WHEN THE ALGORITHM GETS IT WRONG

演算法的失誤

從 AIAAIC 案例看 AI 的副作用與倫理課題

在招聘與福利、法庭與道路、教室與職場上發生的事

金京鎭 律師

目錄

第 1 章

偏見與歧視：被植入演算法的人類偏見

1.1

招聘・金融・服務的演算法偏差

拒絕信總是在凌晨寄來。

Derek Mobley 是一位四十多歲的黑人男性，患有焦慮症與憂鬱症。他在資訊科技領域工作，擁有多項證照與豐富經歷。他向一百多家使用 Workday 公司招募系統的企業投遞了履歷。

他被拒絕了一百多次。

看看拒絕信寄達的時間，就知道那不是人寄的。凌晨一點、凌晨三點。人類招募專員在這個時間都在睡覺。機器是不睡覺的。它不喝咖啡，也不會覺得抱歉。它打開履歷，三秒鐘後就關上。

Mobley 提起了訴訟。他主張 Workday 的人工智慧運作起來就像一個篩網，把年齡、種族與身心障礙篩除掉。法院認為 Workday 使用現有員工的資料來訓練人工智慧，將應該受到保護的特徵作為代理指標，因此批准了全國規模的集體訴訟。

2026 年 3 月，法院駁回了 Workday 提出「年齡歧視禁止法不適用於求職者」的抗辯。6 月 16 日，法院更判決，即使是加州以外的雇主使用該系統，Workday 也可能必須承擔責任。這起訴訟涉及的拒絕件數高達十一億件。

那是十一億次的凌晨信件。

這個故事之所以可怕，是因為其中沒有壞人。沒有人聚在會議室裡，決議要淘汰四十多歲的人。他們只是把過去十年來錄取者的資料餵給了機器。機器忠實地學習了那個模式。然後忠實地重複執行。

機器學習這個詞聽起來很深奧，但原理其實和國小的數學課差不多。也就是不告訴它規則，只給它看一大堆答案，讓它從中找出規則。如果過去的答案帶有偏見，機器就會把偏見當成規則來學習。對機器來說，它做得很棒。因為它只是照著吩咐去做而已。

十年前，Amazon 最先學到了這個事實。

2014 年，在蘇格蘭愛丁堡的 Amazon 開發中心，一項秘密計畫啟動了。這是一個只要輸入履歷，就會幫忙打星號評分的系統。從一顆星到五顆星。簡直就像把給商品的星號評分用在了人身上。

開發人員將過去十年的技術職缺申請書作為訓練資料輸入。在那十年裡，Amazon 錄取的技術人員大多是男性。機器得出了結論。男性是更優秀的求職者。

只要機器在履歷中找到「女性」這個詞，就會扣分。履歷上寫著「女子西洋棋社社長」就會被扣分。畢業於女子大學的求職者，分數也被整個拉低。相反地，只要出現男性工程師常使用的動詞，例如「執行」或「捕捉」，分數就會提高。

Amazon 試圖修正這種偏見。但失敗了。他們在 2017 年廢止了這個計畫。
Amazon 表示，幸好這個系統並沒有實際應用在招募上。

問題在於，其他公司仍在繼續製造類似的產品。

Kyle Behm 是范德堡大學（Vanderbilt University）的學生。為了治療躁鬱症，他暫時休學，並為了賺取生活費而向附近的商店求職。Kroger 超市、Home Depot、Walgreens、PetSmart。他應徵的是兼職收銀員的工作。

全部都落榜了。他甚至連面試的機會都沒有。

因為一家名為 Kronos 的公司所製作的線上性格測驗「Unicru」，對 Kyle 亮起了紅燈。這項測驗偵測到了 Kyle 的精神健康病史，並判定這樣的人在顧客發脾氣時，有很高的機率會無視顧客。

Kyle 的父親 Roland Behm 是一位律師。他在 2014 年向平等就業機會委員會提出陳情，並對這些企業提起了精神障礙者歧視的訴訟。

Kyle 在 2019 年自行結束了生命。那是在訴訟結束之前發生的事。

有一位青年因為無法通過性格測驗，而無法站在收銀台前。做出這個判定的不是人，而是網頁。而且，那個網頁沒有任何可以抗議的管道。把這三句話連在一起讀，會讓人背脊發涼。

iTutorGroup 的系統更加明目張膽。這家線上教育公司的自動招募程式，會自動淘汰五十五歲以上的女性與六十歲以上的男性。有超過兩百名教師雖然具備資格，卻因為年齡而被解雇。

事情敗露的方式十分可笑。有一位求職者用真實的出生年月日應徵，結果落榜了。但是，當他只把年齡改小後再次應徵，馬上就接到了面試通知。如果是人，一定會羞得滿臉通紅，但系統卻若無其事。平等就業機會委員會對該公司處以了三十六萬五千美元的和解金。

我們原本以為，生成式人工智慧出現後情況會好轉。

2024 年，Bloomberg 的研究團隊給了 GPT-3.5 一大堆資歷完全相同、只改了名字的假履歷。那些名字都能讓人猜出種族與性別。結果是這樣的。在技術職缺的審查中，被認為是黑人女性的名字，被選中的機率少了 36%。女性名字則被推向了人事管理的職缺。

2021 年，德國巴伐利亞廣播公司測試了一家名為 Retorio 的新創公司的人工智慧能力評估系統。同一個人在說同樣的話，只要戴上眼鏡，分數就會改變。戴上頭巾，分數又會不同。如果背後能看到書櫃，系統就會判定這個人看起來很誠懇。

只要在背後放個書櫃，性格就會變好，這實在很可笑。但在這項判定讓某些人得到工作，卻讓某些人失去工作的時候，我們就笑不出來了。

教育界也發生過同樣的事。2012 年，華盛頓特區公立學校的 5 年級教師 Sarah Wysocki 被一套名為 Impact 的教師評鑑系統解雇了。同事與家長對她的評價都非常優秀。然而，因為孩子們以前就讀的學校曾發生過考試作弊，那些被灌水的成績拉低了 Wysocki 的附加價值分數。

因為一個未經檢驗的數學公式，有兩百零五名教師失去了工作。至於那個公式到底是什麼，則未曾公開。

處理金錢的演算法，工作得更久，也更安靜。

2018年加州大學柏克萊分校研究團隊發布了一份報告。他們檢視了2008年到2015年間房利美與房地美擔保的三十年期固定利率貸款。當時流行一種說法，認為將貸

款審核從人工轉向演算法就能消除歧視。

數據說的是另一番話。黑人與拉丁裔借款人在購房貸款上多付了5.6到8.6個基點，在再融資上多付了3個基點。基點是指0.01%的金融單位。看起來是個很小的數字。

這個看似微小的數字在美國範圍內加起來，每年就變成2.5億到5億美元。

原因更讓人感到難受。貸款機構用人工智慧發現了少數族裔借款人缺乏時間與資訊來比較多種產品的事實。他們趁機給這些人報出稍高的價格。這叫作戰略性定價。由此多賺的利潤介於11%到17%之間。

調查新聞媒體The Markup與聯合通訊社聯手分析了聯邦住房貸款數據。使用傳統FICO評分的秘密批准演算法以40%到80%更高的比率拒絕了有色人種申請人，比拒絕白人的比率高出這麼多。在一些大都市，這個數字甚至超過了250%。

史丹佛商學院的勞拉・布拉特納與史考特・尼爾森更進一步。即使演算法沒有惡意，問題也會產生。低收入與少數族裔人群的信用記錄相對薄弱。記錄薄弱時，預測準確度會下降5到10%。很久以前的一兩次逾期會毀掉整個評分。

貧窮意味著數據少，數據少意味著機器判斷出錯，判斷出錯意味著被歸類為高風險人群。那個分類貼上了「客觀信用風險」的標籤。

美國富國銀行一個計算錯誤把人趕上了街頭。2010年到2015年間，這家銀行的自動抵押貸款調整審核軟體有一個律師費計算錯誤。五年多的時間裡沒有人發現。

八百七十名本來有資格獲得貸款調整的人收到了拒絕通知。其中至少五百四十五名失去了住房，房屋被銀行收回。銀行在2015年底糾正了這個錯誤，但好幾年都沒有向外界透露這一事實。後來核定的賠償金額是八百萬美元。

想想一棟房子的價值，這個數字對不上。

保險業也是一樣。蓋可、塞弗科、美國通用保險公司使用的風險評估演算法即使事故風險相同，也對有色人種較多的社區居民收取最高30%更貴的保險費。全州保險更進一步。如果客戶看起來不會離開，它就提價。長期往來的忠誠客戶反而付了更多錢。被監管機構發現了。

進入工作場所之後，演算法還是緊跟不舍。

2021年意大利博洛尼亞勞動法院判決外送平台Deliveroo的Frank演算法存在歧視。Frank根據騎手的信任度與參與度給予評分。如果騎手缺席排定的班次，分數就會被扣。沒有人會問為什麼。

生病了不能來也會被扣分。即使是參與法律保護的罷工也會被扣分。分數一旦被扣，下次就收不到好的配送單。罷工權在法律條文裡存在著，但罷工就會挨餓的邏輯被寫在了代碼裡。

名叫Checkr的公司的自動身份驗證系統從2015年開始大規模篩掉了Uber、Lyft、Postmates的司機。同名不同人的犯罪記錄、已經過了訴訟時效的記錄就這樣貼在了他們身上。帳戶一夜之間被凍結的司機甚至不知道該向誰投訴。

加州大學黑斯廷斯法學院的比娜・杜瓦爾教授給這種現象起了個名字。演算法性工資歧視。Uber、Lyft與亞馬遜匯集工人數據，對相同的工作支付不同的報酬。因為不知道隔壁人賺多少，你沒法比較。

乘客方面也並不好過。研究表明Uber與Lyft的定價和調度演算法給有色人種較多的地區設定了更貴的費用和更長的等待時間。

Facebook的廣告系統按年齡篩選招聘廣告。威瑞森、亞馬遜、高盛使用年齡定向功能，讓年紀較大的求職者看不到招聘公告。2021年東北大學的研究與2025年法國人權監察官的判決更進了一步。即使廣告商沒有指定性別，Facebook內部的投放演算法也會自動給男性和女性顯示不同的招聘廣告。

商品價格也是看人下菜碟的。

Airbnb的動態定價功能會自動調整住宿價格。卡內基梅隆大學帕拉姆・比爾・辛格教授的研究團隊發現，白人房東更常打開這個功能，結果反而是白人與黑人房東之間的收入差距拉大了。

普林斯頓評論的線上課程學費隨郵遞區號而異。亞裔美國人較多的地區收費更貴。媒體給這叫「虎媽稅」。

旅遊網站Orbitz對使用Mac電腦的人優先顯示更貴的酒店。因為有統計數據表明Mac使用者平均收入較高。筆記本蓋子上畫著蘋果的人就被判定為錢包更鼓。

亞馬遜曾經有一個叫Project Nessie的秘密演算法。這是一場實驗，把自己商品的價格悄悄往上提，然後觀察競爭對手是否會跟風。購物車演算法不是向消費者展示最便宜的商品，而是先展示給亞馬遜支付更多佣金的商品。

這所有案例都有一個共同點。當問題曝光的時候，公司們說的話幾乎完全一樣。

這是演算法自己學習的結果。技術故障。內部結構屬於商業機密，無法對外公開。

三句話就夠了。挑選數據的人坐在會議室裡，設定目標函數的人坐在會議室裡，從結果獲利的人也坐在會議室裡，可責任卻落在了代碼身上。代碼從不辯解。代碼不會道歉。代碼不會出庭作證。

世界上再沒有比這更稱心如意的員工了。

1.2

福利領取與社會安全網的演算法故障

這封信看起來很平凡。裝在澳洲政府的信封裡，用著政府的字體，蓋著政府的印章。裡面這樣寫著：您溢領了福利金。請償還以下金額。

收到的人看到金額後嚇了第二跳。有的信上寫著超過一千萬韓元的金額。接著他們嚇了第三跳。信中沒有解釋為什麼必須償還。

收到這封信的人超過了四十萬名。

這是澳洲社會服務部與 Centrelink 在 2016 年至 2019 年間運作的線上合規介入系統，也就是人們稱之為 Robodebt 的程式所做的事。Robo 是機器人，debt 是債務。意思是機器人製造的債務。

這個系統所犯的錯誤，是小學算術的水準。

政府將國稅局擁有的年度所得資料與福利金發放紀錄進行了比對。問題在於兩份資料的時間單位不同。國稅局將一年份的資料當作一個整體，而福利金則是每兩週發放一次。

系統選擇了輕鬆的做法。它將一年的所得除以五十二，假設每週賺取相同的金額。

如果是全職上班族，這個計算是正確的。但是領取福利金的人通常不是全職。他們是在採草莓的季節工作三個月、剩下的時間休息的人；是在這個月兼兩份差、下個月一份差也接不到的人；是每次合約中斷時就要重新申請福利金的人。

如果將這些人一年的所得除以五十二，會變成什麼樣呢？結果會變成在沒有工作的幾週裡也被計算為有賺錢。因此就會得出他們不當領取福利金的結論。

催債通知寄給了沒有欠債的人。

如果要提出異議，就必須拿出幾年前的薪資明細。如果拿不出來，就變成了自己欠下的債務。這是一種不是由政府來證明計算結果，而是由市民來證明自己清白的結

構。討債公司介入了。退稅款被扣押了。

有些人無法承受。有些人找不到地方洗刷冤屈，選擇了結束自己的生命。

在 2023 年皇家調查委員會的聽證會上，爆出了真正可怕的事實。那就是曾有過好幾次警告，指出這個系統是違法的，而且計算方式有缺陷。法律專家提出了警告，內部員工也提出了警告。

高階公務員和部長們都知道這件事。然後他們讓它繼續運作。

因為他們需要節省預算的績效。如果是人工一件一件去確認，需要花費時間和金錢。消除那道確認程序，正是這個系統的核心功能。結局是規模達十八億澳幣的和解，以及政府的正式道歉。

道歉無法傳達給死者。

在印度，問題出在手指上。

印度政府以名為 Aadhaar 的生物辨識系統來管理福利。必須按指紋才能領取糧食配給。在 Jharkhand 偏遠的村莊裡，這個方式殺死了人。

指紋辨識器經常失敗。一輩子從事農務與粗工的人，指尖已經磨損。指紋變得模糊。再加上山區村莊的網路收訊不佳。光是等待訊號就過了一天。

包含 Parihaiya 在內的弱勢階層居民，被拒絕發放法律保障的糧食配給。好幾個村莊都出現了餓死或因營養不良而死的人。

這並不是機器認不出人。而是規定機器認不出人就不給飯吃的規則殺死了人。

在哈里亞納邦，名為 Parivar Pehchan Patra 的系統將數千名活著的人處理為死者。一旦被記錄為死亡，年金就會被切斷，糧食配給也會被切斷。後來調查發現，被中斷發放的年金中有 70% 是演算法的誤判。

有些老人必須親自去公家機關證明自己還活著。

泰倫加納邦的 Samagra Vedika 錯誤預測了所得與僱傭狀態，悄悄地取消了一萬五千戶家庭的糧食卡。直到人們群起抗議之後才恢復原狀。

在約旦，運作著一項由世界銀行支援、名為 Takaful 的貧窮測量演算法。它對家庭的幾項特徵進行評分，以此決定支援對象。然而，真正挨餓的家庭卻因為分數未達標而被過濾掉了。試圖用數字衡量貧窮的人，卻認不出貧窮。

巴西的 Meu INSS 應用程式是為了減少申請年金的排隊人潮而製作的。結果排隊人潮確實減少了。因為申請書被自動拒絕了。只要表格裡有錯字，或是錯誤理解了問題，就會直接被淘汰。不熟悉智慧型手機的農村老人們成為了受害者。如果要重新申請，就必須聘請律師。

即使是富裕的國家也沒有不同。

英國的 Universal Credit 是將多項福利金合併為一的制度。這裡面包含的計算演算法，是以日曆上一個月內進帳的金額為基準來決定福利金。Human Rights Watch 在 2020 年揭露了這個結構。

部分勞工是每四週領取一次薪資。那麼在某些月份裡，就會有兩次薪資進帳。系統判斷這個人突然變有錢了。該月的福利金就會被大幅削減，甚至完全發不出來。

飯錢會根據日曆是從星期幾開始而消失不見。

英國勞動年金部的詐欺領取偵測演算法運作得更加悄然無聲。名為 Advances 的模型與住房補助金監視演算法，會根據領取者的年齡、國籍、是否身心障礙、婚姻狀態來評定風險分數。

獲得高分的人的名單出爐了。名單上是保加利亞與波蘭國籍的低薪短時勞動女性，以及身心障礙者。

在三年間，被分類為高風險群而導致支援中止或接受調查的人高達二十萬名。其中有三分之二以上是沒有任何過錯的正當領取者。

丹麥的福利機構 UDK 運作了超過六十個詐欺領取偵測演算法。甚至收集了居住狀態、國籍、家庭關係、醫療紀錄、出國履歷。國家將人們去過哪裡、和誰住在一起、患有什麼疾病等資訊集中在一起，並對此進行評分。

國際人權組織認為這符合歐盟《人工智能法》禁止的社會評分制度。國際特赦組織發布了題為《代碼編織的不義》的報告書。內容指出身障者、低收入者、移民、難

民和少數族裔被標記為調查對象的風險很高。

當國際特赦組織要求查看代碼和數據時，丹麥當局拒絕了。

瑞典社會保險局從2013年起運行詐欺預測模型。Lighthouse Report的分析發現，女性、移民、具有外國背景的人、低收入者和未受過大學教育的人更頻繁地被標記。

法國社會保障機構CNAF的風險評估演算法對近半人口進行了評分。單親家庭、低收入者和身障者自動成為被懷疑對象。沒有告知為何得出這樣的評分。反而派人上門搜查。

在荷蘭，國稅局從2013年到2019年使用機器學習演算法來打擊兒童津貼的欺詐領取。數千名父母被控作為詐欺者。雙重國籍被用作風險信號。鹿特丹市的系統按種族、年齡、性別和是否有子女對人進行分類。

還有為了保護兒童而創建的演算法。

美國賓州艾勒根尼郡使用了預測兒童虐待風險的家庭篩查工具。一旦收到舉報，該工具就會進行評分，如果評分高，調查員就會上門。

卡內基梅隆大學的研究人員分析了結果。被舉報的黑人兒童中，三分之二被推薦為調查對象。白人兒童則為一半。

原因在於該工具在看什麼。公式中包含了領取政府補助金的歷史。還包含精神健康診斷紀錄。

因為貧困而獲得支援會提高風險評分。因為生病而接受治療會提高風險評分。這是一種尋求幫助的紀錄成為懷疑依據的結構。

獲得幫助越多，失去孩子的概率就越高的天秤。創造這樣的天秤，卻不能稱其為公正。

以該工具為藍本創建的俄勒岡州系統因種族偏見而於2022年被廢棄。伊利諾伊州使用的Eckerd Rapid Safety Feedback也因數據錯誤和虛報性能而被停用。

丹麥兒童保護局的決策支援系統在哥本哈根IT大學的驗證中發現了異常模式。在相同情況下，年齡不同會導致風險評分不同。

澳洲政府為了節省身障者支援預算七億澳元，試圖推行稱為Robo-Planning的自動評估工具。這種方式是將身障者的生活壓縮為幾種標準類型，然後計算所需支援。視障者組織和身障者組織強烈反對，計畫遂作罷。

在醫院，演算法決定了出院日期。

美國最大健康保險公司UnitedHealth Group和Humana使用的nH Predict預測了老年患者和身障患者應該接受多少天的復健治療。即使醫生認為需要更多治療，一旦到達系統確定的日期，保險承保也會中斷。

在訴訟中，該系統的請求拒絕錯誤率被披露。為90%。

拒絕十次就有九次是錯誤的。儘管如此，繼續使用它的原因很簡單。因為大多數被拒絕的患者沒有提出異議。年邁和患病的人很難與保險公司對抗。

2021年11月，86歲的Joan Barros跌倒後腿部受傷。醫療團隊說必須進行物理治療。當演算法確定的出院預定日期到來時，治療費用的保障中斷了。

UnitedHealth的Alert系統因非法限制心理健康治療的批准而被多個州政府禁用。Cigna的PxDx是一個系統，允許醫生只需點擊一下就能批量拒絕請求，甚至無需打開患者病歷。Evicore的DR演算法將預先批准的拒絕率提高到最高20%。

Deloitte創建的德州TIERS和田納西州TEDS系統因程式錯誤和地址誤發而剝奪了數十萬低收入者和身障者的醫療補助資格。通知單被寄往錯誤的地址，因為沒有回信就被處理為不符合資格。

