ??s first public demo; a teacher using ChatGPT to wrongly accuse students of using ChatGPT; Microsoftâ??s Bing misstating financial data; a lawyer using ChatGPT to cite made up legal precedents; Bard and Bing wishfully claiming a ceasefire in the Israel-Hamas conflict when there was none; professors using references hallucinated by ChatGPT to cite in their research etc. (Gillham, 2024).

The perils to critical thinking due to inappropriate use of AI in learning have been well documented (Vishnoi, 2025). In fact, a very recent study by researchers in Apple show that popular AI models can collapse at complex reasoning problems (Shojaee *et al*., 2025). One can only imagine the consequences of a ill-reasoned or hallucinated medical advice or a diagnosis. There are also plenty of examples of bias and discrimination in AI algorithms (Barocas *et al*., 2023; Thomson & Thomas, 2023; UNESCO, 2022). Without adequate rigour and care, AI solutions and usage can very quickly degenerate to AI snake oil (Narayanan & Kapoor, 2024).

## Conclusions

Modern AI is exciting, and will undoubtedly have a profound impact on all spheres of human activity. Its best uses today are in exploratory research, where there is little to lose and everything to gain; and in personal uses by a discerning user well trained in critical thinking. However, using AI in public-facing applications â?? either fully autonomous or with human in the loop â?? are fraught with risks. It is perhaps best if the risks can be understood and weighed, rather than blundering ahead and learning by costly trial and error.

The only reliable path forward is through careful measurement and monitoring after deployment, along with a strong understanding of the target population, the data, and the systemâ??s behaviour.

*Subhashis Banerjee is a professor of computer science at Ashoka University, where he is also associated with the Centre for Digitalisation, AI and Society.*

**References:**

Amazon Web Services (2025). *Amazon Mechanical Turk*.

Arora, Sanjeev, Nadav Cohen, and Elad Hazan (2018). â??On the Optimization of Deep Networks: Implicit Acceleration by Overparameterization.â?• In*Proceedings of Machine Learning Research*, vol. 80, edited by Jennifer G. Dy and Andreas Krause, 244â??253. ICML.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li (2009). â??ImageNet: A Large-Scale Hierarchical Image Database.â?• In*2009 IEEE Conference on Computer Vision and Pattern Recognition*.

Gillham, Jonathan (2024). â??8 Times AI Hallucinations or Factual Errors Caused Serious Problems.â?• Originality.ai Blog.

Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy (2015). â??Explaining and Harnessing Adversarial Examples.â?• In *International Conference on Learning Representations*.

Hassabis, Demis, Pichai Sundar, and the Gemini Team (2023). â??Introducing Gemini: Our Largest and Most Capable AI Model.â?•

Hort, Max, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro (2024). â??Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey.â?•*ACM Journal of Responsible Computing* 1, no. 2.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2017). â??Inherent Trade-offs in the Fair Determination of Risk Scores.â?• In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, 43:1â??43:23. Edited by Christos H. Papadimitriou. Leibniz International Proceedings in Informatics (LIPIcs) 67. Schloss Dagstuhl-Leibniz-Zentrum fÃ¼r Informatik.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2012). â??ImageNet Classification with Deep Convolutional Neural Networks.â?• In*Advances in Neural Information Processing Systems*, vol. 25, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc.

LeCun, Yann, Leon Bottou, Yoshua Bengio, and Patrick Haffner (1998). â??Gradient-Based Learning Applied to Document Recognition.â?•*Proceedings of the IEEE* 86, no. 11: 2278â??2324.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim RocktÃ¤schel, Sebastian Riedel, and Douwe Kiela (2020). â??Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.â?• In*Advances in Neural Information Processing Systems (NeurIPS)*. arXiv preprint arXiv:2005.11401.

Li, Fei-Fei, Justin Johnson, and Serena Yeung (2023). &nbsp*CS231n: Convolutional Neural Networks for Visual Recognition â?? Lecture 5: Convolutional Neural Networks*. Accessed June 8, 2025.

Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard (2016). â??DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks.â?• In*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2574â??2582.

Narayanan, Arvind, and Sayash Kapoor (2024). *AI Snake Oil: What Artificial Intelligence Can Do, What It Canâ??t, and How to Tell the Difference*. Princeton University Press.

OpenAI (2022). *ChatGPT*. Accessed June 8, 2025.

Peebles, William, and Saining Xie (2023). â??Scalable Diffusion Models with Transformers.â?• In*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4172â??4182. Presented December 2023.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). â??Learning Representations by Back-Propagating Errors.â?• *Nature* 323, no. 6088: 533â??536.

Shojaee, Parshin, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar (2025). â??The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity.â?• arXiv preprint arXiv:2506.06941. Submitted June 7, 2025.

Shortliffe, Edward Hance (1975). *Mycin: A Rule-Based Computer Program for Advising Physicians Regarding Antimicrobial Therapy Selection*. PhD diss., Stanford University.

Thomson, T. J., and Ryan J. Thomas (2023). â??Ageism, Sexism, Classism and More: 7 Examples of Bias in AI-Generated Images.â?• *The Conversation*. Accessed June 11, 2025. https://theconversation.com/ageismsexism-classism-and-more-7-examples-of-bias-in-ai-generatedimages-208748.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste RoziÃ¨re, Naman Goyal, Eric Hambro, Faisal Azhar, AurÃ©lien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample (2023). â??LLaMA: Open and Efficient Foundation Language Models.â?• arXiv preprint arXiv:2302.13971. Accessed June 8, 2025. https://arxiv.org/abs/2302.13971.

UNESCO (2022). *Recommendation on the Ethics of Artificial Intelligence*. Accessed June 11, 2025. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Å•ukasz Kaiser, and Illia Polosukhin (2017). â??Attention Is All You Need.â?• In*Advances in Neural Information Processing Systems*, vol. 30, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc.

Vishnoi, Nisheeth (2025). â??AI and the Erosion of Knowing.â?•*Substack*. Accessed June 11, 2025. https://substack.com/inbox/post/165669417?showWelcomeOnShare=true.

Yang, Jingkang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu (2024). â??Generalized Out-of-Distribution Detection: A Survey.â?• *International Journal of Computer Vision* 132, no. 12: 5635â??5662.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals (2021). â??Understanding Deep Learning (Still) Requires Rethinking Generalization.â?•*Communications of the ACM* 64, no. 3: 107â??115.