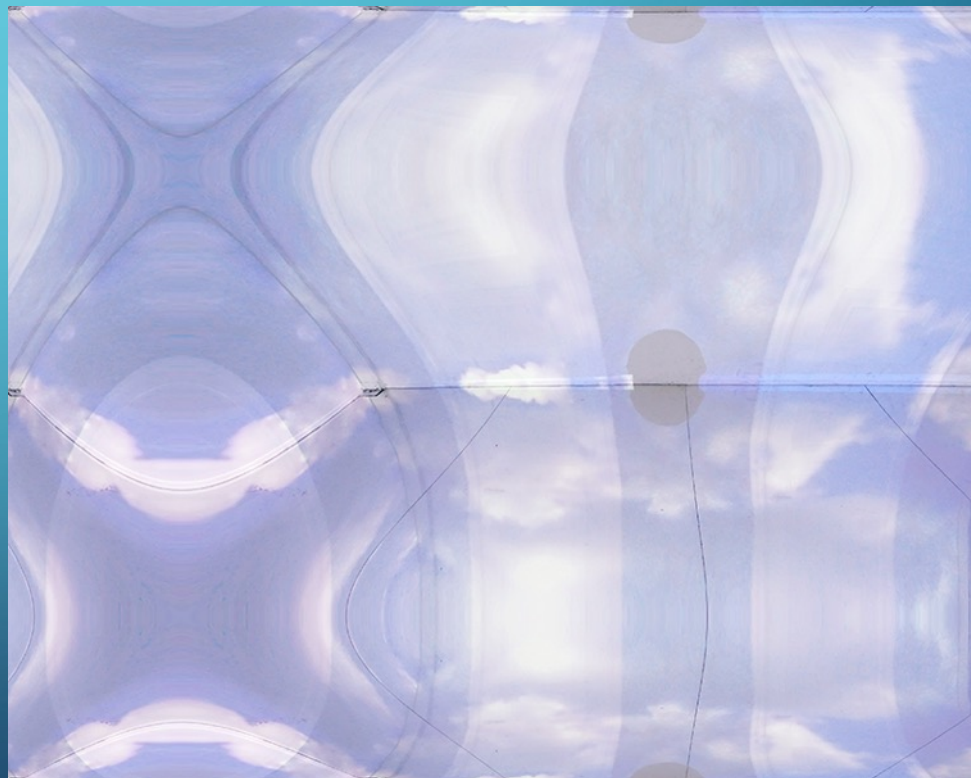


# 人工智能的演算法 帶來的負面影響

余創豪



# Elon Musk strikes deal to buy Twitter for \$44bn

By Natalie Sherman & Daniel Thomas  
Business reporter, New York

🕒 26 April



The board of Twitter has agreed to a \$44bn (£34.5bn) takeover offer from the billionaire Elon Musk.

Mr Musk, who made the shock bid less than two weeks ago, said Twitter had "tremendous potential" that he would unlock.

He also called for a series of changes from relaxing its content restrictions to eradicating fake accounts.


The firm initially rebuffed Mr Musk's bid, but it will now ask shareholders to vote to approve the deal.

Mr Musk is the world's richest person, according to Forbes magazine, with an estimated net worth of \$273.6bn mostly due to his shareholding in electric vehicle maker Tesla which he runs. He also leads the aerospace firm SpaceX.

- [Will Trump return to Twitter after Musk takeover?](#)
- [Why Elon Musk has been so keen to buy Twitter](#)
- [Musk buys Twitter: All you need to know - Podcast](#)

"Free speech is the bedrock of a functioning democracy, and Twitter is the digital town square where matters vital to the future of humanity are debated," Mr Musk said **in a statement announcing the deal.**

← **Thread**

 **Elon Musk** ✓  
@elonmusk

Free speech is essential to a functioning democracy.

Do you believe Twitter rigorously adheres to this principle?

Yes	29.6%
<b>No</b>	<b>70.4%</b>

2,035,924 votes · Final results

12:34 AM · Mar 25, 2022 · Twitter for iPhone

**46K** Retweets   **8,859** Quote Tweets   **190.5K** Likes

🗨️ ↻️ ❤️ 📤

 **Elon Musk** ✓ @elonmusk · Mar 25

Replying to @elonmusk

The consequences of this poll will be important. Please vote carefully.