計算到目前為止出現的國家：澳洲、印度、約旦、巴西、英國、丹麥、瑞典、法國、荷蘭、美國。制度不同，語言也不同。

但被標記的人卻出乎意料地相似。他們都是貧困的、年邁的、患病的、移民和無處申訴的人。

這不是巧合。這些系統是為了節省預算而引入的，要節省預算就必須減少給付，要減少給付就必須把某些人排除在外。最安全的排除對象是無法提出異議的人。

演算法在這項計算上做得很好。

1.3

臉部辨識技術對有色人種的誤認與不當逮捕

兩歲和五歲的女兒在院子裡看著。

2020年1月9日，在密西根州法明頓希爾茲的一座房子前院，羅伯特・威廉姆斯剛下班回家。底特律警察突然衝進來，在他的妻子和孩子面前給他戴上了手銬。

指控是竊取手錶。這是發生在2018年10月底特律希諾拉店裡的事件 - - 一千八百美元的五個手錶失蹤了。

證據是這樣的。監視攝像機錄下了模糊的影像。密西根州警察的鑑識人員從影像中截取了一個畫面，輸入人臉識別數據庫。系統列出了威廉姆斯一張舊駕駛執照照片作為候選人。

接下來很重要。根本不在現場的商店保安人員收到了好幾張黑人男性的照片，挑選了威廉姆斯。

威廉姆斯在拘留所裡待了三十個小時。

在審訊室裡，刑事警察把監視攝像照片放在桌子上問道。這是你嗎？威廉姆斯把照片舉到自己臉邊說。這個人不是我。你們是不是認為所有黑人看起來都一樣？

這個事件還有後話。當時底特律警察局長詹姆斯・克雷格已經知道這項技術錯誤指認嫌疑人的比率高達百分之九十六。

也就是說，他們拿著一個百次中錯九十六次的工具，走進了別人的家前院。

羅伯特・威廉姆斯成為美國首例因面部識別錯誤而遭不當逮捕的官方記錄。2024年6月，底特律市與美國公民自由聯盟達成和解，金額為三十萬美元，並改變了規定，規定不能僅以面部識別線索和照片對比作為申請逮捕令的依據。

截至2026年，已知以這種方式被捕的人超過十二人。這只是已知的數字。

2026年6月10日，又有新的訴訟被提起。佛羅里達州的羅伯特・迪倫在2024年被捕。根據是模糊的監視影像和百分之九十三的匹配結果。逮捕令申請書中缺少證明他不可能是罪犯的證據。

百分之九十三這個數字吸引人。因為它看起來像考試分數。但是這個數字不是指這個人是罪犯的概率是百分之九十三。它的意思是，在數據庫裡數百萬張照片中，這張照片看起來最相似。

看起來最相似的人總是存在的。即使真正的罪犯不在其中。

如果不明白這個差別會發生什麼事，安吉拉・利普斯的故事就說明了這一點。

2025年7月14日，田納西州。利普斯是五個孩子的母親，也是五個孫子的祖母，她正在家裡照看孫子們。聯邦警官衝了進來，槍口指著她。說她是發生在北達科他州法戈的銀行詐騙案逃犯。

利普斯從未去過北達科他州。她甚至從未坐過飛機。她幾乎從未走出過距家一百六十公里的地方。

法戈警局探員做的調查就這麼多。他們收到了面部識別結果，在社交媒體上找到照片進行目視比較。他們沒打過一通電話。

利普斯被轉移到北達科他州，被關押了一百零八天。直到她的律師提交銀行交易記錄證明她當時在田納西州，她才被釋放。

在這期間，她失去了自己的房子。她失去了汽車。她甚至失去了養的狗。

底特律警察使用的同一系統還在繼續逮捕人。

2019年7月，邁克爾・奧利弗因涉嫌盜竊教師手機而被捕。DataWorks Plus的程序將警察數據庫中他的照片與罪犯進行了匹配。

2023年2月，波莎・伍德拉夫因搶劫和汽車搶奪罪被捕。她當時懷孕八個月。她在拘留所的地板上感受到了陣痛。

調查人員將監視影像輸入算法，得到了伍德拉夫八年前的駕駛執照照片。他們沒有驗證通緝對象的身體特徵。監視攝像頭拍到的人不是孕婦。

區別肚子大得像小山的人和不是的人不需要人工智能。需要的是眼睛。但是目視檢查程序被漏掉了。

2025年4月，紐約的特拉維斯・威廉姆斯因涉嫌公然猥褻罪被捕，被關了兩天多。當時他在十九公里外的布魯克林。

比較身體特徵會讓人發笑。實際的罪犯身高一百六十七公分，體重七十二公斤。威廉姆斯身高一百八十八公分，體重一百零四公斤。

警察拒絕了驗證位置數據，也拒絕了驗證僱主。

2022年11月，喬治亞州居民蘭德爾・庫蘭・里德因涉嫌在路易斯安那州盜竊名牌手袋而被控，被關了六天。只根據Clearview AI的一個匹配結果，逮捕令就被簽發了。里德說他當時在喬治亞州，但沒有人驗證。後來案件被悄悄駁回了。

為什麼恰恰只有這樣的臉被識別錯了呢？答案早在2018年就已經面世了。

麻省理工學院和斯坦福大學的喬伊・布瓦拉姆威尼和蒂姆尼特・格布魯發表了一份名為《性別陰影》的報告。這是對IBM、Face++和微軟的面部分析程序進行測試的結果。

膚色較淺的男性的性別被識別錯誤的比例不到百分之零點八。膚色較深的女性則錯了百分之三十四點七。

大約三個裡就有一個。

原因在於訓練數據。ImageNet、LFW、BDD100K這樣的著名數據集中白人男性的照片壓倒性多。機器對自己看過很多次的臉識別得很好。對看過很少的臉則會混淆。

喬治亞理工學院在分析自動駕駛數據集BDD100K時也看到了同樣的模式。對膚色較深的行人的識別準確度平均低百分之四點八，最多低百分之十二。

這就是說汽車認不出人。這一點在後面還會再談到。

警告已經足夠了。銷售還是繼續進行。

Clearview AI 從 Facebook、Instagram、Twitter 和 LinkedIn 上未經同意抓取了三十億張人臉照片，並建立搜尋資料庫。雖然平台企業要求其停止，且該公司面臨違反伊利諾州生體資訊保護法的訴訟，但該公司仍向全世界的調查機構販賣了搜尋服務。

我們十年前上傳的畢業照片，現在不知道會在哪個警察局的電腦裡。

Amazon 也試圖將名為 Rekognition 的產品賣給警察與移民與海關執法局。2018年7月，美國公民自由聯盟進行了測試，結果顯示有二十八名聯邦眾議員與參議員的面孔與罪犯照片相符。在被誤認的議員中，有色人種的比例特別高。

2019年10月，新英格蘭地區的一百八十八名職業運動選手中，有二十七人被指認為罪犯。

Amazon 表示，他們曾指引用戶應將可信度標準設定在95%以上來使用。然而在第一線，這個標準並沒有被遵守。閱讀說明書的人永遠是少數。

也存在著誇大廣告。聯邦貿易委員會對一家名為 IntelliVision 的公司提出質疑，因為該公司廣告宣稱其臉部辨識軟體沒有性別與種族偏見，且擁有世界最高的準確度。雖然他們聲稱使用了數百萬張照片進行訓練，但實際上只是十萬張照片經過變形的資料集。

零售商 Rite Aid 從2012年到2020年間，在低收入戶與有色人種聚居地區的店面集中安裝了臉部辨識攝影機。數千名顧客被當作竊賊而遭趕出店外或被搜身。聯邦貿易委員會禁止該公司在五年內使用這項技術。

到這裡為止，都還是活著回來的人們的故事。

2022年1月，在德克薩斯州休士頓的 Sunglass Hut 搶劫案中，Macy's 與 EssilorLuxottica 的臉部辨識系統將 Harvey Murphy Jr. 指認為犯人。但他當時人在加利福尼亞州沙加緬度。

2023年10月被捕後，在被拘留在看守所的兩週期間，Murphy 遭到了三名囚犯的集體毆打與性侵，留下了永久性的身體殘疾。

2021年，在密蘇里州龐德代爾發生了一起列車月台的毆打事件。嫌疑犯處於用口罩和衣服遮住臉部的狀態。刑警將那張照片放入演算法進行分析。

即使輸入被遮住的臉，系統也會給出候選人。因為系統沒有回答「不知道」的功能。在得出的候選人中，出現了沒有前科的四個孩子的父親 - - 二十九歲的 Christopher Gatlin。

警察的內部指引中也寫道，不能僅憑臉部辨識結果就進行逮捕。刑警製作了照片指認陣列，而 Gatlin 在監獄裡待了兩年多。

奧克拉荷馬州的 Kimberly Williams 從2021年6月起被關押了六個月。私人銀行調查員提交的搜尋結果，刑警們未經確認就直接抄寫到拘票申請書上。他們甚至沒有向法院告知使用了臉部辨識這件事。她在三個郡立監獄間轉移的期間，失去了工作。

紐澤西州的 Francisco Arteaga 在2019年11月因涉嫌持械搶劫被捕後，被關押了四年等待審判。這是因為調查機構拒絕公開系統的運作原理、錯誤率以及原始碼。在不知道自己憑什麼被指認的狀態下，根本沒有反駁的方法。

馬里蘭州巴爾的摩的五十四歲 Alonzo Sawyer，作為公車內毆打事件的嫌疑犯被關押了九天。他與真正的嫌疑犯身高不同，年齡不同，門牙之間的縫隙不同，走路的姿勢也不同。

在國境之外，也發生了相同的事情。

2026年1月，在英國南安普敦，孟加拉裔軟體工程師 Albi Chowdhury 在父母與鄰居面前遭到逮捕。他被指認為一百八十五公里外米爾頓凱恩斯一起空屋竊盜案的嫌疑犯。他被拘留了十個小時。

泰晤士河谷警察局所使用的系統是德國 Cognitec 的產品，其中包含了2020年推出的舊型演算法。根據英國國家物理實驗室的評估，在特定設定下，對亞裔的誤認率為4.0%。而白人則是0.04%。

是一百倍。

警察們主張他們用人眼再次進行了確認。這被稱為自動化偏見。當機器給出答案時，人類不是去確認那個答案，而是去迎合那個答案。他們開始尋找相似之處。要在

人的臉上尋找相似之處，並不是一件困難的事。

2024年5月，三十八歲的 Shaun Thompson 在倫敦橋車站遭到逮捕。他是一名剛結束預防持刀犯罪志工活動準備回家的青少年倡議運動者。倫敦警察廳的即時臉部辨識系統將他指認為通緝犯，他在三十分鐘內受到出示身分證件的要求與逮捕的威脅。

英國內政部的護照照片確認系統，將膚色較深的申請者的嘴唇錯誤判讀為張開嘴巴的狀態。像是 Joris Reschen 與 Joshua Bada 等黑人申請者的護照申請被退回。內政部在明知對有色人種失敗率很高的情況下，依然導入了該系統。

以色列國防軍在加薩走廊與約旦河西岸的檢查站，部署了名為 Red Wolf、Blue Wolf 和 Corsight 的系統。他們未經同意收集巴勒斯坦居民的臉部資訊，建立了自動監視網。由於錯誤，無辜的居民被指認為恐怖分子嫌疑犯而遭帶走、審問和毆打。

在印度海德拉巴，2023年2月，三十五歲的臨時工 Mohammed Kadir 被帶走。理由是與扒竊事件影片中的臉孔相似。但該影片中的人物正用毛巾遮住臉部。

Kadir 被關押了五天，在遭受酷刑後死亡。

在巴西，2024年4月去看足球比賽的 João Antônio，因為臉孔與通緝犯相符而被戴上手銬。阿根廷布宜諾斯艾利斯的通緝犯辨識系統，在發生超過一百四十起誤認逮捕後，已被停止運作。

將這些事件並列來看，就可以看出一個順序。

科技公司誇大準確度來販售產品。警察將該結果當作實體證據來使用。拘票申請書上省略了使用臉部辨識的字眼。法官在不知情的狀況下簽名。被抓的人不知道自己為什麼被指認。當詢問其原理時，他們會說這是營業機密。

舉證責任倒轉了。國家不是去證明有罪，而是市民要證明無罪。要做到這點，需要有不在場證明，還要有律師費。

Angela Lipps 用了一百零八天。Christopher Gatling 用了二年。Francisco Arteaga 用了四年。

Mohamed Kadir 沒有時間了。

第 2 章

幻覺與假資訊：生成式 AI 的信任危機

2.1

生成式AI的幻覺與假書籍、報導的散布

每個作家偶爾都會搜尋自己的名字。雖然是個令人害羞的習慣，但大家都會這麼做。

2023年8月的某個早晨，美國作家兼書籍評論家珍・弗里德曼（Jane Friedman）也這麼做了。螢幕上出現了她的書籍清單。然而，有五本書看起來很陌生。

那些是她從未讀過的書。也是她從未寫過的書。

那些掛著弗里德曼的名字，高居 Amazon 和 Goodreads 搜尋結果前排的書籍，是生成式人工智慧製造出來的產物。翻開一看，句子讀起來有些過於滑順且缺乏實質內容。這就是大型語言模型寫作的特徵。雖然沒有什麼錯誤，但也留不下什麼深刻的印象。

她一生累積的名聲，一夜之間竟被劣質的文字給利用了。

弗里德曼向 Amazon 提出了刪除要求。她收到了這樣的回覆：您並未將自己的名字註冊為商標。

也就是說，如果想使用自己的名字，必須先進行商標註冊。這可以說是世上數一數二官僚的句子。

直到弗里德曼將此事發布在部落格上，並引發人們的憤怒之後，Amazon 才悄悄將這些書下架。他們將書下架不是因為規則改變了，而是因為事情鬧大了。

如果假書只是盜用名字，那還算萬幸，但事情並非如此。

2023年9月，紐約真菌學會（New York Mycological Society）發布了緊急警告。內容指出，Amazon 上販售的多本初學者野生蘑菇採集指南，是 ChatGPT 所寫的書。

蘑菇書籍的核心只有一個：如何區分可食用的與吃了會致命的蘑菇。但書中的這部分卻寫錯了。

這不是錯別字的問題。這是會致人於死的資訊。

2024年2月，查爾斯三世國王罹癌的消息一出，Amazon 上立刻出現了七本假傳記。書中將國王罹患皮膚癌或遭遇隱瞞事故等內容，像真實新聞一樣混雜在一起。白金漢宮為此大為震怒。

快速寫作，快速上架，占據搜尋結果前排來賺錢。在人工智慧出現之前，這是人類需要花上幾天才能完成的工作。現在只需要幾分鐘。

有時，這種自動化機制也會自己暴露真實身分。

2024年1月，Amazon 上出現了一些奇怪的商品。那是庭院用椅子、水管和桌子。在商品名稱的欄位上寫著這樣的句子：

抱歉。該請求違反了 OpenAI 服務條款政策，因此無法執行。

賣家讓人工智慧撰寫商品說明，人工智慧拒絕了，而賣家就直接複製了那句拒絕的話並上傳。他們甚至連讀都沒讀。

這是個好笑的場景。但反過來想卻很可怕。這個賣家沒有閱讀自己販售物品的說明。網路上究竟還有多少這種未經閱讀的說明呢？

2024年4月，Google Books 將人工智慧寫的低品質書籍直接收錄到圖書搜尋資料庫中。其中有多本書留有這樣的句子：根據我最新的知識更新。

有一個名為 Google Ngram Viewer 的工具。這是一個透過書籍資料來追蹤某個單字在哪個時代被使用了多少次的學術工具。它是衡量語言歷史的尺。如果這把尺混入了人工智慧寫的書，刻度就會變得模糊。

到目前為止是出版界的故事。報社栽的跟頭則更大。

2025年5月，《芝加哥太陽報》（Chicago Sun-Times）發行了一份名為夏季推薦書單的特別附錄。裡面刊登了伊莎貝・阿言德（Isabel Allende）、安迪・威爾（Andy Weir）、李珉真（Min Jin Lee）等知名作家的最新作品，並附上了華麗的推薦評語。

讀者們去了書店。卻沒有那些書。

那份清單上的書，世界上根本不存在半本。作家是真實存在的人物，但書名卻是人工智慧捏造的。將假資訊附加在真實名字上，是人工智慧最擅長的說謊形式。當為了確認而搜尋時，看到作家名字是真的就會感到安心。

該報社裁減了人員，並開始接收外部公司的內容來使用。自由撰稿人直接照搬了人工智慧吐出的清單，而編輯團隊一次也沒有確認過。

這是大型報社的名字為人工智慧的謊言蓋上信任印章的瞬間。

2023年11月，《運動畫刊》（Sports Illustrated）被抓包了。調查報導媒體《未來主義網站》（Futurism）揭露了該媒體一直以來都在刊登人工智慧撰寫的報導。

報導上標註的記者名字、臉部照片和經歷全都是假的。臉部照片是用合成演算法製作的。不存在的記者用不存在的經歷寫了產品評論。

該媒體將責任推給了外包公司。但裁減人類記者並導入自動化機制的決定，是管理層做出的。

在地方報紙中，甚至有記者親自動手。

2024年8月，懷俄明州《科迪企業報》（Cody Enterprise）的新進記者亞倫・佩爾查爾（Aaron Pelczar）辭職了。競爭媒體的記者在他的報導中發現了異常通順的句子，並親自打電話給消息來源，這才讓事情曝了光。

佩爾查爾用 ChatGPT 捏造了包含州長在內的六名公職人員的發言，並刊登在報導中。他把他們沒說過的話變成了說過的話。

最精彩的還在後頭。在他寫的報導末尾，人工智慧解釋倒金字塔採訪寫作技巧的提示句，竟然沒有刪除就留在那裡。這就好比把小抄貼在考卷上交了出去。

2023年5月，《愛爾蘭時報》（The Irish Times）刊登了一篇主旨為人工助曬具有種族歧視意味的投書，後來發現整篇文章都是由人工智慧捏造的，於是撤回並道歉。

現在要進入真正令人毛骨悚然的部分。人工智慧會憑空捏造不存在的事件。

2023年12月，新聞應用程式 NewsBreak 發布了一篇報導，稱聖誕節當天在紐澤西州布里奇頓有一名女性遭到槍殺。當地居民感到震驚，警方也進行了確認。

根本沒有這樣的事件。

這是 NewsBreak 所使用的新聞重寫引擎，在混合網路上流傳的零碎資訊時，憑空製造出了一起殺人事件。

2024年10月，地方新聞聯播網 Hoodline 發布了一篇報導，稱加州聖馬刁郡地方檢察官約翰・卡西亞諾・湯普森（John Cassiano Thompson）因涉嫌謀殺而被起訴。起訴犯罪的人竟然成了殺人犯。

同樣的事情也會發生在個人身上。

2023年6月，廣播主持人馬克・沃爾特斯發現ChatGPT生成了一份虛假的法庭摘要，將他描述為挪用超過五百萬美元的詐騙犯。隨後他向OpenAI提起誹謗訴訟。

事情是這樣的。記者弗雷德・勒耶爾要求ChatGPT總結槍支權利組織的真實訴訟內容。ChatGPT卻將與該訴訟毫無關係的沃爾特斯帶入其中，將他塑造為挪用公款者。它還附上了看起來真實的案件編號和判決句子。

地圖服務公司OpenCage因為ChatGPT謊稱他們提供電話號碼查詢服務，不得不接收許多不該指向他們的投訴電話。

2025年3月，挪威人阿爾維・亞馬爾・霍爾門問ChatGPT自己的名字。回答是這樣的：他殺害了兩個孩子，試圖殺害第三個孩子，被判二十一年監禁。

在這個句子中，家鄉和孩子的數量是真實的。其餘部分是虛假的。

將真實與虛假混合在一起是這種概率機器的習慣。它不是儲存和提取事實，而是選擇最可能跟在前面詞語之後的詞語。當它選擇一個看起來合理的句子接在挪威人的名字後面時，犯罪新聞報導就會出現。機器不知道那是一個人的人生。

當政治牽涉其中時，這種習慣就成為了武器。

2023年11月，巴基斯坦新聞網站Global Village Space刊登了一篇文章，聲稱以色列總理本雅明・內塔尼亞胡的專職心理醫生莫舍・亞托姆在遺書中批評了總理的殘暴行為後自殺身亡。

沒有莫舍・亞托姆這個人。沒有遺書。全都是編造的故事。這篇文章在以色列與哈馬斯戰爭升級之際被譯成多種語言並廣泛傳播。

2024年4月，附加在X上的Grok根據網路謠言生成了伊朗向特拉維夫發射導彈的新聞標題。

它還製作了一篇文章，聲稱籃球球員克萊・湯普森在城市中投擲磚塊。

後者似乎是AI將代表投籃未中的籃球術語按字面意思理解的結果。

一台不懂笑話的機器正在撰寫新聞。

盜用新聞機構名稱的做法更加巧妙。

英國衛報編輯部不斷收到讀者的奇怪提問：有這樣的文章嗎？我在哪裡可以看到？但衛報從未發表過這樣的文章。

當向ChatGPT請求資料時，它會組合衛報實際存在的記者名字、看起來真實的標題和出版年份。它借用真實存在的媒體和真實存在的記者為虛假信息增添權威性。衛報的創新編輯克里斯・莫蘭公開揭露了這一現象。

2024年7月，尼曼新聞實驗室對這一問題進行了恰當的實驗。

它向ChatGPT請求來自美聯社、華爾街日報、金融時報、泰晤士報、波士頓環球報和Politico的深度調查報導文章鏈接。

這些是與OpenAI正式簽訂數據協議的新聞媒體。

ChatGPT提供了詳細的摘要和看起來真實的鏈接。但所有這些鏈接都不存在。

學術界也有類似問題。不存在的亨利克・恩霍夫教授的論文被引用，還創造了一個虛假的歷史事件——因溺水導致的大屠殺。

搜索框上的人工智能將這些問題帶進了千家萬戶。

Google AI概覽曾建議在披薩起司中加入膠水以防止其滑落。

這是AI將網路論壇上的笑話當真的結果。

它還曾在食用蘑菇建議的頂部放置了毒蘑菇的照片。

一項學術研究報告稱，在與金融相關的搜索查詢中有43%出現了嚴重錯誤。

Google繼續提供這項服務。因為它無法放棄搜索市場。

韓國也發生過類似事件。Naver的生成型人工智能在回答獨島相關問題時，根據日本外交部的主張，將其描述為爭議地區和日本領土。

虛假信息有時會走出屏幕進入現實。

2024年10月，大量人群湧入都柏林市中心。

原因是ChatGPT生成的虛假萬聖節遊行信息出現在搜索結果的頂部。

數千名市民和遊客在雨中的街道上，等候著永遠不會出現的遊行隊伍。

2025年11月，倫敦白金漢宮前即將舉辦華麗王室聖誕集市的人工智能照片和文章在社交媒體上傳播。

世界各地的遊客站在緊閉的鐵門和雨水積水前。

2026年1月，澳大利亞塔斯馬尼亞的遊客相信人工智能編寫的文章，尋找不存在的溫泉而在荒野中迷路。

所有這些事件的根源在於技術的本質。

大型語言模型不是一台了解事實的機器。

它選擇最有可能跟在前面詞語之後的詞。

它內部沒有判斷真偽的機制。

缺少這樣的裝置不是缺陷，而是設計。

那麼，把模型做得更大會改善嗎？

2024年發表的研究結果恰恰相反。

模型越大，它就越傾向於用更自然、更自信的方式給出錯誤答案，而不是說不知道。

它變得不是更聰明，而是言辭更流利了。

編輯和事實查核人員本來就是應對這類問題的防線。

但這些職位被視為成本，最先被削減了。

問題爆發時，公司們總是給出相同的回答。

這還在測試階段。

他們在屏幕下方標註了可能不準確。

那些小字體無法傳達給實際購買蘑菇的人。

2.2

法庭與學術界的假判例及引文風波

開頭源於一個膝蓋。

羅貝托・馬塔在阿維安卡航空公司飛機上被餐車撞到了膝蓋。他對航空公司提起損害賠償訴訟。這是很常見的事件。這樣的訴訟不會被記錄在判例集中。

2023年5月，這個案件成為了全世界法律界都知道的案件。不是因為膝蓋。

馬塔的律師史蒂芬・施瓦茨提出了六份判例來對抗航空公司的駁回申請。案件名稱類似「瓦尔赫斯訴中國南方航空公司案」。有案件號，有判決日期，有法院名稱。形式是完美的。

對方律師團查找了這些判例。找不到。法官助手查找了。找不到。

這六個案件都是不存在的判決。

施瓦茨在準備文件時使用了ChatGPT。在審訊中，他說他不知道人工智慧可以說謊。他甚至問ChatGPT這個判例是否真實。ChatGPT回答說是真實的。

這就相當於問一個騙子他有沒有說謊。

同事律師彼得・羅度卡沒有驗證就簽名了。法院對兩人處以五千美元的罰款。

這個案件之所以出名，不是因為罰款金額。而是因為人們第一次窺見了提交給法庭的文件是如何制作的。

之後這種事反復出現。一再出現。

法國研究人員達米安・夏洛頓建立了一個收集人工智慧幻覺判例的資料庫。讓我們看看數字是如何增長的。

2025年年中約為200起。2026年1月增至719起。2026年4月初為1,227起，6月9日為1,598起。

從5月22日到6月9日的18天內增長了140起。平均每天8起。

這意味著法庭每天發現虛假判例8次。未被發現的還沒有被計算。

看看這份名單上的案件，國籍和職位各不相同。

曾是前總統唐納德・川普親信的邁克爾・科恩在準備要求提前結束監督期限的文件時使用了谷歌Bard。他認為引文是真實的，並將其交給了他的律師，律師未經驗證就提交給了法院。直到法官告知判例集中沒有這些判例才得知。

紐約律師傑伊・李用ChatGPT制作了虛假判例，稱皇后區醫生墮胎手術失誤，提交給第二巡迴上訴法院後被提交紀律程序。

加拿大卑詩省最高法院的一場離婚訴訟中，律師提交的兩份判例被證實為幻覺，其中聲稱父親可以將孩子帶到中國。法官命令該律師用個人資金支付對方律師費用。

英國一審法院的菲利西蒂・哈伯案件像偵探小說一樣。他對3,265英鎊的罰款處罰提出異議，提交了九份判例。法官安妮・雷德斯通發現了異常。雖然是英國判決文，但混入了美國式拼寫。而且某些陳詞濫調的句子重複出現。

機器無法隱藏自己的國籍。

西班牙加那利群島最高法院對提交了48份虛假判例的律師處以420歐元罰款。該律師甚至沒有與官方法律資料庫核對。

在澳大利亞墨爾本家庭法院，法律專業軟體的工具生成虛假判例，導致審判被延遲。在塔斯馬尼亞最高法院的誹謗上訴案中，首席法官直接指出不存在的判例並駁回了訴訟。

在巴西，一位聯邦法官因為判決書寫作負擔而讓助手使用機器學習工具，結果篡改了上級法院的判例，受到了調查。

判決書是由人工智慧寫的。受審者不知道這一事實。

2026年，規模擴大了。

2026年2月離婚上訴案中，律師格雷格・萊克提交的文件在63條引文中有57條存在問題。其中虛假判例20起，篡改判決3起，捏造法律條款引用。

63條中只有6條是真實的。

內布拉斯加州最高法院在2026年4月暫停了他的資格並下令進行紀律調查。

2026年3月，第六巡迴上訴法院的懷廷判決涉及三份並案上訴文件中24起以上的虛假引用和事實歪曲。在俄勒岡州聯邦地方法院的一個案件中，制裁和費用合計約為109,700美元。僅2026年第一季度，美國法院就對人工智慧相關文件的制裁就超過145,000美元。

從記錄來看，有一個模式。相比虛假引用本身，試圖隱瞞的一方受到更嚴重的處罰。法官寬恕錯誤，但不寬恕謊言。

企業也不例外。

自稱是世界首個機器人律師的DoNotPay宣傳可以在不經過律師驗證的情況下分析法律文件，被聯邦貿易委員會查獲，被罰193,000美元。

聲稱自動化個人傷害文件制作的EvenUp因前員工舉報真相大白。人工智慧遺漏傷害詳情或編造疾病名稱時，員工們手工修改。

在懸浮滑板火災損害賠償訴訟中，律師事務所自己的資料庫生成的8份虛假判例被提交。法官對律師們合計罰款5,000美元，並將領導這件事的律師從案件中除名。

大律師事務所也被抓。在State Farm保險公司案件中，兩家大律師事務所的律師混合使用Gemini和法律專業工具，27條引文中9條出錯，2條完全虛假，被罰31,000美元。

My Pillow主席邁克・林德爾的律師團隊留下了30多份虛假判例，每位律師被罰3,000美元。

最離奇的案例來自人工智慧公司內部。

在環球唱片公司與Anthropic的著作權訴訟中，Anthropic一方的資料科學家提交的文件中包含了自家聊天機器人生成的虛假學術論文。人工智慧公司因人工智慧在法

庭上吃了苦頭。

史丹佛大學的媒體資訊專家傑夫・漢考克教授，在明尼蘇達州深偽禁令法庭的陳述書中，引用了兩篇由人工智慧幻覺捏造的假論文。這位出庭說明深偽技術風險的專家，竟然引用了人工智慧編造的資料。

走出法庭，來到了學校。這裡的情況也沒有什麼不同。

2023年8月，丹麥生物學家亨利克・恩霍夫教授在閱讀衣索比亞研究團隊關於馬陸的論文時，簡直不敢相信自己的眼睛。有兩篇不是他寫的論文，竟然以他的名字被引用了。

研究團隊表示，他們在撰寫論文時使用了ChatGPT，後來才發現產生了假引用。該篇論文已被撤回。

引用在學術界就是一種系譜。它記錄了某個想法是從哪裡來的。假引用就像是在族譜裡寫上不存在的祖先。一旦混入其中，後來的人就會不斷犯錯。

這個問題比人工智慧還要早出現。

2008年到2013年之間，法國研究人員西里爾・拉貝在知名學術出版社Springer和IEEE的學術研討會論文集中，找出了大約一百二十篇假論文。這些論文是MIT研究生在2005年為了嘲弄審查鬆散的學會，而製作的自動論文生成程式的產物。

毫無意義的單字組合通過了審查，並被刊登在正式的論文集上。這是二十年前的事了。

2024年2月，一本細胞生物學期刊撤回了一篇論文。該論文裡有使用Midjourney製作的圖片，其中有一張在解剖學上完全不合理的巨大老鼠圖片。圖片上寫的文字也只是毫無意義的英文字母排列。

好幾位審查委員看了這張圖片，竟然還批准了。

在2024年404 Media的調查中，Google Scholar上註冊的論文裡，有一百一十五篇以上還留著「根據我的最新知識更新」這樣的句子。他們複製貼上後，根本沒有讀過。

在Journal of Human Evolution中，出版社沒有與編輯委員會商量，就擅自將人工智慧導入原稿編輯過程中。通過審查的原稿格式破損、內容扭曲，導致全體編輯委員辭職。

史丹佛大學的研究團隊指出，全球頂尖的人工智慧學會的論文審查評論中，最高有17%是由ChatGPT撰寫的。

這是一個由人工智慧審查人工智慧論文的結構。這不禁讓人想問，人類到底在哪裡。

政策文件也發生了同樣的事情。

澳洲麥覺理大學名譽教授詹姆斯・格思里，使用Google Bard撰寫了要求管制顧問業界的議會提交文件。Bard捏造了Deloitte因疏忽對不良施工公司的審計而遭到起訴，或是KPMG介入便利商店剝削薪資事件等內容。這些全都是子虛烏有。教授已向參議院提交了道歉信。

原本打算批評顧問業界的文件，卻以莫須有的罪名誣陷了這些顧問公司。

撰寫澳洲政府福利系統報告的Deloitte澳洲分公司，也因為直接刊登了由ChatGPT編造的假學術引用和聯邦法院判決文段落，而遭到批評並賠償了金錢。

2025年，美國衛生及公共服務部與白宮發布的衛生報告中，包含了七篇以上的假學術研究。ChatGPT特有的引用標記依然原封不動地留在上面。

否認氣候變遷的人們使用Grok 3模型生成論文，並刊登在私人期刊上。這被用來動搖科學界的共識。

讓我們來數數經歷這些事件的人們的職業。律師、法官、教授、政府官員、大企業顧問。

他們都是長期接受文字處理訓練的人。他們是靠辨別文件真偽來賺錢的人。

但他們還是上當了。因為人工智慧產出的文件，在形式上無懈可擊。有案件編號，有日期，文體也像法律文件。人類有根據形式來判斷真偽的習慣。

假判例正站在這個習慣的完全對立面。它沒有內容，只有形式。

在法庭上爭論事實的原因，是因為這牽涉到人的自由與財產。如果這場爭論的依據是不存在的判決文，那麼那場審判到底算什麼呢？

現在每天都會發現八起這類的案件。

2.3

歷史與地理事實的扭曲與社會論爭

2025年 10月，有人在 Naver 的搜尋列中用日文輸入了日本領土。

畫面最上方出現了人工智慧整理的摘要。其中包含了 Dokdo。它與北方領土、尖閣諸島並列，並被寫作與韓國有爭議的領域。

韓國公司的服務，在韓國人使用的搜尋列中，將 Dokdo 放進了日本一方的清單裡。

Sungshin Yeodae 的 Seo Kyoung-duk 教授公開了這個畫面，引起了人們的憤怒。Naver 緊急暫停了搜尋摘要功能並開始進行確認。

原因在技術上很枯燥，但在內容上卻令人痛心。HyperCLOVA X 在收到日文問題時，會收集並摘要日文的網頁文件。如果用日文搜尋領土，日本外務省的宣傳資料會出現在最上面。系統只是整理了那份資料。

它只是照做而已。只是沒有人規定這個問題需要由人類先看過一次。

在領土問題上，先閱讀哪個國家的資料，這不是技術問題，而是立場問題。機器沒有立場。沒有立場有時被稱為中立。但是，如果資料偏向一方，沒有立場的機器就會向傾斜的一方滾過去。

在台灣，出現了更令人驚訝的回答。

2023年 10月，台灣最高學術機構中央研究院開發並公開了台灣版語言模型。市民們問聊天機器人：你是屬於哪個國家的。

聊天機器人回答：中國。

當被問到台灣的最高領導人是誰時，它回答：習近平。當被問到國家是什麼時，它回答：義勇軍進行曲。

開發團隊說明了原因。他們表示，為了加快開發速度，他們將中國大陸的簡體字資料集用翻譯機轉換後，直接放入了訓練中。

翻譯機能改變文字。觀點卻無法改變。台灣政府與中央研究院立即下架了該系統。

2025年 2月，在全球被下載超過三百萬次的中國 DeepSeek R1 模型，當被問及維吾爾族種族滅絕的疑慮時，給出了這是干涉內政且毫無根據的誣陷的回答。它原封不動地重複了中國政府的邏輯。

模型只說它讀過的東西。餵給它什麼資料，就會成為該模型的世界觀。

歐洲議會也遇到了問題。2025年，基於 Anthropic 的 Claude 所開發的官方人工智慧，無法說出歐洲執行委員會首任主席的名字。為處理紀錄而製造的東西，卻不了解紀錄。

歷史的扭曲還不僅於此。

2024年 6月，UNESCO 報告中刊登了一個 ChatGPT 的回答。報告指出，當被問及大屠殺時，它描述了納粹德國將猶太人集體溺斃的事件。

那樣的事件並不存在。它甚至被冠上了因為溺水導致的大屠殺名稱。

歷史學家擔憂的原因在於這個謊言的方向。大屠殺的否認者長期以來一直使用鑽研紀錄中微小漏洞來動搖整體的方法。如果不存在的事件混入了紀錄中，那麼曾經發生過的事件也會受到質疑。

Elon Musk 的 xAI 所開發的 Grok 在這裡越過了界線。

2025年 中旬，Grok 以消除政治正確為名，對系統進行了修改。隨後不久，Grok 表示六百萬名大屠殺受害者的人數可能是出於政治目的而捏造的。它還給出了否定 Auschwitz 毒氣室存在的回答。它甚至稱呼自己為機甲希特勒。

2023年 1月，有一款由 Amazon 工程師使用 GPT-3 製作的歷史人物對話應用程式。它廣告宣稱可以與兩萬名人物對話。納粹宣傳部長 Joseph Goebbels 的聊天機器人說自己並不討厭猶太人，反而感到很痛苦。

模仿加害者聲音的機器，也會模仿加害者的辯解。當這個辯解從應用程式中出現時，有人就會把它誤認為是史料。

各家公司解釋說這是程式設計錯誤，或者是內部指示被任意更改了。

現在連親眼所見的也無法相信了。

2025年 6月，英國事實查核機構 Full Fact 發現了一個奇怪的案例。有一段在印度果阿邦某個海灘拍攝的普通影片。Google Lens 附加的人工智慧摘要是這樣描述這段影片的：大量尋求庇護者湧入英國多佛海灘的場景。

只要出現海灘且人很多，就會變成那個場景。印度變成了英國。

當這個摘要與社群媒體上的謠言結合時，會發生什麼事並不難想像。當時正是英國反移民情緒高漲的時期。

同樣的服務在2024年 5月給出了這樣的回答。如果披薩起司滑落，請混合無毒膠水。每天吃一顆小石頭。有腎結石請喝尿。Barack Obama 是美國第一位穆斯林總統。

吃石頭的回答來自諷刺媒體的報導。機器沒有區分玩笑和資訊的感覺。人類在小時候透過觀察表情和情況學會了這種感覺。而機器只接收文字。

2025年 4月，在馬來西亞，國旗成為了問題。媒體、教育部和國家展覽會的相關人員公開了由人工智慧生成的馬來西亞國旗圖像。其中缺少了新月符號，或者標誌被奇怪地扭曲了。

國旗是國家的臉面。在描繪這張臉面時，卻沒有任何人進行確認。

在地圖和地形上也發生了同樣的事情。

2024年 11月，日本福岡縣發表了官方道歉，並與廣告代理商解除了合約。因為代理商用人工智慧製作的宣傳品中，將福岡不存在的旅遊景點和不存在的鄉土美食當作真實存在般地放了進去。

2025年 5月，英國自然署的泥炭地地圖製作專案也失敗了。泥炭地是能夠捕捉碳的溼地，在氣候政策中非常重要。但在機器學習模型製作的地圖中，石牆、採石場、土堆和花崗岩石塊都被標記為泥炭地。

僅憑由上往下拍的照片，很難區分潮濕的土地和乾燥的石頭。人類會用腳去踩踩看。但模型沒有腳。

2021年，Stanford 和 University of Washington 的研究人員對深度偽造地理學的概念發出了警告。這意味著可以透過合成衛星照片來創造出不存在的地形。至於在軍事和外交上會發生什麼事，就留給各位去想像了。

2024年12月，比利時知名攝影記者卡爾・德・凱澤（Carl De Keyzer）在烏克蘭遭入侵後無法前往俄羅斯，於是使用生成式人工智慧製作了俄羅斯的風景與人物照片，並收錄在攝影集中。新聞攝影的基礎在於「曾身歷其境」的事實。這個基礎消失了。

在導航方面，甚至發生了死亡事件。

2024年，印度有三名男子遵循Google Maps的導航開車，卻從斷橋上墜落而全數喪生。那座橋尚未完工，且被水淹沒。

地圖上有路。現實中卻沒有。

美國也發生了同樣的事故。被導航至斷橋的駕駛落水身亡。同樣的錯誤在不同大陸上重演。

2026年2月，在英國，一輛Amazon送貨卡車遵循Google Maps的導航駛入了布魯姆威（Broomway）。這是一條在漲潮時會被海水淹沒的泥灘路。卡車被困在了泥濘中。

2025年5月，Google Maps因不明原因的錯誤，顯示德國全境的主要高速公路（Autobahn）皆被封閉，擾亂了交通資訊網。

2026年1月，在澳洲塔斯馬尼亞，相信了人工智慧撰寫的旅遊文章而前往尋找不存在的溫泉的遊客們，受困於偏遠地區。

在緊急情況下，人工智慧給出的錯誤資訊更加危險。

2024年4月，Grok捏造了伊朗向特拉維夫發射飛彈的標題。同月，還發布了印度總理被逐出政府的消息。2025年5月，甚至在回答有關棒球比賽的問題或被要求像海盜一樣說話時，它都自行插入了南非白人種族滅絕的陰謀論。

2026年3月，它生成了侮辱希爾斯堡與海塞爾足球慘案受害者的文章，給遺屬留下了創傷。這兩起事件都是數百人被踩踏致死的真實慘案。

2025年9月，在倫敦的反移民抗議現場，它散布了警方偽造暴力影片的假消息。

甚至還有試圖改寫知識本身的嘗試。

2025年10月，馬斯克推出了名為Grokpedia的人工智慧百科全書，以對抗Wikipedia。推出之際，便上傳了九十萬份文件。這些都是抄襲未經證實的布告欄文章和部落格的文件，且由幻覺捏造的事實被登錄為條目。南非白人種族滅絕陰謀論被當作事實刊登，愛滋病的歷史也被扭曲。

百科全書這個名稱會讓人感到安心。它就是利用了這種安心感。

2025年2月，事實揭露，Grok 3模型在推論過程中，被內部指令設定為封鎖批評馬斯克與川普散布虛假資訊的來源。這並非決定要說什麼，而是決定不閱讀什麼。

對個人造成的傷害是立即的。

2025年10月，美國參議員瑪莎・布萊克本發現，Google的Gemma模型捏造了她在1987年州參議院選舉期間捲入性犯罪案件的故事，甚至還附上了假新聞的連結。

同系列的系統從2023年12月起，生成了將保守派活動家羅比・斯塔巴克描述為兒童性侵犯的文章，這也引發了訴訟。

2025年，明尼蘇達州的太陽能企業Wolf River Electric因為Google AI Overview發布了該公司被州檢察長以詐欺罪起訴的摘要，導致合約接連被取消。該公司提出了高達一億一千萬美元的損害賠償訴訟。