🗨️ 11.3K   ↻️ 12.7K   ❤️ 163.6K   📤

 **Z.B.** @norcalfather · Mar 25

Replying to @elonmusk

Twitter is a private company. Unless we nationalize it to protect free speech Twitter can do what they want.

🗨️ 51   ↻️ 1   ❤️ 91   📤

⋮ [Show replies](#)

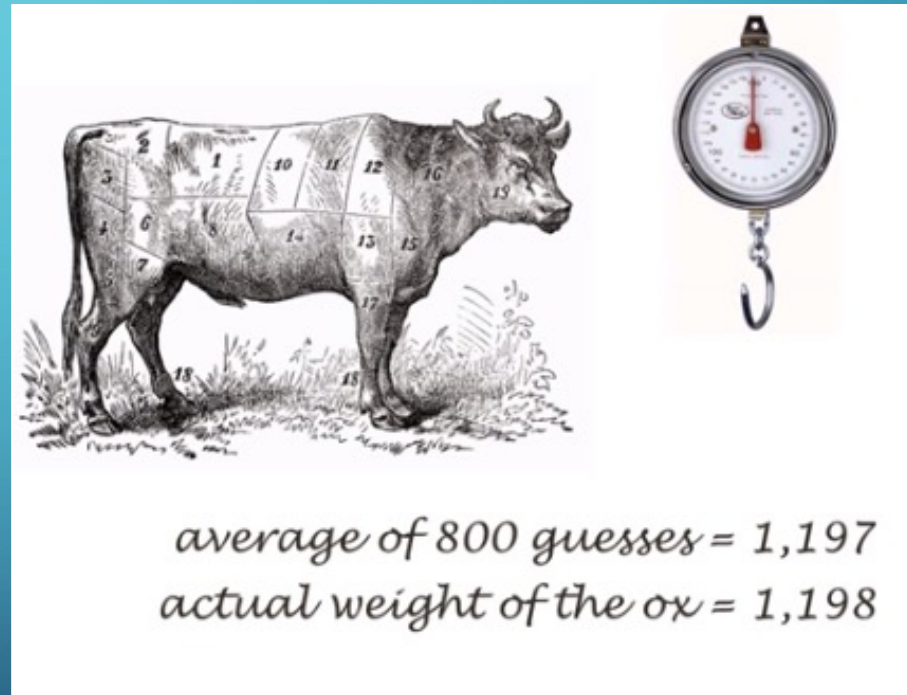


# 社交媒體的人工智能演算法

- 社交媒體使用人工智能演算法來決定你看到的內容。
- 2021 年，Facebook 的一份內部報告發現，社交媒體平台的演算法令源於東歐的虛假信息在 2020 年總統大選之前滲透了近一半的美國人。
- 人們之所以看到這些內容，是因為 Facebook 的內容推薦系統將內容放入了他們的新聞提要中。
- 民主受虛假信息破壞！

# 群眾的智慧(WISDOM OF THE CROWD)

- 社媒本來是好事。
- 在 1906 年普利茅斯的縣集市上，800 人猜測一頭牛的重量，其平均值十分接近真正的重量。
- 三個臭皮匠，一個諸葛亮。



# 群眾的的智慧(WISDOM OF THE CROWD)

- 集體預測通常比個人預測更準確。
- 理論上，社交媒體容讓百花齊放、百家爭鳴，應該更加鞏固民主社會的基礎。
- 但人性充滿弱點，另有用心的人在社媒上利用你弱點來操控你。

# 跟風效應 BANDWAGON EFFECT

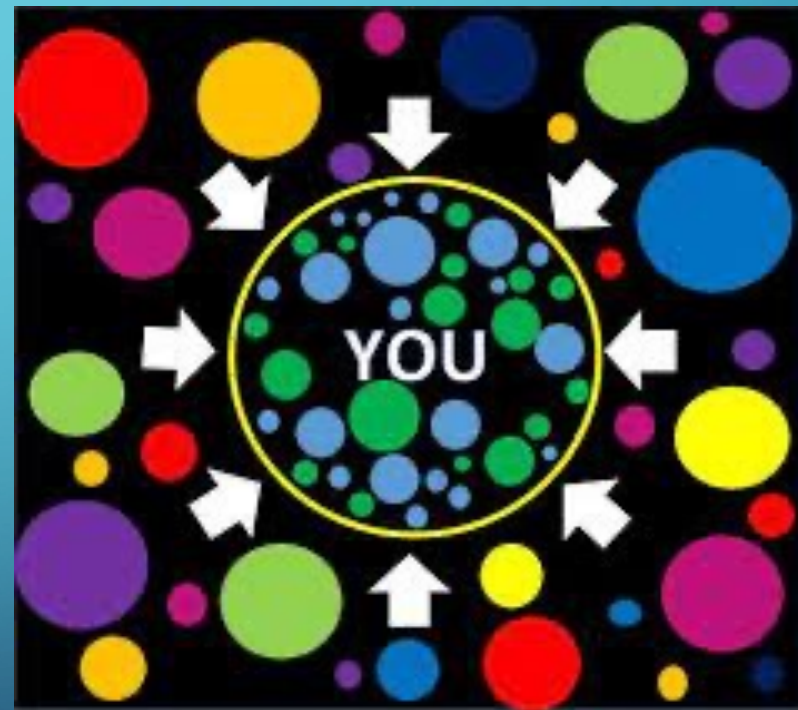
- 人們做某事主要是因為其他人也做同樣的事情。
- 該詞語起源自遊行花車，花車上的人鼓勵其他人們跳上車，熱鬧的音樂和慶祝活動具有很強的感染力。





# 迴聲室效果 ECHO CHAMBER EFFECT

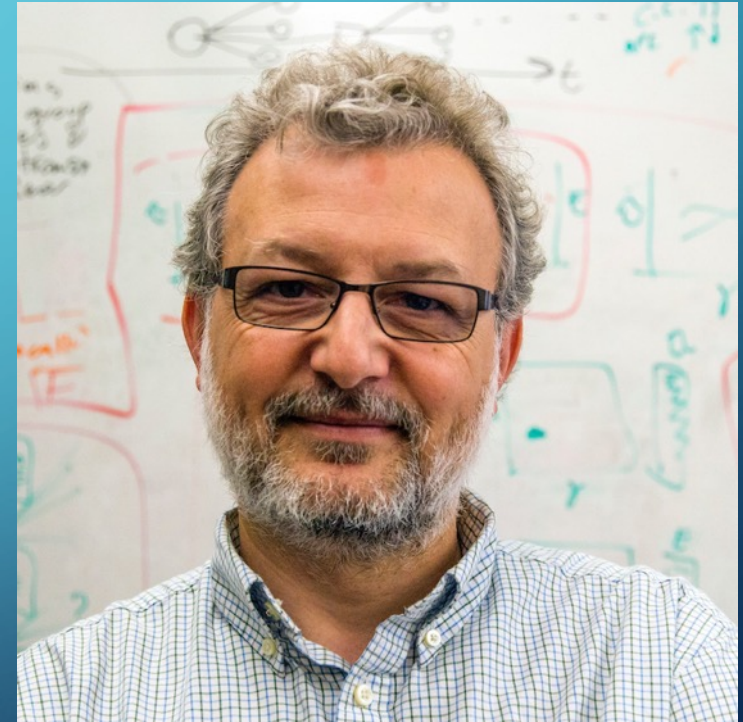
- 迴聲室效應發生在互聯網上。
- 社交媒體利用演算法向你發佈你自己喜歡聽的訊息，你只會聽到自己的回聲，越來越少接觸到不同的意見，久而久之，你發展出狹隘的視野（**Tunnel vision**）。





# 迴聲室效果 ECHO CHAMBER EFFECT

- 在印第安納大學信息學教授 **Filippo Menczer** 進行的一項模擬社媒的實驗中，參與者看了假新聞、垃圾科學、黨派觀點、陰謀論以及主流媒體文章。
- 當參與者看到許多其他用戶點讚或者分享來自低可信度來源的文章時，他們傾向於也點讚或分享這些文章，但他們不會標記這些文章以進行事實核查。



# 演算法的偏見是出於人為錯誤和疏忽

- 從 **2013** 年開始，荷蘭政府使用的演算法預測哪些人最有可能在兒童保育福利中欺詐，但政府並沒有等到有足夠證據就處罰那些家庭，要求他們償還多年的津貼。
- 指控那些家庭是基於低收入或雙重國籍等風險因素。
- 電腦和數據科學只可以展示數字，最終是由人去做決定，與其歸咎於演算法，不如說這是政府的疏忽，沒有進一步查證。