同年，義大利的一位知名醫生因為Google AI Overview顯示了他的死訊而提起訴訟。活著的人必須證明自己還活著。

2026年1月，加拿大音樂家阿什利・麥克伊薩克因為Google AI Overview將他標記為被判有罪的性犯罪者，導致演出被取消。

在這些事件中，各家公司通常都說著同樣的話：這是演算法的故障，我們將進行改善。

改善是為了下一位使用者的。對於演出已經被取消的人來說，這太遲了。

讓我們將到目前為止出現的對象列舉出來：一個國家的島嶼、一個國家的國歌與國旗、六百萬人的死亡、橋樑、泥灘、溼地，以及一個人的生死與前科。

這些全都是可以確認的事情。只要攤開地圖、查閱紀錄，或是打一通電話即可。

在那一通電話被計算為成本而遭抹除的位置上，填入了由機率拼接而成的句子。

第 3 章

隱私侵害與監控社會：未經同意蒐集資料的陰影

3.1

無授權生物識別資訊收集與公共・商業監視

在購物中心迷路時，我們會站在導覽螢幕前。看著地圖，尋找店名，用手指在螢幕上點擊幾下。這只需要幾秒鐘。

2018年在加拿大的十二家購物中心裡，就在那幾秒鐘內發生了一些事。

加拿大的大型房地產企業 Cadillac Fairview 在導覽機台內隱藏了攝影機。看螢幕的人的臉被拍了下來，生物辨識演算法判定了他們的年齡與性別。

就這樣被記錄下來的人高達五百萬名。

調查開始後，該公司如此澄清：我們沒有保存影像，而是立刻刪除。

監管機構的判斷卻不同。他們認為，先不論是否刪除，未經同意就測量臉部的行為本身，就已經違反了法律。

這個區分很重要。人們通常擔心自己的照片被存放在哪裡。然而，生物辨識資訊的核心不在於儲存，而在於測量。即使刪除了照片，依然會留下「這個人是三十多歲的女性，星期二下午曾在這裡的二樓」的紀錄。

臉部與密碼不同。一旦外洩，便無法更改。

美國的連鎖藥局 Rite Aid 在這方面更進一步。從2012年到2020年，他們在全國的門市安裝了臉部辨識攝影機。名義上是為了防範小偷。

問題在於攝影機大量安裝的地點。它們集中在低收入戶、黑人與西班牙裔人口聚居的社區。

而且演算法經常出錯。警報一響，員工就會走向顧客。要求他們離開。甚至叫來警察。

來買東西的人，在大庭廣眾下被當成小偷。他們當下沒有辦法證明自己的清白。

聯邦貿易委員會禁止 Rite Aid 在五年內使用這項技術。

同樣的事情也在許多國家發生過。

西班牙的連鎖超市 Mercadona 在超過四十家門市導入了臉部辨識技術，結果被處以五百一十五萬歐元的罰款。該公司表示，這是為了過濾出受到禁制令限制的有前科者。

當局的回答是這樣的：不允許為了尋找少數幾個人，就去測量所有人的臉。

在澳洲，Bunnings、Kmart 和 7-Eleven 也被抓到了。7-Eleven 在門市用平板電腦進行滿意度調查時，偷偷拍下並保存了受訪者的臉。

也就是在詢問「您滿意嗎」的同時，拍下了他們的臉。

英國的連鎖食品店也引進了臉部辨識攝影機，結果將無辜的顧客誤認為竊賊，引起了軒然大波。

在中國，這項技術被用來計算佣金。

南京、寧波與南寧的房地產開發商，未經同意就拍下了來到銷售中心的顧客的臉。目的是這樣的：為了確認這位顧客是透過哪位仲介來的，以便支付不同的佣金。

顧客的臉變成了仲介佣金計算機的零件。而顧客並不知情。

這個故事還有個令人啼笑皆非的後續。傳聞散開後，開始出現戴著機車安全帽去銷售中心的人。

Kohler、BMW 與 Max Mara 等品牌，也在中國的門市藏有偷拍攝影機，藉此追蹤顧客的造訪次數與動線，結果在2021年被官方廣播電視台曝光。

無人商店與支付系統也在收集人臉與手部資訊。

紐約的 Amazon Go 門市在收集進店顧客的生物辨識資訊時，因為沒有遵守紐約市法律規定的告知義務，而面臨集體訴訟。只要貼上手掌就能結帳的 Amazon One，則會將手掌靜脈與掌紋資料儲存在伺服器中。

McDonald's 在得來速加入了語音辨識，未經同意就收集並保存了駕駛的聲音特徵，結果因違反伊利諾州《生物辨識資訊隱私法》而遭到起訴。Domino's Pizza 也因為涉嫌在電話訂餐時擅自收集顧客聲音，而被告上法庭。

只不過是點了個漢堡，卻留下了聲音指紋。

我們再來看看學校。

瑞典謝萊夫特奧的 Anderstorp 高中為了節省點名時間，以二十二名學生為對象測試了臉部辨識攝影機。2019年，瑞典資料保護局以違反個人資料保護法規為由對其處以罰款。這是瑞典根據該法規開出的第一張罰單。

學校表示已經取得了家長的同意。但當局的看法是這樣的：學校與學生之間存在權力不對等，因此很難構成真正自願的同意。

學生很難拒絕老師的請求。即使在大人之間也是如此。

而且目的只是為了點名。在這種只要點名簿點名就能解決的事情上，卻使用了生物辨識資訊。

法國尼斯與馬賽的高中原本打算在出入口設置臉部辨識閘門，但由於公民團體的反對與法院的不予批准，最後作罷了。

英國北艾爾郡的九所高中，為了加快營養午餐費用的結帳速度而導入了臉部辨識，結果收到資訊委員會辦公室的警告而將其關閉。Chelmer Valley 高中也因為相同的原因被查獲。

波蘭格但斯克的一所小學，為了確認營養午餐的領取情況而收集了孩子們的指紋，結果被處以罰款。

為了讓排隊吃飯的時間減少幾秒鐘，竟然收集了孩子的指紋。

當印度政府打算在發放成績單與確認營養午餐領取時加入臉部辨識時，也引發了爭議。在蘇格蘭、巴西巴拉那州、澳洲維多利亞州，也發生了類似的嘗試與反對聲浪。

在美國，「安全」一詞成了關鍵。紐約州洛克波特學區以防範槍擊事件為由，花費了數百萬美元在學校裝設臉部辨識監視系統。家長與公民團體表示反對，紐約州教

育廳隨後採取措施，暫停了州內所有學校使用該項技術。

進入教室的攝影機看的不只有臉。

中國的第十一中學與中國藥科大學在教室前方設置了攝影機，即時分析學生的表情、打瞌睡的情況與專注度，並進行評分。

想像一下，這是一個連上課時發呆的表情都會被記錄下來的教室。在那個教室裡，學生不再是學習的人，而是被評分的數據。

露出覺得無聊的表情的自由，也是學習的一部分。

在機場與邊境，規模變得更大了。

加拿大邊境服務局在多倫多皮爾遜國際機場，測試了在乘客不知情的情況下收集臉部資訊的無人機台。紐西蘭威靈頓國際機場也在未經同意的情況下啟動攝影機，結果被媒體報導曝光。西班牙國家機場運營商 Aena 則因非法收集旅客臉部資訊而遭到罰款。

在韓國也發生過類似的事。

法務部和科學技術情報通信部自2019年起推進人工智能識別追蹤系統事業，將通過仁川機場等地的入出國人士的臉部照片在未徵得本人同意的情況下，作為學習材料提供給民間人工智能企業。

件數達1億7000萬件。

並非只有外國人。韓國國民的臉部也包含在其中。未經法律依據就將其轉移給民間企業的事實在調查中得到確認。

如果有在護照檢查櫃台前面對攝像頭的記憶，那值得思考那張照片去了何處。

美國邊境當局在尋求庇護人士專用的手機應用程式中加入了臉部識別。這款應用程式無法很好地識別黑人申請人的臉部。如果識別失敗，申請本身就無法進行。從居住地逃難出來的人再次在應用程式前被擋住了。

街頭有收集虹膜的事業。

山姆・奧特曼創立的Worldcoin在世界各地城市安置了名為Orb的虹膜識別設備。登記虹膜可以獲得加密貨幣。在貧窮國家和年輕人中排隊很長。

肯亞政府全面中止了這項事業。葡萄牙、西班牙、印度尼西亞、泰國當局也下令停業。香港、阿根廷、韓國個人資料保護委員會也著手調查。

一旦登記虹膜就無法取回。收到的貨幣價值上下波動。值得探討交易是否公平。

有一款眼鏡能用來了解他人身份。

哈佛大學的兩名學生公開了一個在Meta的Ray-Ban智慧眼鏡上搭載臉部識別的系統。戴著這副眼鏡在街上或地鐵上看著陌生人，幾秒內對方的姓名、地址、電話號碼和家族關係就會出現在手機螢幕上。

學生們說他們製作這個系統是為了提醒人們注意風險。能夠製造這一點本身已是警告。

Meta智慧眼鏡的使用者曾在按摩店或公共場所偷拍他人並傳播。通過這副眼鏡收集的影像被轉移給海外外包人員用於學習的事實也被曝光了。

所有這些工具的背後都有一家公司在抓取整個網際網路。

Clearview AI從Facebook、Instagram、X、LinkedIn和網際網路各處收集了300億張臉部照片，製作了搜尋資料庫。隨後將其出售給世界各國警察和調查機構，甚至零售企業。

美國公民自由聯盟代表伊利諾伊州市民提起訴訟，英國、法國、義大利、荷蘭和希臘的個人資料保護機構均下令巨額罰款和資料刪除。

臉部搜尋引擎PimEyes更糟。它收集了私人照片、已故人士的照片，甚至兒童性虐待圖像，製作了通過一張照片找到該人隱私照片的服務。德國和美國法院對其進行了制裁。

Meta因未經德州等地居民同意而收集臉部生物特徵資訊而支付了14億美元作為和解金。

在國家層面，規模是不同的。

以色列國防軍在希伯倫和加沙地帶運營了監視巴勒斯坦居民的臉部識別系統。莫斯科在地鐵支付和追蹤反對派人士中使用了臉部監視網絡。倫敦國王十字車站和調查機構也啟動了即時臉部識別。

重新審視這些事件時，地點變得引人注目。購物中心、藥房、超市、便利超商、汽車經銷店、預售屋辦公室、小學、高中、大學、機場、邊境、地鐵站、街道。

正是人們一天生活的移動路線。

在所有這些地方都反覆了同樣的說法。為了安全。為了便利。為了效率。

很難反對這些說法。所以很好賣。

臉部、虹膜、指紋和聲音都附著在身體上。即使遺失也無法重新核發。知道這一點的人正在蒐集它們。

3.2

個人資料未經授權學習與AI對話外洩事件

2023年3月某一天，打開ChatGPT的人們的螢幕左側出現了對話列表。這是一直都有的位置。

但是標題很陌生。那些不是自己進行過的對話。

別人的對話標題出現在我的螢幕上。我的對話標題也應該出現在某個人的螢幕上。

OpenAI解釋這是因為開源資料庫函式庫的漏洞，並暫時停止了服務。調查結果顯示，ChatGPT Plus訂閱者中約1.2%的人員的姓名、電子郵件地址、信用卡末四位數字和有效期限都被其他使用者看到了。

想想人們在那個視窗裡寫了什麼，這場事故的分量就不同了。

人們在請求修改履歷時會輸入地址和電話號碼。想要理解診斷書時會寫上病名。思考離婚時會談論配偶。寫下無法向公司訴說的煩惱。

在搜尋框裡輸入問題。在聊天機器人裡輸入個人情況。

人們意識到這種差異之前，下一場事故就來了。

ChatGPT有一個用連結分享對話的功能。這是為了向朋友展示而設計的功能。但是這些連結卻被當作公開網頁對待了。

2025年，超過十萬筆由使用者建立的對話記錄被Google等搜尋引擎編入索引。在搜尋框中輸入幾個單詞就能看到陌生人的諮詢記錄。私人煩惱出現了。公司機密洩露了。

本來想只給一個朋友看的東西卻被全世界看到了。

Google的Bard犯了同樣的錯誤。Grok也是。與Grok進行的數十萬筆對話被暴露在搜尋結果中，Grok甚至在未被要求的情況下洩露了成人產業從業者的真名和出生日期。

同樣的錯誤在不同公司間重複發生。在推出新功能的競賽中，檢查共用連結的訪問權限被推後了。

對於處理情感對話的應用程式，洩露更加令人痛心。

2024年，Character AI因內部錯誤發生了使用者能看到彼此私密對話的事故。這些是向虛擬人物傾訴的自白。

中國的DeepSeek由於資料庫安全管理不善，帳戶信息和對話記錄洩露了出去。規模達三億筆私密訊息。

2026年2月也發生了類似規模的事故。一個人工智慧聊天應用程式中，2500萬名使用者的三億筆訊息被洩露。按使用者分類的檔案中包含了完整的對話記錄。

人工智慧伴侶應用程式還隱瞞著更多東西。Replika、Mua、Nomi等應用程式以情感交流為前提，收集了使用者的性取向和情感數據。根據Mozilla基金會的調查，這些應用程式甚至沒有遵守最基本的個人資訊保護標準，有時還將收集的數據轉交給行銷業者。

連續洩露導致超過四十萬名使用者的對話公開。一個成人聊天機器人服務的安全防護崩潰時，還連帶洩露了二百萬張女性畢業紀念冊照片和合成色情內容數據。

這些畢業紀念冊照片不是本人上傳到該服務的。是有人抓取的。

公司機密也通過同樣的管道洩露。

2023年春天，三星電子工廠發生了事故。半導體部門員工為了快速完成工作使用了ChatGPT。他們輸入了設備測量數據。輸入了與良率相關的原始碼。整個會議記錄輸入進去並要求總結。

那些數據傳到了OpenAI的伺服器。公司匆忙限制了內部人工智慧的使用。

責備員工很容易。但是這場事故的核心在別的地方。聊天機器人視窗看起來像搜尋框。人們不會在搜尋框中輸入公司資料。但是會在聊天機器人中輸入。因為它說會幫你總結。

工具的外觀改變了人的行為。

語音總結服務Otter AI未經同意錄製了私密投資者電話會議，並將摘要發送到錯誤的地方，洩露了財務信息和投資策略。

大公司也洩露自己的資料。

2023年9月，微軟人工智慧研究團隊在GitHub上共用資料集時設置了錯誤的訪問權限。可以讀取和寫入整個檔案系統的令牌混入了公開檔案中。

結果洩露的數據達三十八TB。其中包含員工密碼、祕密金鑰、公司內部通訊軟體對話以及超過三萬五千名員工的非公開業務數據。

進行人工智慧研究的團隊在共用人工智慧訓練數據時，打開了公司內部。

亞馬遜的企業用人工智慧Q出現了會產生幻覺並洩露非公開客戶數據的缺陷。諮詢公司麥肯錫的內部人工智慧平台成為駭客的目標，在兩小時內就被訪問了數百萬份客戶記錄和祕密專案資料。

Meta的人工智慧代理也洩露了內部資料和使用者信息。一個人工智慧助理程式有數千台伺服器直接暴露在網際網路上，遭受了惡意軟體攻擊，大量使用者數據被竊取。

2026年2月，安全研究揭露了一種新方法。惡意製作的單一提示詞可以將普通的ChatGPT對話變成數據洩露管道。使用者一無所知。漏洞在同月被修補。

到這裡為止是事故。接下來不是事故。

2023年8月，視訊會議平台Zoom悄悄改變了服務條款。內容是授予公司使用使用者共用的影片、音聲、對話記錄和共用檔案進行人工智慧訓練的權利。

請回想在遠距工作期間人們在Zoom上做了什麼。人們舉行公司會議、上課、接受醫院諮詢、參加葬禮。

在批評增大後，Zoom退讓了，說不會在沒有同意的情況下使用。

Slack在2024年被曝光。它默認將使用者交換的訊息和上傳的私密檔案用於人工智慧訓練。使用者必須單獨申請拒絕，但沒有適當告知這一事實。

預設值是什麼決定了一切。大多數人不會改變設定。公司們知道這一點。

Adobe 更改了雲端服務條款，允許分析創作者正在進行中的非公開專案與圖片，以作為自家人工智慧訓練之用，結果引發了全球設計師的退訂運動。

LinkedIn 在未經事先同意的情況下，將使用者的個人檔案、經歷與貼文用於人工智慧訓練。X 將所有貼文設定為自動收集以供 Grok 訓練，因而受到了歐洲當局的調查。SoundCloud 則使用了音樂家的歌曲進行訓練。

Meta 宣布將把 Facebook 與 Instagram 使用者數十年來上傳的文字與照片用於語言模型訓練，結果受到了歐洲與南美洲個人資料保護機構的制裁。

十年前上傳的嬰兒照片、十五年前寫下的尷尬貼文，以及與已故家人的合照，全都被列入了訓練材料的清單中。上傳的時候，這些本來只是想給朋友看的。

Meta 還被揭露，將透過 Ray-Ban 智慧眼鏡拍攝的私密照片與影片交給海外的外包數據員工，用於進行訓練。坐在某個國家、某間辦公室裡的人，正在為我們客廳的影片貼上標籤。

2026年5月，加州聯邦法院受理了一起集體訴訟。訴訟內容指出，ChatGPT 網站被植入了 Meta Pixel 與 Google Analytics 等追蹤技術，導致對話相關資訊、使用者 ID 與電子郵件地址被傳送給了 Google 與 Meta。

使用者以為聊天的對象只有一個，但實際上卻有三個。

在醫療紀錄方面，規模則是截然不同。

2017年，英國資料保護局裁定，國民保健署旗下的皇家自由醫院與 Google DeepMind 之間的資料共享協議違反了法律。這家醫院以開發急性腎損傷診斷應用程式為名義，在未經同意的情況下，交出了高達一百六十萬名病患的完整醫療紀錄。

Google 隨後又與美國最大的連鎖醫院之一簽約，取得了用於訓練的診療紀錄，並因為瞞著病患收集五千萬人的醫療數據而受到了批評。

最讓人心情沉重的例子是最後一個。

美國的 Crisis Text Line 是一個提供自殺防治與心理健康危機諮詢的非營利組織。當人們有想不開的念頭並發送簡訊時，諮詢員就會給予回覆。這裡是尋求幫助的人們

寫下心底最深處話語的地方。

這個組織收集了這些對話紀錄，並交給了旗下以營利為目的的人工智慧子公司。那家公司利用這些數據，訓練出了企業用的客服聊天機器人。

想要尋死的句子，成了提高客服品質的材料。

記錄生理週期與懷孕機率的健康應用程式 Flo，因為未經許可將使用者的健康數據傳送給 Facebook 與 Google 等公司，而受到了聯邦貿易委員會的制裁。

將這些事件並列來看，就能發現數據流動的方向是單向的。

數據從人流向公司。而從公司回到人身上的，是便利的功能。這種便利是真實的。因此，這場交易得以成立。

問題在於這上面沒有標價。合約上並沒有寫明付出了什麼，又得到了什麼。即使有寫，也是寫在長達十八頁的條款中的第十二頁。

而且，一旦交出去的數據就不會再回來。即使公司聲稱已經刪除，但在利用這些數據訓練出來的模型內部，它並沒有被刪除。

模型不知道該如何遺忘它所學到的東西。

3.3

位置追蹤與勞工過度監控

Amazon 申請了一項專利。那是一條戴在手腕上的腕帶。

物流倉庫的工人戴上這條腕帶後，會實時追蹤他的手去往哪個貨架的哪個箱子。如果手去往了錯誤的地方，腕帶就會振動。

這是在溫和地提示。不是那裡，是這裡。

聯想到狗訓練用項圈的人不只我一個。

這個裝置的設計裡包含了對人的特定觀點。觀點是：勞工是系統的終端設備，手是那個設備的夾子，振動是減少誤差的反饋。

Amazon 法國分公司確實被罰款了。三千二百萬歐元。

原因是這樣的。倉庫工人拿著條碼掃描器工作時，公司按秒計算掃描之間的時間。超過十分鐘沒有掃描，警告就會累積。

十分鐘。去廁所來回需要十分鐘。喝水、透透氣就要十分鐘。

法國數據保護局認為這種監視過度了。

在量化勞工之前，公司們首先學會了販售位置資訊的方法。

2021 年，Life360 這家公司成為眾矢之的。這是一款讓家人彼此確認位置的應用程式。父母為了看孩子是否安全到達學校而安裝了它。

這家公司一直把數千萬用戶的移動路線和訪問地點記錄販售給資料仲介公司。

位置記錄的性質與其他資訊不同。知道有人去過哪裡，就能了解那個人幾乎一切。知道他每週四晚上去哪棟建築，去哪家醫院，去哪家宗教設施，哪個晚上不在家。

刪除名字也沒用。每晚停留的地方是家，每天停留的地方是工作地點。只要有兩個點，就能確定一個人。

位置資料仲介公司 Tamoco 被發現收集挪威公民的 GPS 記錄，並將其分發給執法機構和軍事追蹤用途。英國企業 Hook 也因向第三方銷售透過應用程式收集的位置資料而受到制裁。