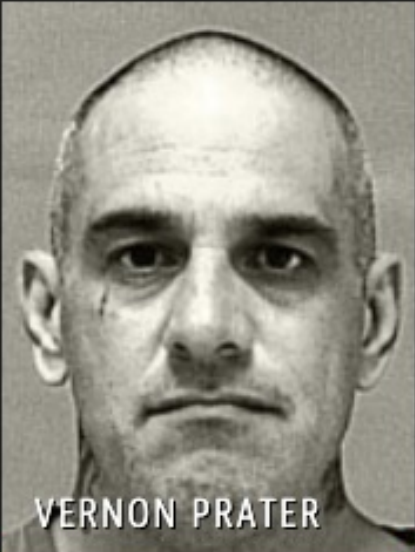
# 演算法的偏差是出於人為錯誤和疏忽

- **COMPAS: Correctional Offender Management Profiling for Alternative Sanctions**
- 一種用於預測犯人再犯風險從而做出保釋或假釋決定的軟件。
- **137**個因素，包括囚犯的年齡、性別、犯罪史等。
- 已在紐約、威斯康星、加利福尼亞、佛羅里達使用了很多年，但最近才檢查它的預測是否準確。

# 演算法的偏見

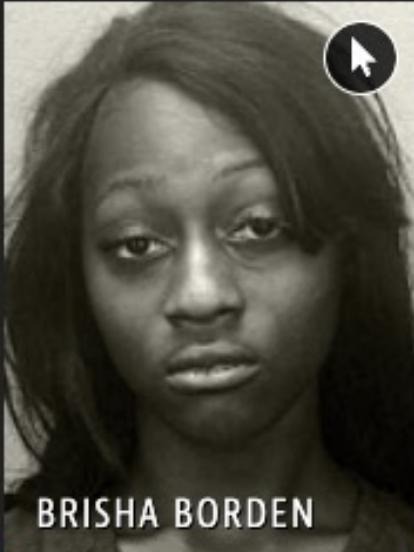
- 白人傾向於被判定為低風險，黑人傾向於被判為高風險。
- 但實際結果和預測差距很大。

### Two Petty Theft Arrests



VERNON PRATER

LOW RISK **3**



BRISHA BORDEN

HIGH RISK **8**

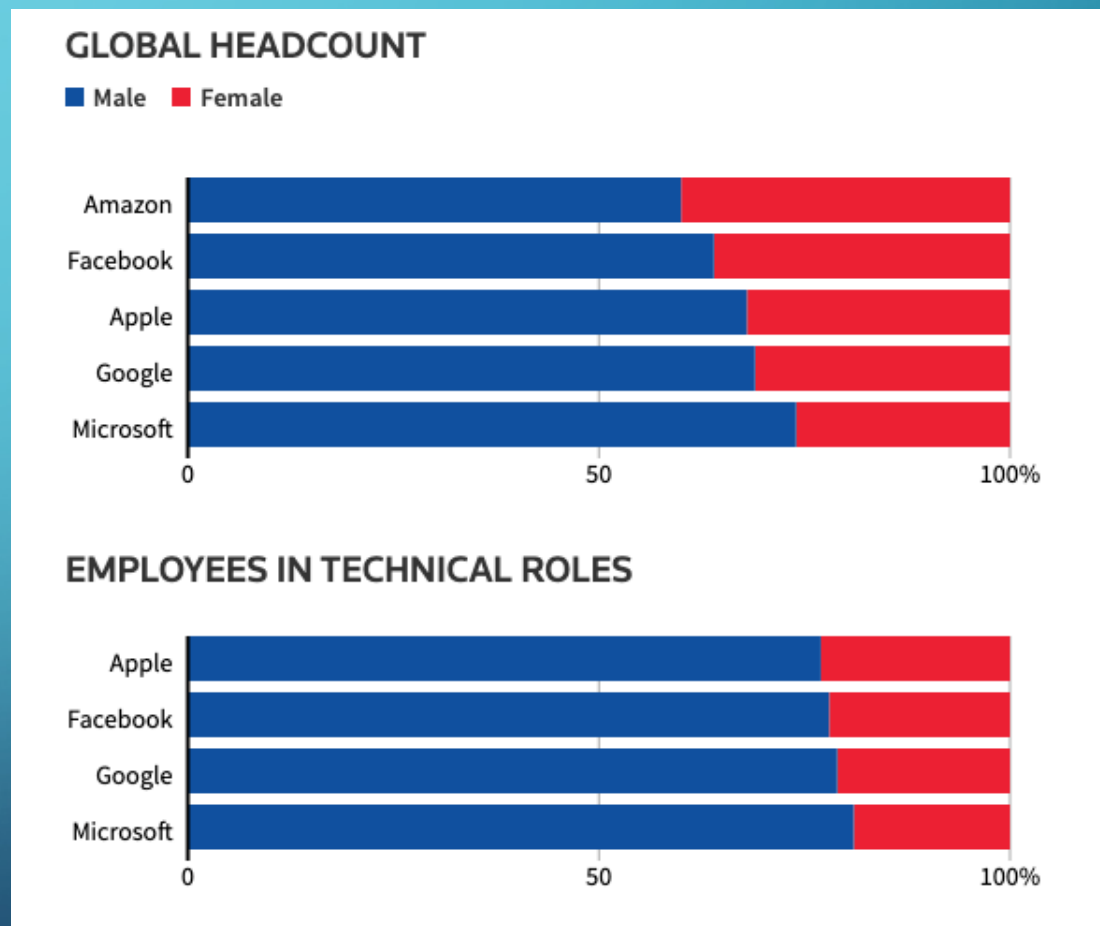
*Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.*

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%



# 演算演算法強化偏見及刻板印象

- 2018 年，亞馬遜放棄了其人工智能簡歷篩選工具，因為它對女性的偏見。
- 亞馬遜的計算模型是基於過去十年提交給公司的簡歷，但大多數申請人都是男性，這反映了男性在整個科技行業的主導地位。

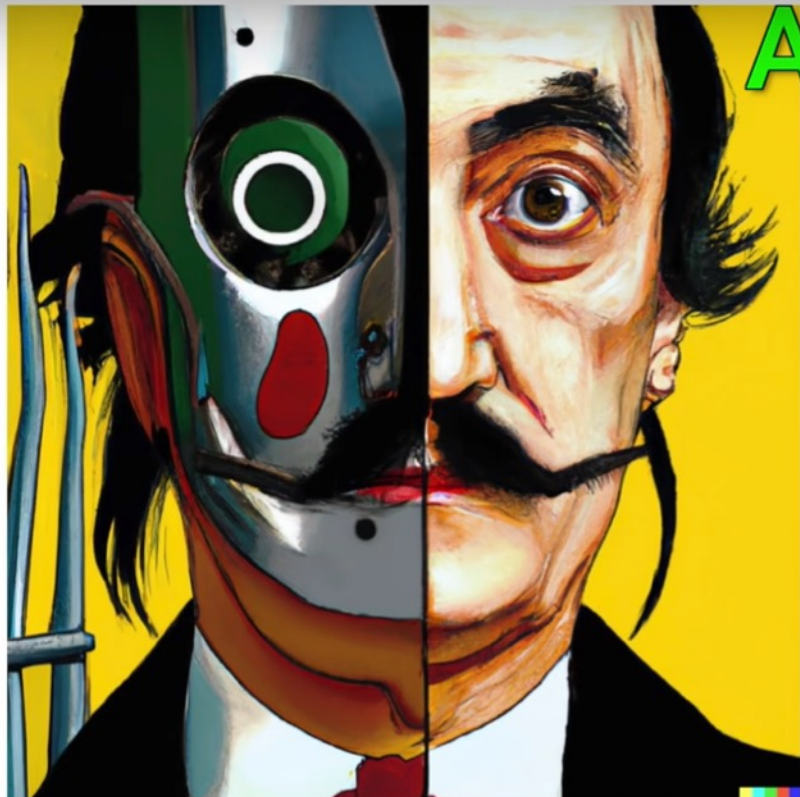


# 演算法強化偏見及刻板印象

- 2021 年 1 月，OpenAI 推出了能夠製造逼真圖像的 AI 系統 DALL-E。
- 2022 年 4 月，其第二版 DALL-E2 的巨大改進震驚了世界。
- DALL-E2 從互聯網上收集了 6.5 億張圖像，學習如何根據不同的特徵去繪畫。
- 用戶只需將描述輸入系統（例如，“畫一個像碧姬芭鐸的法國女孩”），然後DALL-E2就可以根據輸入創造出栩栩如生的圖像。
- [https://www.youtube.com/watch?v=X3\\_LD3R\\_Ygs](https://www.youtube.com/watch?v=X3_LD3R_Ygs)