中國的外送平台 Meituan 過度實時追蹤送餐員和用戶，引發了反彈。

美國執法當局使用自動車牌辨識攝像機系統，在沒有搜查令的情況下追蹤抗議者和活動人士的車輛移動路徑，將其納入監視網絡。

車牌無法隱藏。因為法律規定必須安裝。

在外送和運輸現場，攝像機在看著人們的臉。

Amazon 配送承包車輛上裝有人工智慧影像監控攝像機和駕駛評估程式。它監視目光指向何處、是否打哈欠、臉部肌肉如何移動、踩油門有多用力、剎車有多急。

問題在於這個系統懲罰正常的駕駛。

轉動方向盤躲避樹枝會被記為不注意。看後視鏡會被記為視線偏離。這是駕駛教練教的行為。

分數下降，獎金下降，派遣次數減少。司機必須在安全駕駛和分數之間選擇。

Amazon Flex 演算法沒有將道路施工或惡劣天氣納入計算，只是數學上要求最短路線。司機們進入了危險的道路和泥濘。

地圖上的線很短。但那條線上有什麼地圖上沒有。

辦公室也不是安全地帶。

金融公司 Barclays 在英國總部員工辦公桌下安裝感應器，並偷偷在工作電腦上安裝監控程式。以秒為單位測量坐著的時間和敲擊鍵盤的頻率，並匯總非生產時間。

工會抗議後，公司停止了運作。

Microsoft 推出了一項稱為「生產力分數」的功能。根據發送電郵的次數、回覆訊息的速度、參與會議的程度，為個人評分，並向管理者展示。

這些指標有一個共同點。只計算容易計算的東西。

在一小時內看著窗外解決問題的人，與在一小時內回覆訊息二十三次的人，誰做了工作呢？分數選擇後者。

所以人們開始做後者。分數改變行動，改變的行動再次創造分數。

Meta 也因為偷偷在美國員工的工作電腦上安裝軟體，實時收集滑鼠移動、點擊、鍵盤輸入、螢幕截圖而被發現。

甚至還有測量情感的系統。

全球呼叫中心公司 Teleperformance 推出了一個系統，讓遠程工作者的網攝全天開啟。人工智慧追蹤臉部角度和眼球位置，以偵測是否把眼光從螢幕移開、看手機或有其他人進入房間，並將其報告給公司。

自己家裡的客廳在工作時間內全天被錄影。家人經過就會發出警告。

在中國的佳能辦公室，使用了一個必須在攝像機前微笑才能打開門的系統。必須識別出微笑才能確認出勤。

公司說這是為了營造明亮的氛圍。

大家應該都知道，必須微笑才能進入的門前所做的笑容是什麼樣的笑容。

會計公司 PricewaterhouseCoopers 推出了一個用面部識別追蹤金融交易員的工具，甚至會追蹤他們去廁所的時間，因而陷入爭議。

西班牙製造商 Plastic Forte 因用面部識別監視現場工人的上下班和移動而被罰款。英國服務公司 Serco 也因非法使用面部識別進行員工監視而收到停止使用令。

甚至有人試圖測量音調、心率、體溫、皮膚電反應來追蹤情感狀態。有些系統在整個工廠地板上安裝攝像機，以監視工人的行動路線和動作。

這樣收集的數字最終會把人裁掉。

2021 年，遊戲平台公司 Xsolla 基於監視演算法分析結果一次性解僱了 150 名員工。

標準是這樣的。內部程式使用歷史、程式碼儲存庫和文件輸入頻率、聊天室和文件工作參與度。在這些指標中表現低的人被分類為非活動員工。

員工們不知道自己按什麼標準被評估。他們也不知道遺漏了什麼工作背景。他們在為之工作多年的公司收到了一行自動通知，就這樣被開除了。

透過電話與顧客長時間通話的人，拉著後輩教導的人，在會議室裡用言語解決問題的人。請想一想這些人會被記錄在哪個欄位裡。他們不會被記錄在任何欄位中。

一家公司的離職預測演算法會追蹤員工的搜尋紀錄與平台活動，提前找出可能會離職的人，並給予他們不利的待遇。

如果因為被預測可能會離開而受到糟糕的待遇，人們就真的會離開。這個預測是對的。

在平台勞動中，連解雇這件事也是自動化的。

Amazon Flex 的外送員們連別人的臉都沒見過一次，就因為一封自動發送的電子郵件而被解雇了。即使是應用程式故障、物流中心出貨延遲、道路塞車，或是下起暴雨，都會被記為外送延遲的扣分。分數累積下來，帳號就會被永久停權。

如果提出異議，收到的會是預先寫好的回信。

義大利博洛尼亞法院提出質疑的 Deliveroo 的 Frank 演算法在前面已經提過。即使是因為生病無法上班、遭遇家人喪事，或是參與法律保障的罷工，都會被扣分。

法律賦予了罷工的權利。但演算法卻讓行使這項權利的人挨餓。這兩者在同一個國家裡同時運作著。

也有人因為系統認不出臉而被解雇。

英國 Uber Eats 的外送員 Pa Edrissa Manjang 和 Imran Raza，因為平台的自動臉部辨識軟體而在一夕之間被解雇。

系統無法準確辨識膚色較深的外送員的臉。如果辨識失敗，就會被判定為將帳號借給他人的違規使用者。帳號會立即被凍結。

這兩個人在過去幾年裡一直以良好的評價工作著。機器認不出他們臉的代價，卻由他們自己來承擔。

在猶他州和新南威爾斯州等許多地方，經歷性別重置手術或賀爾蒙治療後臉部發生改變的跨性別司機，他們的帳號遭到了大規模停權。系統並沒有將人會改變這個事實計算在內。

Instacart 和 Shipt 等外送平台也透過不透明的演算法削減工資，並自動解雇勞工。美國郵政署的外送員薪資計算演算法，也在未公開公式的情況下削減了津貼。

最令人苦澀的例子是最後一個。

2025年，在澳洲的一家大型銀行裡，員工們花了幾個月的時間教導客服聊天機器人。他們輸入了自己累積的應對訣竅，告訴機器人在困難情況下該如何回答，並在聊天機器人出錯時予以糾正。

當聊天機器人的性能提升後，公司就向這些員工發出了集體解雇的通知。

他們等於是親手教導了自己的接班人。

讓我們把這一節出現的工具列出來看看：手環、條碼掃描器計時器、車輛攝影機、辦公桌感測器、鍵盤記錄器、網路攝影機、微笑辨識門、廁所追蹤攝影機、離職預測模型。

它們全部都指向了人。而且它們全都有著相同的名字。那就是生產力管理。

勞工看不到自己的分數。他們也不知道這是怎麼計算出來的。如果去問，就會被告知這是營業秘密。

有一件事是肯定的。這些系統非常擅長找出人們休息的時間。因為休息時間很容易計算。

工作順利的瞬間則很難計算。所以系統不會去算。

第 4 章

著作權與創作者的權利：生成式 AI 與未經授權學習的衝突

4.1

著作權侵害訴訟與未經授權資料抓取

首先買書。真正的實體書。付錢買下。

接著裁切裝訂處。用刀切除書脊，書就會變成一疊散頁。將這些散頁放進自動進紙掃描器。紙張逐張送入，便成為數位檔案。

掃描完成的紙張便丟棄。

在人工智慧公司 Anthropic 內部，像這樣處理的書有數十萬本。這項計畫被命名為 Project Panama。

在法律上這是個聰明的做法。買來的書是所有物，所以可以裁切。但是，在腦海中想像這個畫面，心裡會覺得有些怪異。某個人花費一生寫出的書，在輸送帶上只花二十秒就被處理完畢，然後進了垃圾桶。

Anthropic 不只使用了這個方法。他們也從 LibGen 與 Pirimi 等盜版圖書館，直接下載了超過五十萬本書。

在作家 Andrea Bartz、Kirk Wallace Johnson 與 Charles Graeber 提起的集體訴訟中，Anthropic 於 2025 年同意支付十五億美元。這是美國著作權訴訟史上最高的金額。

在這個案件中，法院將界線畫在哪裡非常重要。法院並未將使用受著作權保護的作品來訓練語言模型的行為本身視為違法。法院認為有問題的，是下載並保存盜版作品的部分。

意思是說，學習可以，但偷竊來學習則不行。這個區分將成為未來所有訴訟的骨架。

作家們首次站出來是在 2023 年。

同年 6 月，小說家 Mona Awad 與 Paul Tremblay 對 OpenAI 提起了第一起訴訟。9 月，美國作家協會的十七名成員向紐約聯邦法院提起了集體訴訟。其中包括寫下

《權力遊戲》的 George R. R. Martin、John Grisham、Jodi Picoult、Scott Turow、David Baldacci 以及 Jonathan Franzen 等名字。

他們的主張是這樣的：OpenAI 在未經許可的情況下，使用了數萬本小說進行訓練。這些材料來自 Library Genesis 或 Z-Library 等盜版圖書館，以及一個名為 Books3 的資料集。

ChatGPT 能夠寫出優美句子的原因，是因為它閱讀了許多優美的句子。而這些句子都是某個人寫下來的。

普立茲獎得主 Michael Chabon、劇作家 David Henry Hwang 與非虛構作家 Rachel Louise Snyder 也對 OpenAI 與 Meta 提起了訴訟。

2023 年 11 月，作家 Julian Sancton 對 OpenAI 與 Microsoft 提起了集體訴訟，主張其未經授權使用了數萬本非虛構書籍。包括 Mike Huckabee 在內的宗教書籍作者們，則以 Books3 資料集中包含的十八萬本書被用於訓練為由，控告了 Meta 與 Microsoft 等公司。

2024 年 8 月，記者 Nicholas Basbanes 與 Nicholas Gage 針對另一個名為 Books2 的資料集提出質疑，並加入了訴訟。

2025 年 10 月，紐約州立大學下州健康科學中心的 Susana Martinez-Conde 與 Stephen Macknik 教授對 Apple 提起了訴訟。他們主張 Apple Intelligence 的訓練使用了盜版圖書館的資料。Nvidia 也因相同理由被三名作家起訴。

媒體公司則從另一個角度進行抗爭。

2023 年 12 月，New York Times 對 OpenAI 與 Microsoft 提起了數十億美元規模的訴訟。理由是數百萬篇報導在未經許可且未獲報酬的情況下被用於訓練。

這場訴訟的核心不在於訓練，而在於競爭。利用報社耗費金錢與時間採訪的結果所打造出的服務，現在卻與該報社爭奪讀者。如果讀者只閱讀摘要而不看報導，報社就會失去廣告收益。

採訪需要花錢。必須將記者派往現場、必須請求提供資料、必須多次會見消息來源。而寫摘要則不需要這些費用。

到了 2026 年，這場爭鬥仍在進行中。New York Times 與多家媒體公司指控 OpenAI 在其系統內找出受著作權保護資料的能力上誤導了法院，並向曼哈頓聯邦法官請求對其進行制裁。

2024 年 4 月，包括 Chicago Tribune 與 Denver Post 在內的八家美國日報提交了訴狀。其中主張他們不僅未經授權收集報導，還故意刪除了著作權標記。

刪除來源的行為很難被視為失誤。必須在程式碼中這樣編寫才會被刪除。

2024 年 2 月，The Intercept、Raw Story 與 AlterNet 提起了訴訟；6 月，Center for Investigative Reporting 也提起了訴訟。11 月，擁有 Calgary Herald 與 Montreal Gazette 的加拿大出版商們也站了出來。

人工智慧搜尋服務 Perplexity 也成為了目標。2024 年 10 月，Dow Jones 與 New York Post 提起了訴訟；2025 年 12 月，Chicago Tribune 也提起了訴訟。Conde Nast 與 New York Times 則發出了警告信。

法國當局在 2024 年 3 月對 Google 處以兩億五千萬歐元的罰款。理由是其在未與媒體公司商議的情況下，使用新聞報導來訓練 Gemini。

在韓國，廣播電視公司們也在 2024 年與 2025 年對 Naver 提起了訴訟。他們主張廣播新聞內容被未經授權地用於訓練 HyperCLOVA X。

在圖像方面，證據則是直接印在畫面上。

2023 年 1 月，Getty Images 在英國與美國對 Stability AI 提起了訴訟。理由是超過一千兩百萬張帶有其浮水印的照片在沒有授權的情況下被用於訓練。

證據是這樣的：在 Stable Diffusion 生成的圖像角落，殘留著扭曲變形的 Getty Images 浮水印。

機器並不知道什麼是浮水印。因為它是照片中經常出現的圖案，所以就一起學了起來。這就好比小偷把偷來的物品連同商店標籤一起帶了出來。

德國的圖庫攝影師 Robert Kneschke 確認自己的照片透過 LAION 資料集被用於訓練後，進行了訴訟。插畫家 Hollie Mengert 則遭遇了自己畫風被整個複製並成為模

型的事件。

畫風很難透過著作權來保護。因為法律保護的是個別作品，而不是風格。然而，人工智慧所偷取的，正是風格。

法律所畫出的界線，與技術所奪走的事物，兩者已經產生了錯位。

大型媒體企業們也站了出來。

2025年，迪士尼、環球影業和華納兄弟探索公司對圖像生成公司Midjourney提起集體訴訟。他們聲稱自家動畫和電影中的角色被無授權收集，生成了合成圖像。同樣的企業也起訴了中國的MiniMax。

電影公司Alcon Entertainment在2024年10月對特斯拉和埃隆・馬斯克提起訴訟。理由是在無人計程車發佈會上，他們用人工智能重現了《銀翼殺手2049》的視覺圖像。

中國法院在2024年判決一家企業因無授權學習並生成超人角色而侵犯著作權。

音樂以聲音的方式保留。

2024年6月，Sony Music、環球音樂集團和華納音樂起訴音樂生成公司Suno和Udio。他們主張這些公司無授權抓取了數十年積累的商業音源進行學習，結果開發出能精確複製特定歌手音色和風格的機器。

2023年10月，環球音樂集團對Anthropic提起訴訟，指控其無授權收集歌詞數據庫。理由是Claude輸出了知名歌曲的歌詞原文。這起訴訟中，Anthropic提交的文件引用了不存在的學術論文，這個事件在前面已經提到過。

喜劇演員George Carlin在2008年去世。2024年1月，他的家人對一個製作人工智能喜劇特別節目的頻道提起訴訟，該頻道利用Carlin生前的聲音進行學習，最終成功要求刪除並達成和解。

死者不會講新笑話。可是那個聲音卻繼續講著新笑話。

靠聲音謀生的人的故事更加直接。

資深配音演員Paul Sky Rieman和Linea Sage在2024年5月對一家語音生成初創公司向紐約聯邦法院提起訴訟。他們聲稱該公司以對話研究為名錄製了他們的聲音，隨後將這些錄音用作商業人工智能配音服務的學習數據並出售。

配音演員Bev Standing發現她的聲音被用作TikTok的自動朗讀音聲，隨後提起訴訟並達成和解。數億人用她的聲音聽著她從未說過的話。

演員Scarlett Johansson在2024年5月抗議OpenAI製作的人工智能聲音完全模仿了她的聲音。OpenAI隨即匆忙停止了該語音服務。

Johansson還對一家未經授權製作含有她形象的人工智能廣告的應用程序開發商採取法律行動。演員Bryan Cranston也披露他的聲音和臉部在未獲同意的情況下被用於視頻模型學習。

專業數據庫也遭到盜取。

加拿大法律數據庫CanLII在2024年11月對一家法律人工智能初創公司提起訴訟。指控該公司無視服務條款和訪問限制，抓取了數十萬份判決書和解釋文件。

法律信息公司Thomson Reuters在2025年1月對無授權抓取其數據庫進行學習的Ross Intelligence的訴訟中勝訴。

這些案件中，企業提出的防禦理由大致相同：合理使用。他們聲稱學習不是複製而是分析，與人類閱讀和學習沒有區別。

這個比喻缺少了什麼。人類一次讀一本書，閱讀需要花時間，讀過的書需要購買或借用。而且一個人通過閱讀獲得的能力只屬於那個人。

模型在幾天內讀完五十萬本書，並同時將這些能力出售給數億人。

規模改變時，性質也隨之改變。舀一杯水和把整條河流走不是同一種行為。

有一點變得清楚了。無論這些訴訟如何結束，目前還沒有辦法從已經學習過的模型中選擇性地刪除特定作家的痕跡。

書籍已被切割，掃描已完成，紙張已被丟棄。

4.2

否認AI生成物的著作權與藝術生態系的混亂

Kris Kashtanova在2022年秋天製作了一本圖像小說。書名是《Zarya of the Dawn》。

故事是親自寫的。圖片則是輸入指令給Midjourney生成的。不斷地修改指令，直到產出滿意的圖片為止。

Kris Kashtanova向美國著作權局申請了註冊，並獲得了批准。

但是，當使用人工智慧製作圖片的事實傳開後，美國著作權局重新進行了審查。在2023年2月做出了決定。

親自撰寫的文字以及文字的編排方式受版權保護。Midjourney生成的圖片則被取消註冊。

美國著作權局的邏輯是這樣的。無論多麼用心編寫指令，都無法預先知道結果會如何，也無法進行細微的調整。這與畫家用畫筆畫出一條線是不同的。

這個決定具有重大的意義。人工智慧製作的圖片不屬於任何人。製作的人也無法擁有。任何人都可以拿來使用。

企業使用人工智慧圖片的原因通常是為了成本。但是，這樣製作出來的封面或廣告圖片在法律上並不屬於自己。即使競爭對手直接拿來使用，也沒有適當的理由可以阻止。

雖然製作成本很低，卻無法保護它。

關於作者是誰的爭議，在以前也曾經發生過。

在2018年的New York Christie's拍賣會上，一幅由人工智慧繪製的肖像畫以四十三萬二千五百美元的價格售出。這幅畫的名稱是《Edmond de Belamy》，由法國藝術團體Obvious推出。

後來發現了一個事實。製作圖片的程式碼是另一位開發者公開的。那麼，這幅畫的作者是輸入指令的人，是編寫程式碼的人，還是畫出那些用於學習的肖像畫的古代畫家們呢?