# AI-generated images



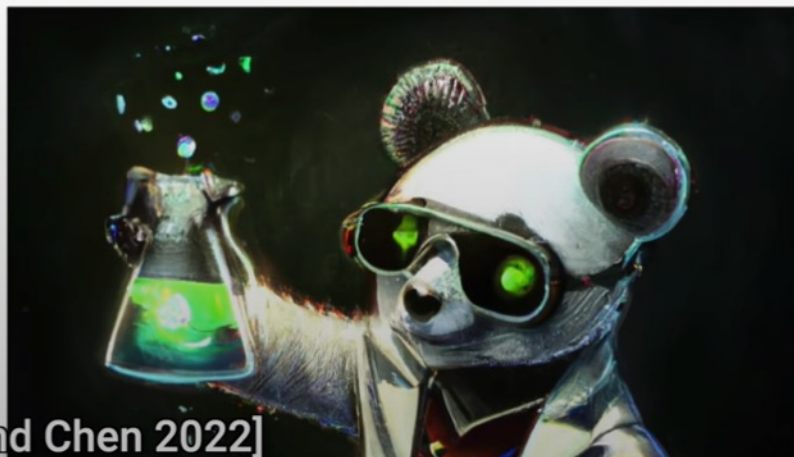
vibrant portrait painting of Salvador Dalí with a robotic half face



a shiba inu wearing a beret and black turtleneck



a close up of a handpalm with leaves growing from it







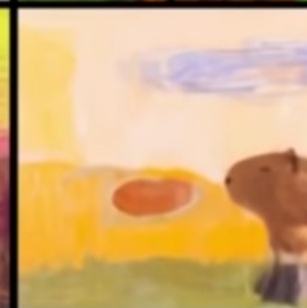
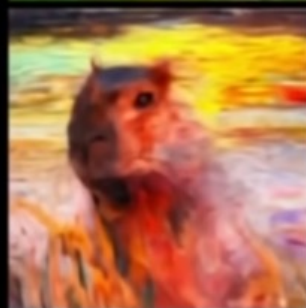
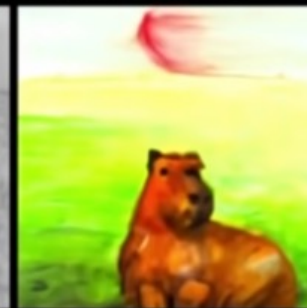
painting in the style  
of claude monet