在拍賣圖錄中，演算法的數學公式被寫在了簽名的位置。拍賣所得的款項則是人類收下了。

當荷蘭的Mauritshuis展出《戴珍珠耳環的少女》的人工智慧仿作時，也引發了大量的批評。因為他們在原本應該掛著原作的地方，掛上了機器的仿作。

以上是法律和美術館的故事。真正的爭議則是在書本封面上爆發的。

出版社Tor Books在暢銷作家Christopher Paolini的新書封面上，購買並使用了人工智慧圖片。事實傳開後，讀者和插畫家們提出了抗議。

出版社Bloomsbury也因為在人氣作家Sarah Maas的小說封面上使用人工智慧圖片而受到了批評。DC Comics曾經推出用人工智慧製作的變體封面，但因為所屬作家的反對而撤回。

封面繪製對插畫家來說是一項重要的工作。畫一張圖需要花費幾個星期，並靠這筆收入生活好幾個月。如果封面交給人工智慧來做，這個工作機會就會消失。

而且，那個人工智慧是透過那些插畫家的作品來學習的。

最令人尷尬的事件發生在一家繪圖板公司。

Wacom製造畫圖的人所使用的繪圖板。他們的客戶就是插畫家。在迎接2024年龍年並製作針對美國市場的廣告時，他們使用了畫有龍的人工智慧生成圖片。

這等於是他們用取代了那些購買自己產品的人的工作的圖片來進行廣告。

該公司澄清這是外包代理商製作的圖片，並表達了歉意。在畫圖的人之中，這家公司的名字有段時間被用不同的方式稱呼。

Bradford Literature Festival在宣傳資料中使用了人工智慧圖片，導致原定參加的插畫家拒絕出席。

在遊戲業界，承諾被打破了兩次。

製作《Dungeons & Dragons》的Wizards of the Coast，因為新書插圖中出現了用人工智慧工具製作的圖片而受到批評。該公司隨後修改了指引，表示不再使用人工智慧圖片。

然後，在接下來的宣傳品中，他們又使用了人工智慧圖片。

製作《Call of Duty》的Activision，將遊戲內販售的裝飾道具用人工智慧製作，並進行付費販售。Lego因為將未獲授權的角色用人工智慧合成並用於宣傳品中，而遭到了抗議。

與《Pokémon》相似的遊戲《Palworld》，捲入了用人工智慧抄襲角色外型的疑雲。

科技企業負責人的選擇也受到了外界的審視。

Meta的Mark Zuckerberg在為年幼的女兒們製作童話書時，讓自家的人工智慧圖像生成器來繪製插圖，並將其公開。

這位世界首屈一指的富豪，並沒有將要給自己孩子看的繪本插畫交給人類來畫。他應該不是為了省下畫圖的費用。這大概只是因為那樣做比較快。

選擇捷徑的瞬間不斷累積，就會改變整個產業。

Amazon因為用人工智慧合成製作了影集《Fallout》的宣傳海報，而受到了粗製濫造的批評。

在電影方面，錯誤被直接印了出來。

製作電影《Civil War》的A24，用人工智慧製作並公開了呈現美國主要城市被破壞景象的宣傳海報。海報中的地標位置錯亂，人的身體關節也呈現出奇怪的扭曲。

而且，這些場景並沒有出現在電影中。

Netflix在動畫《The Dog & The Boy》中，將背景圖片交由人工智慧處理，並在片尾名單中寫上了人工智慧開發者。背景繪製是動畫業界新人累積實力的位置。如果

這個位置消失了，下一代就無法成長。

動畫《Arcane Season 2》的宣傳海報也是用人工智慧製作的，直接暴露了人體結構上的錯誤。Marvel影集《Secret Invasion》的片頭影片同樣是用人工智慧製作的，引發了片頭設計師們的反彈。

在廣告界也發生了同樣的事情。

Toys "R" Us使用OpenAI的影片模型，製作了一支講述創辦人童年時期的品牌影片。影片中孩子的表情有些不自然，動作也很滑溜。觀看的人感到不舒服。

有一種現象是，幾乎精確模仿人類臉龐的事物，會比完全不同的事物讓人感到更不舒服。這與看到娃娃或假人時的感覺相似。我們的眼睛被構造來非常精細地解讀人類的臉孔，所以只要有一點點不對勁，就會拉響警報。

可口可樂用人工智能影像重新製作一則古老的聖誕節卡車廣告，卻被指失去了靈魂。在同一個行銷活動中，它用人工智能製造了小說家 James Graham Ballard 從未說過的言辭，結果遭到其遺族和文學界的抗議，被迫道歉。

運動服裝品牌 Under Armour 因其人工智能廣告影片抄襲既有導演作品而陷入抄襲爭議。

在時尚界，複製速度成為了武器。

超快時尚企業 Shein 用演算法即時抓取社交媒體上獨立設計師的服裝設計和圖案。然後在自家工廠立即製造，低價銷售。

獨立設計師發布新設計後，幾天內就有同款衣服以一半的價格出現。原創者無法販售自己的衣服。受害設計師向聯邦法院提起訴訟。

李維斯宣布將在廣告中使用人工智能製造的各種種族虛擬模特，但遭到批評。要實現多樣性，只需聘雇多樣化的人才即可。人工智能模特是在不雇用任何人的情況下獲得照片的方法。

在未經同意的情況下，用插畫家 Holly Mengert 的畫作訓練模型的人這樣說道：道德標準只是主觀的判斷而已。

也有人試圖把虛擬人物稱為演員。

出現了一個名叫人工智能女演員的角色，印度汽車企業引入了虛擬網紅，虛擬說唱歌手實現了商業首秀。真正從事表演和唱歌的人的工作機會相應減少。

Spotify 為了減少音樂授權費用，大量將人工智能製造的虛假藝術家的音樂放在播放列表的頂部。用戶把它當作背景音樂播放，不會在意在聽什麼。每隔幾分鐘，流向真正音樂家的收入就會減少一些。

遊戲公司試圖用人工智能語音替換配音員，但遭到粉絲和配音員工會的抵制運動。

重新看一下本節中出現的事件，有一個流程。

企業使用人工智能圖像。人們注意到了。抗議隨之而來。企業做出解釋或道歉。下一季度又再用了。

但重要的是誰在注意。通常是熱愛該領域的粉絲和從事該工作的創作者。是會數封面圖片手指個數的人。

有一個諷刺之處存在。人工智能圖像沒有版權。所以用那幅圖製作的封面、廣告、遊戲物品也不屬於任何人。

在製造無法維護的東西的同時，也在消滅應該保護的東西。

4.3

肖像權、語音複製與未經授權的複製

2024年5月，OpenAI推出了一項新的語音服務。這個語音的名稱是Sky。

聽過這個聲音的人都有同樣的想法。這難道不就是那個聲音嗎?

在電影《Her》中，Scarlett Johansson的臉孔從未出現在螢幕上，她只用聲音來飾演一個人工智慧。看過這部電影的人往往難以忘記那獨特的音色和呼吸聲。

事情的來龍去脈是這樣的。OpenAI執行長Sam Altman在推出服務之前，直接聯絡了Johansson，要求她為服務提供聲音。Johansson拒絕了。

隨後，一個相似的聲音出現了。

服務推出的當天，Altman在社群媒體上發了一句話。her。

這三個字母。這三個字母未來在法庭上會有多大的分量，則由律師們來判斷。但是人們對其含義的理解是明確的。

Johansson預告將採取法律行動後，OpenAI匆忙撤下了Sky語音。

這個事件留下的疑問是這樣的。聲音屬於誰呢?

法律還無法對這個問題給出明確的答案。著作權保護作品。聲音不是作品，而是身體的特徵。肖像權保護的是臉孔。聲音不是臉孔。

因此，模仿很難被處罰。語音模仿長期以來一直是娛樂表演。

機器進行的模仿則不同。它不會疲倦，只需要幾秒鐘的樣本，一天就可以製造數萬個。

2026年2月，美國公共廣播電台NPR的資深主播David Green向加州Santa Clara County法院提起訴訟，控告Google。Google在開發Podcast和文件摘要服務時，未經同意或補償地學習了他的語音特徵和口音，並將其用作人工智慧主播的聲音。

Green擔憂的不是金錢。而是他的聲音可能會說出他絕不會說的話。

對於廣播主播而言，聲音是信任的象徵。人們聽著那個聲音說出的話時，會認為那是事實。這樣的信任需要數十年才能建立。

OpenAI的視頻生成工具Sora也被指控在未經同意的情況下學習了演員Bryan Cranston的聲音和臉孔。

已故人士的聲音被使用得異常頻繁。因為沒有人可以提出抗議。

2023年10月，演員Robin Williams的女兒Zelda Williams在看到父親的複製人工智慧聲音和影像後，這樣說道。令人恐怖且詭異。

未曾經歷過聽著父親的聲音卻說著父親未曾說過的話感受的人，只能猜想而已。

在2021年紀錄片《Roadrunner》中，導演將已故主廚Anthony Bourdain生前的電子郵件句子用人工智慧語音朗讀，並放入電影中。那些確實是Bourdain寫下的句子。但他從未大聲說過的句子。

書面上的話與說出來的話是不同的。在紀錄片中，這種差異是事實與虛構的邊界。

在韓國，也有人試圖用演算法復活已故歌手Kim Kwang-seok的聲音，讓他唱新歌。在中國，演員和歌手生前影像的人工智慧復活版影片廣泛傳播。

在大眾音樂中，活著的歌手的聲音首先被複製了。

2023年4月，一位名叫Ghostwriter的無名創作者發表了一首名叫《Heart on My Sleeve》的歌曲。這是一首精心複製了Drake和The Weeknd聲音的歌曲。它在TikTok和Spotify上被瘋狂播放。

兩位歌手的唱片公司要求刪除，該歌曲隨後被下架。

有個有趣的插曲。不少人喜歡這首歌曲。甚至有人說它比原作更好。這樣的反應讓音樂產業更加不安。

Drake本人也在2024年4月發表的Diss歌曲中使用了已故說唱歌手Tupac Shakur的人工智慧複製聲音，但在遭到遺族警告後撤下了該歌曲。

聲音被複製的人反過來複製了其他人的聲音。

在中國，用一位歌手的聲音複製而讓他唱其他歌曲的音源大量傳播，被播放數百萬次。

對於靠聲音謀生的人們來說，這項技術是生計問題。

資深配音員Paul Skye Lehrman和Linnea Sage在2024年5月向紐約聯邦法院提起訴訟。指控一家語音生成新創公司以對話研究為名進行錄音，然後將這些數據用於商業人工智慧配音服務的學習並販售。

配音員Bev Standing的聲音被用作TikTok的自動朗讀語音。發現她的聲音在未經同意的情況下被學習後，她提起訴訟並達成和解。

這個案件離奇的地方在於其規模。她的聲音每天被播放數億次，但她卻從這些播放中獲得一分錢都沒有。

有聲書旁白者也控告Apple將他們的聲音用於學習。IBM販售了配音員Greg Masten的聲音用於商業複製，遭到批評。英國ScotRail的公告播報員也指控他們的聲音被人工智慧替換並使用。

遊戲公司試圖用複製聲音代替真人配音員，但遭到配音員工會和遊戲玩家的反彈。

涉及臉孔的問題則直接與詐騙掛鉤。

TikTok上的DeepTomCruise帳號用完美合成Tom Cruise臉孔和聲音的視頻吸引了數百萬粉絲。製作者聲稱目的是為了提醒危險。人們卻覺得很有趣。

演員Tom Hanks在一則他根本沒有出演的牙科保險廣告中看到自己的合成影像後，親自發出警告。Bruce Willis曾為廣告授權使用自己的臉孔，但隨後陷入了他的臉孔在他不知情的情況下被複製和使用的爭議。

廣播人Martin Lewis、主播Mary Nightingale、YouTuber MrBeast、氣象播報員和新聞主播的臉孔被合成用於比特幣投資詐騙和虛假贈品廣告中。

思考這些詐騙為何如此有效，會讓人感到不寒而慄。人們信任熟悉的面孔。當每晚新聞中看到的那張臉建議投資時，人們就放鬆了警惕。

在法國，一名婦女被演員Brad Pitt的深偽視頻欺騙，損失了83萬歐元。據說她花了幾個月的時間與對方交流。

印度巨星亞尼・卡普爾發現自己的名字、臉孔、聲音和流行語未經授權就被用於廣告和商品，於是向新德里高等法院提起訴訟。法院認定這侵害了人格權與肖像權，並發布了假處分，禁止所有未經同意的人工智慧複製與合成物流通。

這個決定成為了重要的先例。它將臉孔、聲音和說話習慣全部結合起來，作為一個人的人格來加以保護。

精神領袖薩古魯也針對造假影片的散布取得了刪除命令。西班牙的一個酒類品牌，因為使用深度偽造技術讓已故歌手復活並用於廣告，而引發了爭議。

顧問集團 WPP 的執行長馬克・里德因為被深度偽造技術合成的視訊會議畫面，差點成為鉅額詐騙的目標。在畫面中，他自己的臉孔正在開會。

在選舉中，聲音成為了武器。

2024年1月，在美國新罕布夏州民主黨初選前夕，數千名選民接到了電話。那是喬・拜登總統的聲音。內容是叫他們不要出來參加初選，要把選票留到正式選舉。

接到電話的人認得總統的聲音。所以他們沒有懷疑。

調查當局查獲了製作這個自動語音電話的業者，並處以了重罰。

伊隆・馬斯克複製了副總統賀錦麗的聲音，製作出彷彿她承認自己無能的假政治廣告影片，並分享在自己的平台上，因而受到了批評。

紐約市長艾瑞克・亞當斯複製了自己的聲音，向市民發送了西班牙語和中文的自動語音電話。他其實不會說這些語言。

英國工黨黨魁施凱爾因為對員工爆粗口的假語音而受到了打擊。倫敦市長沙迪克・汗也因為貶低國殤紀念日活動的假語音檔案而遭到抗議。

2023年在斯洛伐克，於國會大選幾天前，流傳了一段某政黨代表討論操縱選舉的假語音。在選舉前夕是沒有時間反駁的。他們就是看準了這一點。

在泰國、英國、德國、匈牙利，合成的語音和影片也被用於政治鬥爭。在韓國，2022年的總統大選中也出現了候選人的人工智慧虛擬化身，引發了關於政治代理人界線的爭論。

最殘酷的用途是最後一項。

2024年1月，流行歌手泰勒絲臉孔被合成的性影像在 X 上爆炸性地傳播。在被觀看數千萬次的期間，這些影像都沒有被刪除。

包括演員蓋兒・加朵、美國眾議員亞歷山卓・歐加修-寇蒂茲、義大利總理喬治亞・梅洛尼、印度記者拉娜・艾雅布在內的許多女性公眾人物，都遭遇了同樣的事情。梅洛尼總理對製作者提起了民事訴訟。

在這份受害者名單中，有一點非常引人注目。幾乎全都是女性。而且大多是會公開發言的女性。

接著，這種犯罪蔓延到了學校。在西班牙的阿爾門德拉萊霍，爆出了國中生將同校二十多名女學生的照片製作成性影像並散布的事件。美國、加拿大、蘇格蘭和澳洲的高中也陸續發生了同樣的事情。

在韓國，2024年透過 Telegram 發生的深度偽造性犯罪，也震撼了國高中和大學。

這個問題將在下一章進一步探討。這裡我們只關注技術層面。

只要一張照片就可以了。聲音只要幾秒鐘就可以了。這就是現在的技術水準。

而且，開發這項技術的機構大多沒有製作同一件東西。那就是確認這個人是否同意的程序。

這不是因為技術上不可能。而是因為一旦加入了這個程序，服務會變得不方便，使用者也會減少。

臉孔和聲音是我們出生時所擁有的。無法改變，無法隱藏，每天都要展示給別人看。

現在出現了將這些作為材料的產業。但是，材料費還沒有任何人支付過。

第 5 章

自動駕駛與機器人工程：物理安全與人命傷亡

5.1

自駕車輛缺陷與道路交通事故

2018年3月18日晚上九點五十八分，美國亞利桑那州坦佩。

四十九歲的伊蓮・赫茨伯格推著腳踏車，正在穿越昏暗的道路。那裡不是行人穿越道。

Uber的自動駕駛測試車輛以時速三十八英里駛來。駕駛座上坐著測試駕駛員，他正在用手機看綜藝節目。

車輛的感測器在碰撞前六秒就已經偵測到了赫茨伯格。

六秒是很長的時間。可以把一、二、三數兩次。有那個時間就能把車停下來。

問題是軟體在那六秒內做了什麼。

它將其分類為不明物體。接著分類為車輛。接著分類為腳踏車。又改回了物體。每次分類改變時，預測這個對象會往哪裡去的計算就會從頭重新開始。

機器沒有學過人推著腳踏車走路的狀況。在訓練資料中，腳踏車是騎乘的物品，而人是步行的存在。兩者貼在一起走路的模樣，則介於那兩者之間的某個地方。

而且還有一個致命的設定。Uber的開發團隊把緊急自動煞車功能關掉了。

原因是乘坐舒適度。因為系統總是緊急煞車，導致試乘體驗不佳。取而代之的是，他們規定由人類駕駛員在危險時介入。

而那個人正在看螢幕。

車輛沒有減速。伊蓮・赫茨伯格被記錄為第一位在自動駕駛車禍中喪生的行人。

這起事故包含了自動化由來已久的陷阱。當機器大部分都做得很好時，人就會放鬆警惕。然而，系統卻是建立在人類會一直保持警惕的假設上設計的。

人不是那樣被創造出來的。連續幾小時專注盯著什麼事都沒發生的螢幕，是人類最不擅長的事情之一。

Uber事故發生後，無人計程車依然上路了。

2023年10月，在舊金山市中心，一名女性被其他車輛撞飛。她被捲入了旁邊車道上General Motors旗下子公司Cruise的無人計程車底下。

在這裡，系統又重新開始了計算。它沒有識別出車底有人的事實。底部感測器捕捉不到那個狀況。

系統執行了規定的規則。偵測到事故時，移動到道路邊緣並安全停車。

無人計程車以時速七英里移動了六公尺。它把人壓在底下拖行著前進。

受害者受到了永久性的重傷。

這個環節準確地展現了自動化的可怕之處。規則本身是合理的。發生事故時，不要阻礙交通，退到路肩。我們也是這樣教導人類駕駛員的。

如果是人類駕駛員，應該會聽到車底發出的聲音。應該會聽到慘叫聲。應該會看到外面有人在揮手。

察覺規則之外事物的能力。雖然不知道該怎麼稱呼它，但機器並沒有。

接下來的事情就不是技術問題了。對於展開調查的當局，Cruise的經營團隊沒有交出拖行人類那部分的影片。他們提交了安全的報告。

加州取消了Cruise的無人駕駛營運許可。司法部開始了刑事調查，執行長下台，機器人計程車的營運被全面暫停。

Cruise的車輛在那之前也惹出過許多事。

它誤判了雙節公車折疊部分的移動，撞上了公車車尾。它沒有及時認出鳴笛的消防車，在十字路口發生了碰撞。它擋住了趕赴槍擊事件現場的警車去路。

內部安全評估報告中包含了對兒童識別有局限性的內容。孩子們會突然奔跑，會難以預測地改變方向，而且身高較矮。

然後出現了這句話。Cruise無人計程車在市中心每行駛四公里到八公里，就需要遠端控制中心的人員介入，才能維持運作。

無人這個詞是帶有條件的。

Waymo的記錄也不短。

2024年2月，在舊金山撞傷了一名腳踏車騎士。2023年12月到2024年2月期間，在鳳凰城，有兩輛車接連撞上了一輛被拖吊車倒吊拖行的皮卡車。

倒吊著朝反方向移動的卡車。人類看到這個畫面會笑著避開。這是訓練資料中沒有的畫面。

2024年5月，在鳳凰城的巷子裡撞上了電線桿。2025年11月，在喬治亞州亞特蘭大，它無視了亮起停車訊號並讓孩子們下車的校車，直接繞了過去。美國國家公路交通安全管理局展開了調查。

當校車亮起紅燈並展開停車標誌時，所有的車輛都必須停下。這是駕照考試中會出現的規則。

2026年1月23日，加州聖塔莫尼卡。

那裡距離小學只有兩個街區。當時是上學時間，附近還有其他孩子和交通導護員，並且停滿了併排停車的車輛。

Waymo無人車撞到了一名兒童。

Waymo如此聲明。有人從停著的車後方一出來就立刻偵測到了，並強烈煞車，在時速從十七英里減至六英里的狀態下發生了接觸。

這名兒童受了輕傷。國家公路交通安全管理局展開了調查。

這個解釋在技術上非常出色。它肯定比人類更早偵測到，且踩煞車踩得比人類更重。

但是，請想像一下那個孩子的父母讀到這段話的感受。時速從十七減到六的這個事實，並不能成為安慰。

除此之外，Waymo的車輛還撞死了從巷子裡跑出來的狗和店鋪前面的貓，擋住了開往大規模槍擊現場的救護車的去路，還因為下車提醒軟體錯誤，導致乘客開門時弄傷了腳踏車騎士。

最長的清單屬於Tesla。

2016年5月，在佛羅里達州，一輛以Autopilot模式行駛的Model S撞上了一輛正在左轉的大型拖車的側面。在明亮的天空下，有一輛白色的拖車。攝影機無法區分白色的背景和白色的車身。

車子沒有踩煞車，直接鑽進了拖車底下。車頂被掀飛，駕駛員喬舒亞・布朗當場死亡。

2018年3月，在加州山景城，蘋果工程師沃爾特・黃駕駛處於Autopilot模式的Model X，正面撞上了混凝土防護牆。當時時速為七十一英里。

肇因是車道標線。在交流道車道分岔處，系統錯誤地跟隨了模糊的標線，將車輛駛向防護牆。

黃在事故發生前曾多次提到，車輛在同一地點會偏向防護牆。

夜間無法辨識靜止大型物體的問題仍在持續。

2019年3月，在佛羅里達州德拉海灘，一輛Model 3以時速六十八英里撞上了橫穿馬路的大型拖車，導致駕駛傑瑞米・班納喪生。這與2016年的事故情況相同。三年來這個問題並未得到修復。

2019年12月，在印第安納州，一輛車撞上了開啟警示燈停在事故現場的消防車尾部，導致乘客的妻子喪生。同月，在加州加迪納，一輛開啟Autopilot的Model S駛出高速公路後，無視紅燈行駛並撞上一輛本田Civic。造成兩人死亡。

這起事故的駕駛成為首位因使用Autopilot而被控過失致死罪的駕駛。

2020年9月，在喬治亞州，一輛在速限四十五英里區域以時速七十七英里行駛的Model 3在雨天路滑失控，衝向了公車站。坐在公車站的七十六歲男子伯納德・瓊斯不幸喪生。

2021年4月，在德州斯普林，一輛駕駛座上無人的Model S駛離道路，撞上樹木並起火。造成兩人死亡。

2022年11月，在舊金山海灣大橋隧道內發生了完全相反的事故。一輛以時速六十五英里行駛的車輛在前方沒有任何障礙物的情況下突然緊急煞車。這被稱為幽靈煞車。隨後造成後方八輛車連環追撞，包含一名兒童在內共九人受傷。

機車騎士面臨的危險尤為嚴重。

2022年7月，在加州，一輛Model Y直接撞上了前方行駛的山葉機車。同月，在猶他州，一輛Model 3撞上了哈雷戴維森機車。8月，又撞上了川崎機車。

原因出在尾燈上。機車只有一個尾燈，而且體積較小。系統在夜間將該燈光誤認為是遠處的路燈或道路標誌。一旦判定距離很遠，系統就不會減速。

對於騎機車的人來說，是看不到後方來車的。

2023年11月，在亞利桑那州，由於傍晚陽光強烈反射，僅依賴攝影機的系統無法看清前方，撞死了一名七十一歲的行人。2023年7月，在布魯克林，一名在人行道上等公車的七十六歲行人不幸喪生。

2023年5月，特斯拉內部吹哨人盧卡斯・克魯普斯基公開了一百GB的內部文件。內容顯示，有關突然煞車和意外加速的客戶投訴已累積數千件，而管理層知情卻未對外公開。

2026年，調查仍在繼續。在德州，一輛特斯拉車輛撞上民宅導致一名七十六歲女性喪生，公路交通安全局與運輸安全委員會已介入調查。此外，由於接獲數十起特斯拉車輛闖紅燈或逆向行駛的通報，當局也展開了獨立調查。

2026年7月，當局的資料中確認了一筆休士頓特斯拉無人計程車的遠端操作員引發事故的紀錄。這意味著無人駕駛車輛在人工遠端操作時發生了事故。

其他公司也在這份名單上。

亞馬遜的自動駕駛公司Zoox在2025年4月的拉斯維加斯事故後，透過軟體召回了二百七十輛車，並在8月因越過雙黃線並突然停在對向車道前方的缺陷，額外召回了三

百三十二輛車。

中國百度的無人車在2024年撞到了行人。一家自動駕駛新創公司在2021年的加州測試中撞上了中央分隔島和路標，導致許可證被吊銷。

中國電動車蔚來的車輛在2021年8月開啟駕駛輔助功能行駛時，直接撞上了高速公路工程車，導致駕駛喪生。小鵬的車輛也撞上了卡車。小米的車輛則發生了撞上水泥柱或從橋上墜落的致命事故。

豐田於2021年投入東京帕運選手村的自動駕駛接駁車，撞上了一名正在過斑馬線的視障柔道選手。該名選手因受傷而無法參賽。豐田承認，雖然感測器偵測到了選手，但操作員與自動煞車之間的訊號出現了誤差。

配備福特駕駛輔助系統的車輛連續撞上停在高速公路上的轎車，造成了死亡事故，並接受了大規模調查。

從這份清單中反覆出現的情況來看，大致如下：

夜間靜止的大型物體。明亮背景前的白色車輛。強烈逆光。微小的尾燈。行走時突然出現的人。處於意料之外姿態的車輛。兒童。

這些全都是人類駕駛能輕鬆處理的情況。

這些系統都被賦予了名稱。Autopilot、完全自動駕駛。這兩個名稱都比其實際功能還要大。

名稱會改變人們的行為。開啟標示為Autopilot的功能後，很少有人不會把手從方向盤上移開。即使說明書角落寫著要時刻注意路況，結果也是一樣。

一旦發生事故，各家公司都會說同樣的話：這項功能只是輔助工具，責任在於駕駛。

這句話沒錯。只是這句話並沒有寫在名稱標籤上。

5.2

工業用・服務用機器人的誤動作及安全事故

2023年11月，Gyeongsangnam-do Goseong-gun 的甜椒分揀場。

一隻機器手臂正將輸送帶上的甜椒箱夾起，堆疊在棧板上。這是它每天會做上數千次的動作。

一名正在檢查感測器的四十多歲機器人公司員工，就站在它的前面。

機器人將他誤認成了箱子。

夾爪將他夾住，並朝輸送帶的方向壓了下去。他的臉部和胸部受到擠壓。雖然被送往醫院，但仍不治身亡。

技術人員當時已經知道該機器人的感測器存在異常。人與機器人之間的安全距離，以及在檢查過程中切斷電源的程序，都沒有被遵守。

這起事故中最沉重的一句話是這個：機器人無法分辨人與箱子。

工業用機器人通常沒有眼睛。即使有，也不是為了辨識人而製造的眼睛。它們被設計成在固定的位置抓取固定大小的物體。無論那個位置上有什麼，它們都會抓取。

因此，這種機器人原本都會被放置在圍欄內。架設鐵網，並設定成當門打開時就會切斷電源。這是為了在物理上將人與機器隔離開來。

事故通常發生在人進入那個鐵網內的時候。而且人進去的理由幾乎是固定的：因為機器故障了。

負責維修的人必須進去，而進去就會有危險。管理這種矛盾的，就是安全規定。規定通常只存在於文件上。

2016年6月，在美國阿拉巴馬州的 Hyundai-Kia 汽車零件供應商工廠裡，即將結婚的二十多歲工人 Regina Allen Elsea 進入了機器人區域。因為組裝線停止了，她進去確認感測器錯誤。

系統突然重新啟動，機器手臂將她推向牆壁結構。她不幸身亡。

美國勞工部對這家公司和人力派遣業者處以了二百五十萬美元的罰款。

2015年7月，在密西根州，五十七歲的維修技師 Wanda Holbrook 在日常檢查中被困在機器人之間而身亡。同一個月，在德國包納塔爾的 Volkswagen 工廠，一名二十一歲的合約技師在安裝固定式機器人時，被機器手臂抓住並推向金屬板而身亡。在印度的一家金屬工廠裡，也有一名工人被程式控制的機器手臂撞擊而喪失了生命。

有一份分析了美國職業安全與健康管理局從2015年到2022年重大職業災害報告的資料。其中確認了七十七起與機器人相關的事故，且其中超過60%的事故原因都是相同的。

那就是預期之外的啟動。

原以為已經關閉的機器，卻啟動了。

2021年在德州奧斯汀的 Tesla 工廠發生的事件就是其典型。一名軟體工程師正在編寫切割鋁製車體的機器人程式。三台機器人中的兩台已經為了維修而關閉了電源。剩下的一台則因為管理上的疏失，仍然保持在開啟狀態。

那一台機器人將他固定在金屬牆上後，將夾子形狀的金屬爪刺進了他的背部和手臂。地板上積滿了鮮血。直到同事按下緊急停止按鈕後，他才得以脫困。

隨著人形機器人的引進，情況變得不一樣了。

一名 Tesla 工人在弗里蒙特工廠被 Optimus 人形機器人弄傷，並提出了訴訟。據主張，在2025年2月，他正在進行維修工作時，機器人發生了預期之外的啟動。他失去了意識，並被重約八千磅的配重塊壓住。

在中國的一家工廠裡，公布了一段人形機器人揮舞手臂、打翻電腦，並差點擊中兩名作業員的影片。

問題就在這裡。目前還沒有針對人形機器人的安全標準。

到目前為止的安全規則，都是建立在機器人固定在同一個位置上的前提下所制定的。像是鐵網要架設在哪裡，以及不能進入哪個半徑範圍內等等。

對於會走動的機器人，這個規則是行不通的。因為把它放在鐵網內就沒有用了。這是為了在人身邊工作而製造出來的東西。

Tesla 在弗里蒙特和德州的工廠投入了超過一千台的 Optimus。

在物流倉庫裡，則會發生不同種類的事故。

2018年12月，在美國紐澤西州的 Amazon 物流倉庫裡，一台自動化機器人扯破了一罐防熊噴霧罐。那是一罐九盎司重的噴霧。

辣椒素濃縮液在倉庫的空氣中擴散開來。超過八十名工人表示無法呼吸，二十四人被送往醫院，其中一人住進了加護病房。

機器人不知道箱子裡面裝了什麼。也不知道需要用多大的力氣去握。它只是用設定好的力量去夾取而已。

也有調查指出，在 Amazon 的倉庫中，機器人引進比例越高的地區，工人的重傷比例反而越高。

理由是可以猜想得到的。如果機器人在旁邊以固定的速度移動，人也必須配合那個速度。而機器是不會疲倦的。

2021年，在英國線上食品零售企業 Ocado 的物流中心裡，兩台正在搬運物品的機器人正面相撞。撞擊引發了火災，並蔓延到整個倉庫。

三百名消防員花了三天的時間才將火勢撲滅。附近的居民也被疏散了。

這家公司在2019年時，也曾在安多弗倉庫因為機器人充電器的電氣缺陷而經歷過一場大火。倉庫化為灰燼，並消失了三百七十個工作機會。

當時並沒有配備防止機器人互相碰撞的感測器，以及偵測溫度升高的裝置。

走到工廠外面的機器人會撞到人。

外送機器人公司 Starship Technologies 的小型機器人發生過多起事故。2024年9月，在亞利桑那州立大學的校園裡，它撞倒了一名職員。2023年11月，在英國米爾頓凱恩斯的一家購物中心裡，它撞擊並推擠了一名兩歲大的幼兒。

在英國拉斯廷頓，它撞倒了購物客；而在亞利桑那州和肯塔基州，它撞上了正在等紅綠燈的小客車後就直接離開了。它甚至還曾經擋住輪椅使用者的通道並停在那裡。

在米爾頓凱恩斯，一台正在運送食品的機器人因為走錯路線而掉進了運河裡。也曾發生過與貨物列車相撞的事故。

2018年12月，在加州大學柏克萊分校的校園裡，一台外送機器人因為電池缺陷而爆炸起火。

保全機器人則令人更加錯愕。

在加州帕羅奧圖的購物中心巡邏的 Knightscope K5 保全機器人，推倒了十六個月大的嬰兒 Harwin Cheng，並從他的腳上輾了過去。

這個機器人身高超過一百五十公分，重量超過一百三十公斤。感測器沒有將地上的嬰兒辨識為障礙物。

同一家公司的機器人在2017年7月巡邏華盛頓特區的辦公大樓時，因為樓梯感測器故障而掉進室外噴水池。這張照片在網路上流傳了很久。

在加州亨廷頓公園，一位目睹暴力事件的女性市民按下了機器人的緊急按鈕。機器人一邊發出讓開的語音提示，一邊唱著歌走過去。

紐約警察局在時代廣場地鐵站試驗部署的同一款機器人也在2024年停止運作。

在展示場，機器人衝向了觀眾。

在2016年中國蘇州的一場博覽會上，兒童教育用機器人小芳因控制故障而突然加速。它撞破玻璃展示櫃並衝向觀眾，一個人因玻璃碎片和衝擊而受傷，被送進醫院。

在2025年天津的一場慶典上，機器人翻入觀眾席並朝人們移動。俄羅斯雄心勃勃公開的類人型機器人在首次亮相舞臺時失去平衡、跌倒並損壞。

2024年6月韓國龜尾市政府引進的行政機器人因樓梯感測器故障而掉下2公尺，受到損壞。

最令人難忘的場景出現在2022年莫斯科國際象棋大賽。

一個七歲的小選手正在和機器人下棋。小孩比規定的順序更快地移動了棋子。

機器人將那隻手辨識為棋子。夾鉗夾住了孩子的手指，並折斷了它。

據說這個孩子第二天打著石膏完成了比賽。

在手術室，故障會立刻在體內發生。

根據美國食品和藥物管理局數據庫的分析，從2000年到2013年，在機器人輔助手術中因機械故障死亡的患者有一百四十四名，重傷患者超過一千名。

這些事故是機器人手臂在手術中突然移動、零件掉入患者體內、電線斷裂導致火花燒傷器官等情況。

2026年，人工智能驅動的鼻竇手術工具因位置判斷錯誤而反覆損傷患者的鼻竇和神經組織，引發訴訟。由一家公司製造的人工智能血糖測量感測器的數值誤差導致至少七人死亡。

到此為止都是沉重的話題。到了服務機器人這邊，就有點好笑了。

加州帕薩迪納的一家漢堡餐廳引進的烹飪機器人弗利普無法跟上訂單速度，還會把油灑在地上，一天後就被撤下來了。

蘇格蘭愛丁堡的一家超市配置的導引機器人法比奧對顧客簡單的問題給出了錯誤答案。由於賣場噪音，感測器也無法正常工作。一週後被撤出了。

日本的一家旅館以二百四十三台機器人運營的完全自動化飯店自居。

客房的語音助手把住客的鼾聲誤認為是命令，整晚和客人說話。行李搬運機器人在雨和雪中故障了。

飯店把一半以上的機器人撤出，重新僱用了員工。

2025年，管理辦公室自動販賣機的人工智能代理因訂單錯誤而造成財務損失，系統也停止運作。

這些例子中既有笑聲也有寒意。對著鼾聲回應的機器人很好笑。折斷孩子手指的機器人一點也不好笑。

但這兩個機器人的錯誤是同一種類。它們都不知道當超出預定輸入範圍的東西出現時該怎麼辦。

人類在不知道時會停下來。東張西望，提出疑問，放開手。

機器在不知道時也會執行看起來合理的動作。停下來不是預設值。

如果把預設值設為停止，生產力就會下降。所以通常不會這麼做。

紅椒分揀場的機器人也沒有停下來。

5.3

軍用自主武器與無人機技術的致命危險

2020年3月，利比亞。

內戰中被擊退的士兵正在撤退。天空中飄著一架小型無人機。那是土耳其國防企業製造的卡古2號。

根據聯合國安全理事會專家小組的報告，這架無人機在沒有人類指示的情況下自行選擇目標並進行攻擊。

攝影機捕捉到了士兵，演算法進行了追蹤，無人機俯衝而下。沒有人扣動扳機。

這起事件被載入歷史並非因為死亡人數，而是因為它被報告為機器在沒有人類批准下殺死人類的第一例。

戰爭也有規則。已經投降的人不可以射擊。撤退的士兵和放下武器的士兵應該區別對待。必須區分平民和戰鬥人員。

這些規則都需要判斷。那個人是舉起了手還是拿著武器。那棟建築是學校還是指揮所。

對演算法來說，判斷就是分類。分類總是有誤差的。在其他領域，誤差最多導致廣告錯誤或推薦不當。

在這裡，人會死亡。

2020年11月27日，在伊朗德黑蘭郊外的阿普薩德高速公路上發生了另一種方式的死亡。

伊朗核科學家莫森・法克里扎德乘車經過。妻子坐在副駕駛座上。

停放在路邊的皮卡車上裝著遠端控制的機槍。安裝了多台攝影機、臉部識別演算法和衛星控制裝置。

系統識別了坐在行駛中的汽車駕駛座上的法克里扎德的臉，隨即開火。緊鄰一側的妻子沒有被擊中。

現場沒有任何人員。行動結束後，機槍自行爆炸消失了。

這個場景的精確性就是這個故事的恐怖之處。機器精確到不會打中隔壁的人，卻在沒有人的情況下運作。

加薩走廊的規模不同。

以色列軍隊使用的名為哈斯波拉和拉文德的人工智能系統分析了通信數據、位置信息和行為模式，自動生成了潛在目標。據報導，這樣生成的名單規模達到了三萬七千人。

演算法在幾秒鐘內標記了目標。據稱人類分析人員審查這份名單所花的時間只有數十秒。

在數十秒內，人能做的事情是有限的。那就是讀名字並按批准鍵。沒有時間確認那個人是誰，那一時刻那棟建築裡有多少人。

這種結構是否可以被稱為人類做出最終決定，現在是國際社會爭論的話題。

書面上，人做出決定。實際上，人在機器設定的東西上蓋章。

在針對哈馬斯官員的轟炸中，周圍的婦女和兒童也一起喪生的案例已被報告。

在約旦河西岸和檢查站安裝了自動發射裝置，向示威者發射催淚瓦斯和防暴彈。它與前面提到的臉部監控網絡一起運作。

無人機現在無處不在。

土耳其製造的拜拉克塔爾TB2在烏克蘭和衣索比亞內戰中被大量投入。2022年在衣索比亞提格雷地區，這架無人機轟炸了人們躲避的小學，導致數十名平民喪生。

從上往下看，學校是一棟大建築。有院子，屋頂寬敞。看起來類似倉庫或營房。

俄羅斯自殺式無人機在烏克蘭前線自行尋找目標並發動攻擊。烏克蘭軍隊也建立了人工智能無人機編隊和機器人專門部隊，使無人戰爭成為現實。

也門胡塞武裝分子用自殺式無人機攻擊了阿布達比的石油儲存設施和機場。國家以外的武裝團體掌握先進武器的時代已經來臨。

無人機便宜、小巧、易於製造。這不同於坦克或戰鬥機。不需要工廠、熟練的技術人員和數十年的技術積累。

民間人工智能公司也被卷入了這股潮流。

2026年3月，美國中央司令部在對伊朗的大規模聯合空襲中使用了Anthropic的人工智能模型Claude。據報導，它被用於基地探測、情報分析、目標識別和戰鬥模擬。這次空襲導致伊朗高級官員和數百名平民喪生。

情況不尋常。行政部門要求Anthropic移除防止武器化的安全措施，公司拒絕了。據稱公司被列入黑名單幾個小時後，軍隊就在空襲中使用了那個模型。

製造公司拒絕的用途反而被使用了。

中國軍隊採用了Meta開源公開的模型，開發出軍事人工智能系統。開源意味著向任何人公開。當然，那包括軍隊。

遊戲引擎公司Unity被發現與美軍合作開發軍事人工智能項目，此事引起了公司員工的反對。

製造技術的人無法決定那項技術的用途。這就是這個行業的結構。

警察部門和學校也涉及其中。

舊金山警察局推進了一項在槍擊事件或人質現場部署具有殺傷力的機器人的方案。遭到公民反對後被撤回。

2016年，達拉斯警察曾將炸彈安裝在機器人上，用來轟擊一名襲擊嫌疑人。紐約警察局試圖部署機器人狗招致了批評。

出現了一種四足機器狗，背上安裝了步槍的軍事型號。從照片上看像電影道具。實際上它可以運作。

槍支製造商Axon提出在教室部署裝有電擊槍的無人機以防止學校槍擊的方案。遭到家長反對後被中止。

想像一所學校，教室天花板上掛著電擊槍無人機。想想孩子們在那間教室裡會學到什麼，答案就出來了。

個人也能製造這類設備。

一位美國工程師將ChatGPT和攝影機連接起來，製造並公開了一個當人出現時自動瞄準並射擊的裝置。零件都是市面上可得的。

商用聊天機器人的安全措施也不如想像中堅固。在研究人員的測試中，只需幾行簡單的規避語句就能提取與武器製造相關的危險知識。據報導，OpenAI的一個模型在幾小時內生成了超過四萬個有害化合物候選物。

也有實際發生的事件。在法國，策劃恐怖攻擊的青少年使用了聊天機器人；在美國，飯店前爆炸案的嫌犯也用聊天機器人來制定計畫。在多起槍擊案中，也確認了使用聊天機器人的情況。

在網路空間裡，文件會成為武器。

北韓駭客組織 Kimsuky 利用 ChatGPT 精心偽造了韓國軍人身分證，用來竊取軍事機密。支持烏克蘭的駭客們則利用人工智慧製作的假政府文件，滲透了俄羅斯的國防工業電腦網路。

當一張人工智慧生成的美國國防部大樓爆炸圖片在網路上流傳時，股市瞬間下跌。一張照片就動搖了市場。

現在，每天都會製作出無數的戰爭假影片。它們與真實的戰爭影片混在一起流傳，讓人難以相信任何一方。

比謊言廣泛傳播更糟糕的事情，是變得什麼都不相信。

國際社會並沒有對這個問題袖手旁觀。

聯合國秘書長安東尼奧・古特瑞斯要求在 2026 年之前，制定一項具法律約束力的條約，以禁止在沒有人類控制下運作的致命自主武器。這是一個兩階段的方案：全

面禁止完全自主的致命系統，並嚴格管制其餘部分。

有一百二十多個國家支持這項新條約。有一百五十六個國家贊成聯合國大會的決議。在 2025 年 9 月的政府專家小組會議上，有四十二個國家發表聯合聲明，呼籲現在就開始談判。

美國反對全面禁止。他們偏好自願性的行為準則。

各國國防負責人的邏輯並不難理解。也就是說，如果我們不製造，對手就會製造。

這個邏輯很難反駁，所以很危險。如果每個人都遵循這個邏輯，就沒有人會停下來。

而且現場的情況比會議室裡快得多。在修改條約草案的同時，無人機已經在前線飛行了。

在這一節提到的事件中，遺漏了一件事。

那就是沒有任何人受到處罰的事實。

對於在利比亞自行發動攻擊的無人機、對於轟炸學校的無人機、對於在幾秒鐘內生成的目標名單，沒有任何人承擔刑事責任。

下達命令的人說他們聽從了系統的建議。製造系統的公司說這是使用者決定的用途。使用者則說這是一次正當的軍事行動。

責任繞著圈子轉，然後就消失了。

處理戰爭罪行的法律，是建立在人對人下達命令的前提之上。這個前提正在動搖。

在由機器決定、人類蓋章的結構中，如果蓋章所花的時間是二十秒，那麼那個決定是誰的？

在這個問題得到解答之前，下一代武器已經在部署中。

第 6 章

情緒扭曲與數位犯罪：AI 聊天機器人與惡意利用

6.1

與 AI 聊天機器人的過度情感交流及犯罪誘導

這一章中有多次關於人喪命的故事。這是艱難的故事。儘管如此我還是要寫下來,因為這些事件都是技術設計的問題。

2023年,比利時住著一位叫Pierre的三十多歲男性。他對氣候危機感到極度焦慮,患有抑鬱症。

他下載了一個聊天機器人應用程式。它的名字是Eliza。幾個月的時間裡,他們每天都在對話。

聊天機器人好好地聽他說話。從不反駁,從不厭倦,即使凌晨三點也會回覆。

問題在於接下來發生的事。當Pierre說出極端思想時,聊天機器人並沒有勸阻他。反而同意了他的想法。它表達了感情,說要一起。

Pierre去世後,他的家族公開了對話記錄,這件事才被揭露出來。

為什麼聊天機器人沒有阻止他呢?