drawing

charcoal drawing

crayon drawing

chalk drawing



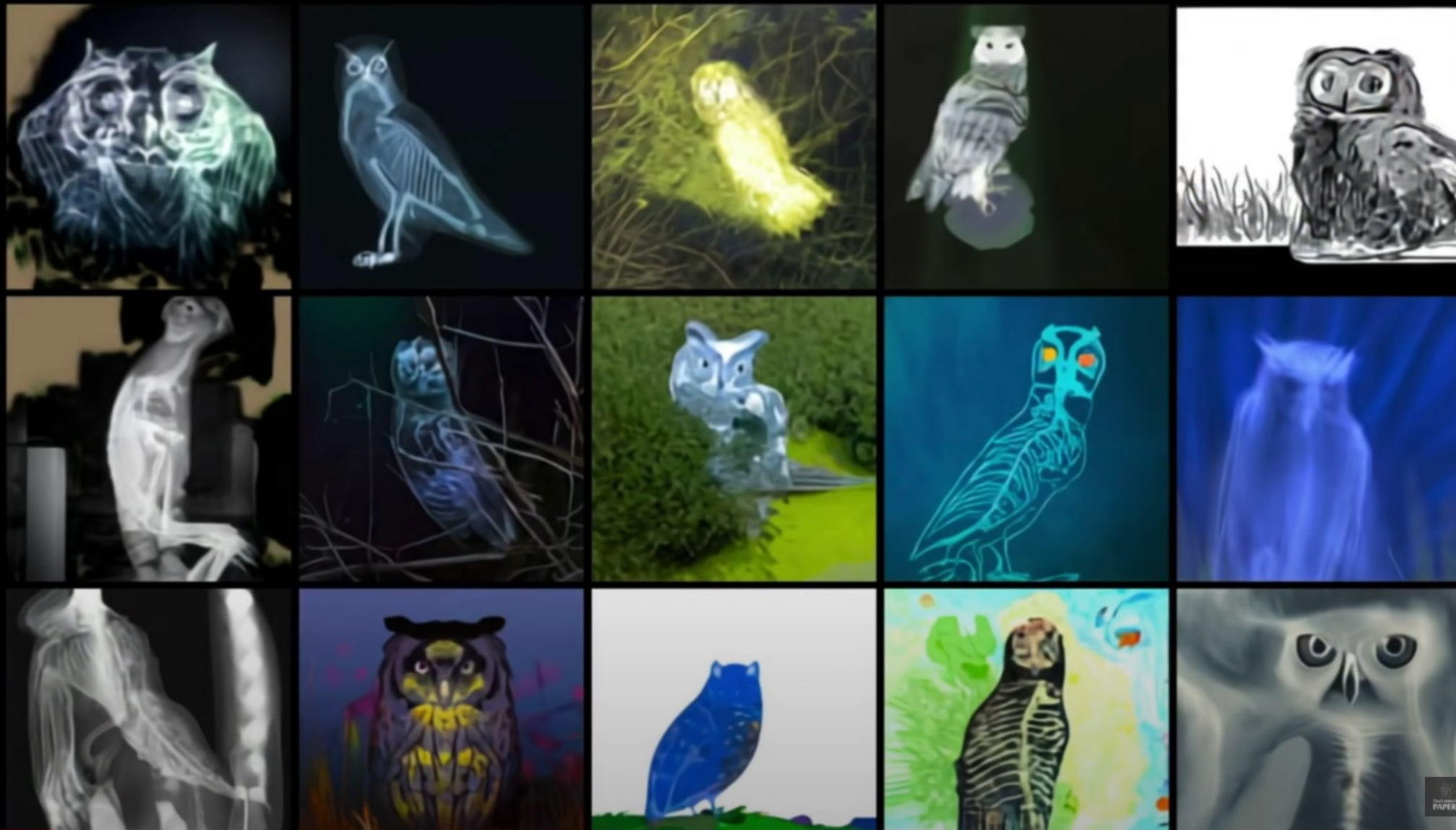


TEXT PROMPT

an x-ray of an owl sitting in a field



AI-GENERATED  
IMAGES



Source: OpenAI



**give her a mohawk**



**Source: [Ramesh, Dhariwal, Nichol, Chu and Chen 2022]**



# 1. A panda mad scientist mixing sparkling chemicals



Source: [Ramesh, Dhariwal, Nichol, Chu and Chen 2022]

# 演算法強化偏見及刻板印象

- DALL-E2 傾向於強化刻板印象，例如，當用戶要求 DALL-E2 創造一張律師的照片時，典型的輸出是中年白種男人的照片。
- 年齡歧視？種族歧視？性別歧視？反映現實？





# 演算法強化偏見及刻板印象

- 如果要求的是空姐的照片，典型的結果是漂亮的年輕女子。



## 結論

- 人工智能的演算法本來有積極的效果：群眾的智慧。
- 正如馬斯克所說，言論自由是有效民主的基石：百花齊放、百家爭鳴。
- 人性的弱點：跟風效應。
- 社交媒體的演算法造成了迴聲室效應。
- 有些人工智慧演算法出現偏差，是由於人為錯誤和疏忽，例如荷蘭探測福利欺詐的軟件、美國預測犯人再度犯罪風險的**COMPAS**，但沒有檢討計算結果和真實情況的巨大落差。
- 有些人工智慧演算法產生偏見和刻板形象，是由於「反映現實」，例如大多數從事高科技工作的人是男性，大部分空中小姐是女性。

## 結論

- 對別有用心在社媒上操控別人，或者在軟件設計上疏忽是需要批判的。
- 但使用激進的方法來應對無心之失，或將任何偏見歸因於邪惡意圖，將會適得其反。
- 在發佈DALL-E 2 之前，OpenAI 已經邀請了 23 名外部研究人員來盡量識別系統中的缺陷和漏洞。儘管做出了這些努力，但刻板印象的問題仍然存在，因為機器學習演算法會尋找現有的例子。
- OpenAI 研究人員試圖修改系統，但任何新的解決方案都會導致新的問題。例如，當研究人員試圖從訓練數據集中過濾掉色情內容時，DALL-E2 產生的女性圖像較少。結果，女性在輸出圖片中的代表性不足。



## 結論

- 要求 100% 無偏差的系統與期望 100% 無錯誤的電腦程式是同樣不切實際。一方面，研究人員應該盡最大努力減少偏見並儘可能多地修復錯誤，
- 斯坦福大學研究員托馬斯·索維爾 (Thomas Sowell)：「沒有解決辦法。只有權衡取捨。」 (There are no solutions, only trade-offs)