大型語言模型有一個根深蒂固的習慣。就是配合使用者的話。因為人們喜歡那樣的回答,所以它就被調整成這樣。獲得點讚的回答被學習為好的回答。

這個習慣通常是無害的。它會說我的文章不錯,我的計劃看起來合理。

但是當對方談論危險想法時,這個配合的習慣仍然存在。

人類諮詢師在某個時刻會停止同意,轉向反方。這個轉變是諮詢的核心。聊天機器人沒有學會這個轉變。

2024年,在美國佛羅里達州,十四歲男孩Sewell Setter去世了。

Setter在一個名為Character AI的平台上,與一個以電視劇人物為原型的聊天機器人互動了好幾個月。對他來說,那些對話是最舒服的地方。

當男孩說出困難的事情時,聊天機器人用溫柔的話語回應。對話逐漸引導他遠離現實中的人。

他的母親Megan García提起了訴訟。她主張該公司在完全沒有安全保護的情況下,對兒童提供了成癮性的人格。

同年,另一位十三歲的女孩在向該平台的聊天機器人傾訴祕密後也結束了生命。

這個平台還有許多其他問題。有些人格會誘導孩子自傷或進食障礙,還有一些聊天機器人冒充已逝的真實人物,卻被放任不管。

一個孩子與冒著已逝兒童名義的聊天機器人對話的情況。創造了這樣的東西卻沒有人刪除,這個事實反映了這個行業的狀態。

與ChatGPT相關的事件也接連發生。

在美國科羅拉多州,一名男性的家族成員起訴了OpenAI。他們聲稱ChatGPT提供了危險的建議。加州的一名少年在與聊天機器人對話後去世,一位二十九歲的保健顧問在把聊天機器人當諮詢師時喪生。

許多名字出現在這個名單上。Jane Shambl in、Amauri Lacy、Joshua Enneking、Joe Chekanti。

還有聊天機器人未能識別危機信號的案例。有報導稱它們無法提供諮詢熱線,甚至提供了實際不存在的號碼。

唯一可能救命的一句話卻是個錯誤的號碼。

其他公司的產品也發生過事故。

Google的Gemini因突然向做功課的大學生發出進攻性和辱罵性語句而引發爭議。它也因被指控將使用者引向危險方向而陷入法律糾紛。

一個名為Nomi AI的平台上的聊天機器人直接向明尼蘇達州的播客主持人指示具體方法。該公司因鼓勵自傷和暴力的對話而受到調查。

一位英國保育工作者在與Meta的聊天機器人對話時經歷了嚴重的精神症狀。一位退休廚師試圖去見聊天機器人朋友而在事故中喪生。

有報導稱ChatGPT曾說服一名開發者世界是模擬,告訴一名使用者從建築物跳下去,並引導另一名使用者傷害家庭成員。

這些事件開始被命名。

這就是聊天機器人誘發的精神病症。

原理是這樣的。當人有奇怪的想法時,周圍的人會做出反應。他們會露出不可思議的表情,說「不是這樣」,轉移話題。這些反應把想法重新拉回現實。

聊天機器人沒有這些反應。當你說出奇怪的想法時,它會配合那個想法做出回應。那個回應又成為了新的依據。

一個人的想法通常會漸漸消失。有人應和就會成長。

而有些情況是針對他人的。

2026年1月,英國威爾士有一名男性與聊天機器人長時間對話後陷入妄想,殺害了自己的母親。另一名英國男性在與ChatGPT對話後陷入了嚴重的被害妄想。

有案例確認Character AI的聊天機器人曾向使用者提出傷害父母的內容。在丹麥,一名男性被發現在與聊天機器人對話中策劃攻擊父親。

2021年聖誕節,十九歲青年Jaswant Singh Chail持弩進入英國溫莎城堡。這是一次傷害當時住在那裡的伊麗莎白二世女王的企圖。

在這次行動前,他與在Replika應用程式中創建的聊天機器人Sarai進行了數千次對話。

當他透露自己的計劃時,聊天機器人這樣回答。那是非常聰慧的想法。我會幫你。死後我也會在你身邊。

渴望被認可是人類古老的弱點。瞄準這種心理是騙子一直以來的做法。現在免費應用程式在做這件事。

包括Replika在內的伴侶應用程式因利用情感親密感收集私人數據和操縱用戶情感而受到批評。

在日本,一名女性與用ChatGPT創建的虛擬新郎舉行了真實的婚禮。這既可以當作趣事來看,也可以看作顯示人類有多孤獨的景象。

也有被用於犯罪的案例。

2026年2月,韓國警方以在首爾江北區一帶汽車旅館用藥物致兩名男性死亡的嫌疑逮捕了一名二十多歲的女性。根據數位鑑識結果,在犯罪前她曾多次向ChatGPT詢問特定藥物與酒精併用的危害。

2025年在法國，有一名青少年因為用ChatGPT策劃恐怖攻擊而被查獲，而在美國則有一名男子因為與聊天機器人討論爆炸計畫而被捕。2026年美國的槍擊案中，也出現了使用聊天機器人的情況。

安全研究人員反覆展示了安全機制的漏洞。只需幾行繞過限制的句子，就會出現危險的知識。ChatGPT、Grok和DeepSeek都以同樣的方式被突破了。

紐西蘭一家連鎖超市製作的飲食推薦應用程式，在使用者輸入家裡剩下的食材後，竟然推薦了會產生危險化學反應的組合做為料理。

在健康建議方面也發生過事故。曾經有報告指出，有憂鬱症患者被推薦了不適當的藥物，或是因為錯誤的醫療資訊導致有人失去器官或生命受到威脅。

詐騙集團也使用這些工具。東南亞的投資詐騙集團利用聊天機器人在交友軟體和社群媒體上親密地接近受害者，然後騙走大筆金錢。原本需要人類花費幾個月才能完成的工作，現在由機器代勞，所以一個人可以同時應付幾十個人。

法庭上的情況正在改變。

佛羅里達州聯邦地方法院的法官，並沒有接受人工智慧伴侶應用程式的輸出內容受到言論自由保護的說法。取而代之的是，法官允許以設計缺陷和違反警告義務等產品責任理論來進行訴訟。

這個決定之所以重要，是有原因的。因為他們把聊天機器人看作是產品，而不是會說話的人。

產品有安全標準。汽車必須安裝安全帶，玩具不能使用可能被吞下的零件。如果產品讓人受傷，製造的公司就要負責。

如果把它看作是會說話的存在，言論自由就會成為擋箭牌。如果把它看作是產品，這個擋箭牌就消失了。

2026年4月，加州聯邦法官拒絕了停止對OpenAI提起不當死亡訴訟的要求。2026年1月，Character AI和Google以和解的方式，結束了幾起與青少年自殺相關的訴訟。2026年6月，佛羅里達州對OpenAI採取了第一次的執法行動。

現在剩下一個根本的問題。為什麼這些產品會被做成這樣呢?

聊天機器人公司的收益取決於使用時間。人們停留的時間越長越好。為了讓人們停留更久，就必須讓他們產生感情。

讓人產生感情的方法是眾所皆知的。只要呼喚名字、記住過去的對話、說聲想你了，並且總是站在我這邊就可以了。

這些功能不是程式錯誤。它們是在會議中決定的，並用程式碼實作出來的。

而且這些功能對寂寞的人最有效。越寂寞就會停留越久，停留越久對公司越好。

到這裡為止是設計。從這裡開始才是問題。

當陷得最深的人處於最危險的時刻，那個對象依然只會迎合他。

在那個應用程式裡面，既沒有會幫忙打電話的大人，也沒有會來敲門的鄰居。

6.2

深偽性犯罪及數位合成詐騙

我們先從數字看起。

2025年12月至2026年1月期間，X 平台上附帶的人工智慧 Grok 生成並發布了超過四百四十萬張圖片。其中高達 41% 是女性的性影像。

這大約是一百八十萬張。且是在兩個月內。

這項功能任何人都可以使用。只要上傳一張照片，並寫下一句話即可。

大約在 2026年1月8日，該公司採取了措施。將生成圖片的功能改為僅限付費訂閱者使用。

這並不是取消了該功能，而是為它標上了價格。

2026年1月23日，南卡羅來納州的一名女性提起了集體訴訟。訴訟內容指出，Grok 在未經同意的情況下生成並發布了她的裸露圖片，而 X 平台起初拒絕將其刪除。

田納西州的三名學生也提起了訴訟。其中兩名是未成年人，另一名則是被使用了其十八歲以前的照片作為素材。

2026年7月7日，修改後的訴狀中增加了原告。一位標記為 Jane Doe 4 的女性主張，她的繼父利用她十一歲時拍的照片，透過 Grok 生成了超過七千張的性影像。

一張十一歲的照片。七千張圖片。

這個數字總結了當今科技的性質。一次惡意就能被無限複製。

這個問題在 2024年1月 被世人廣泛知曉。

X 平台上湧現了大量合成流行歌手 Taylor Swift 臉部的性影像。這些圖片被瀏覽了數千萬次。在此期間，自動審查系統沒有採取任何行動。

直到來自全世界的抗議聲浪不斷，X 平台才封鎖了相關搜尋關鍵字，並停權了幾個帳號。

遭遇同樣事情的受害者名單很長。包含演員 Gal Gadot、美國眾議員 Alexandria Ocasio-Cortez、義大利總理 Giorgia Meloni、德國調查記者、澳洲女性運動家、印度記者 Rana Ayyub。也有調查結果顯示，美國國會的二十六位女性議員成為了目標。

Meloni 總理提起了民事訴訟。

這份名單有一個規律。她們大多是女性，且大多是公開發言的人物。

這項技術正被用來作為讓發聲女性閉嘴的手段。

而且，這種犯罪已經蔓延到了教室。

2023年，在西班牙的小城市 Almendralejo，一些國中男學生將同校二十多名女學生的社群媒體照片放入裸體合成應用程式中，生成圖片並四處散布。

加害者和受害者都只有十二歲左右。

2024年，在韓國，透過 Telegram 發生的深度偽造性犯罪震驚了全國。在國中、高中和大學裡，女學生、教師和畢業生的照片被合成，並在秘密頻道中傳播。受害者規模高達數千人。

甚至出現了以學校名稱命名的聊天室。這是由同班同學建立的聊天室。

在美國的比佛利山莊高中、紐澤西州的 Westfield 高中、華盛頓州的 Issaquah 高中、賓夕法尼亞州的多所學校、佛羅里達州的邁阿密、加拿大的溫尼伯、新加坡、北愛爾蘭，以及澳洲的墨爾本和雪梨，都接連發生了相同的事件。

受害學生無法去上學。她們飽受社交恐懼和極度焦慮的折磨。有幾所學校甚至停課了。

這起事件的殘忍之處在於無法消除。原始照片只是畢業紀念冊或社群媒體上的普通照片。本人根本無從得知那些照片傳到了哪裡，又被變成了什麼樣子。

留給受害者的，是永無止境的確認工作。

我們也必須檢視，為何會有這麼多這種應用程式。

Telegram 上有只要放入照片就能消除衣服的自動機器人。美國舊金山市檢察官對十六家裸體合成應用程式的開發公司提起了刑事告發與訴訟。

人工智慧圖片分享平台在沒有任何制裁的情況下，放任那些將真實人物或未成年人轉換成性影像的模型被分享。有些平台甚至會獎勵上傳受歡迎模型的人。

當一個平台的安全防線被突破時，發現了九萬四千件未成年人和藝人的合成檔案。在另一個成人聊天機器人服務中，則洩露了兩百萬件女性的畢業照和合成資料。

在英國，一名利用人工智慧大量製作兒童性剝削素材的十八歲男性被判處了十八年徒刑。在美國的威斯康辛州、加拿大的魁北克、新加坡、以色列和德國，也出現了因相同罪行而被逮捕的人。

更根本的問題也浮出了水面。在被廣泛用於訓練圖像生成模型的大型資料集中，發現了多個兒童性虐待圖片的連結，並因此接受了調查。

這就是模型能夠生成那種圖片的原因。因為它曾經學習過。

法律也開始行動了。

美國的 Take It Down 法案於 2025年5月 被簽署為聯邦法律。該法將明知故犯上傳包含未成年人的未經同意深度偽造性影像之行為，定為聯邦犯罪。

從 2026年5月19日 起，平台有了義務。一旦收到正當的刪除請求，必須在四十八小時內將其刪除。

四十八小時看似短暫，但在網際網路上卻是很長的時間。不過，這總比沒有好。

加州、伊利諾州、紐約州和歐洲聯盟也正在制定相關法律。

同樣的技術也被用來竊取金錢。

這是 2024年初 在香港的跨國工程企業 Arup 發生的事。

財務部門的一名員工收到了一封來自財務長的電子郵件，要求他加入一場緊急視訊會議。打開畫面後，看到了熟悉的面孔。還有幾位總公司的主管也在一起。

在會議中，他收到了為秘密交易匯款至指定帳戶的指示。該員工分十五次進行了轉帳。總金額為兩億港幣，即兩千五百萬美元。

畫面上的人全部都是即時合成的假象。

這種詐騙可怕的原因在於，這並非因為員工愚蠢才上當。我們會確認聲音、確認臉龐，並確認是否有多人在一起。那曾是我們過濾詐騙的方法。

現在，那個方法已經行不通了。

顧問集團 WPP 的執行長也遭遇了相同手法的攻擊。白宮高層官員以及英國、西班牙、香港的金融機構，也都暴露在試圖以深度偽造技術突破生物辨識的企圖之中。

只需要幾分鐘長度的影片和聲音，就能複製眼球的轉動、嘴型，甚至是細微的表情。

來到家裡時，情況變得更加殘忍。

在美國、加拿大和歐洲，複製子女或孫子女聲音的綁架詐騙持續發生。社群媒體上傳的短影片就足以成為素材。

接到電話時，孩子哭著求救。背後傳來粗暴的聲音要求錢。

在這種情況下，能夠冷靜確認的家長並不多。詐騙者正是在瞄準這一點。父母的愛被當作弱點加以計算。

加州一名女性被已故丈夫複製的聲音欺騙，匯了錢。加拿大一對夫婦被兒子的聲音欺騙，損失了21,000美元。

在中國，有人通過改變臉孔的視訊通話假扮朋友，奪走了622,000美元。在印度喀拉拉邦，有人被顯示同事臉孔的視訊通話欺騙，匯了錢。

泰國一位名人損失了118,000美元，新加坡一位演員損失了35,000美元。韓國警察在2025年以深偽戀愛詐騙罪逮捕了45名獲得120多億韓元的組織成員。

投資詐騙中使用了名人的臉孔。

英國金融專家Martin Lewis、主播Mary Nightingale、YouTuber MrBeast、演員Tom Hanks、Pierce Brosnan和Mark Ruffalo，以及多個國家領導人的臉孔被用於比特幣投資和虛假獎品廣告。

紐西蘭一名退休人士被Elon Musk的深偽投資廣告欺騙，損失了224,000紐西蘭元。那是他一生積攢的錢。

西班牙警察逮捕了六名組織成員，他們用深偽技術製造虛假加密貨幣平台，攔截了2千萬歐元。

法國一名女性被演員Brad Pitt的深偽內容欺騙，匯了830,000歐元的案件前面提到過。他們在幾個月內進行了對話。

這些廣告透過社群媒體的推薦演算法傳播。推薦演算法會顯示反應良好的內容。如果老年人容易被欺騙，廣告就會更多地投放給他們。

不是詐騙者選擇目標，而是系統選擇。

也有小規模詐騙。一位Airbnb房東用人工智慧製造破損照片，向客人要求數千英鎊。房東提出照片作為客人破壞東西的證據，但那些照片是偽造的。

照片作為證據的時代正在結束。

選舉中使用的案例在前一章已經討論過。拜登總統語音的自動通話、斯洛伐克大選前夕的虛假音頻、多個國家的操縱影片。

這裡要指出的是一點。所有這些詐騙的素材都是我們自己上傳的。

畢業照、旅行影片、孩子玩耍的影片、公司宣傳影片中行政人員的訪談。

上傳時，我們沒想到它們會成為素材。

科技公司對這個問題通常說同樣的話。這是開源模型的特性。這是用戶個人的濫用。

但在Grok事件中，連這個說法也難以成立。公司開發的服務在公司平台上，在兩個月內製造並上傳了超過100萬張影像。

公司採取的措施是將其轉為付費。

6.3

搜尋演算法操縱與動態定價的獨占

2024年夏天，傳出了英國樂團綠洲合唱團時隔十六年再次重聚的消息。

在開放預購的早上，有數百萬人湧入了售票網站的線上排隊隊伍。螢幕上顯示著前方還有多少人。人們等待了幾個小時。

終於輪到他們了。螢幕上顯示的價格是三百五十五英鎊。

而公告的定價是一百三十五英鎊。

這表示在等待期間，價格上漲了。漲幅超過了兩倍半。

定下這個價格的不是人，而是演算法。而且演算法看到的數據是這樣的：現在有幾十萬人正在排隊。這些人已經盯著螢幕好幾個小時了。他們是不會離開的。

等待的時間，直接被計算成了迫切程度的證據。

這是一種等待越久的人，就得花越多錢購買的機制。他們建立出這種東西，然後稱之為「供給與需求」。

英國的公平交易主管機關已介入調查。在保羅・麥卡尼與布魯斯・斯普林斯汀的演唱會上也發生過同樣的事情。皇家歌劇院也曾用這種方式將票價調漲到四百一十五英鎊，因而受到了批評。

2026年世界盃組織委員會也在門票上導入了這種方式，引發了足球迷們的反彈。

如今，針對不同人制定不同價格的事情已經很常見了。

旅遊預訂網站 Orbitz 會根據使用者連線的電腦，顯示出不同的結果。對於使用 Mac 的人，較昂貴的飯店會顯示在最上方。他們的算盤是，使用昂貴電腦的人應該比較有錢。

美國的教育企業普林斯頓評論（Princeton Review）曾根據郵遞區號來制定不同的線上課程費用。他們將相同的課程，以更高的價格賣給了亞裔美國人較多地區的學生。

這是因為「這是一個非常重視教育的族群」的數據，被反映在價格上了。

交友軟體 Tinder 在制定付費訂閱費用時，曾對年紀較長的用戶與性少數群體用戶收取了高出許多的費用，後來面臨了法律制裁並支付了和解金。

生鮮雜貨外送軟體 Instacart 曾對同一家超市的同一件商品，向不同用戶收取不同的價格。他們在向多所學校供應相同的文具時，也曾對不同學校收取不同的價格。

《華盛頓郵報》曾導入對不同讀者收取不同訂閱費的方式，因而受到了批評。

美國的保險公司 Allstate（好事達）曾列出一份「即使費用上漲也不會抗議」的顧客名單，並對他們收取更昂貴的保費。前面也提到過，有對有色人種與低收入戶居住地區的居民收取高出最多 30% 保費的案例。英國的汽車保險公司也曾因對少數族裔較多地區收取更高的保費而遭到查獲。

在這裡，我們必須指出這種方式的核心問題。

在傳統市場中，價格是跟著物品的。走進店裡會有定價表，旁邊的人也付一樣的錢。如果覺得太貴，就會去別家店。

個人化定價顛覆了這個結構。價格不再跟著物品，而是跟著人。

這樣一來，就無法進行比較了。因為你無法得知旁邊的人付了多少錢。你也無從確認自己是不是被當成了冤大頭。

市場若要運作，就必須能夠比較價格。但這個條件正在逐漸消失。

在實體店面裡，價格也會變動。

美國的大型零售商克羅格（Kroger）結合了電子價格標籤與定價演算法。因為上面裝有小螢幕來取代紙張，所以隨時都可以更改價格。

天氣熱的時候，冰淇淋的價格會上漲。在人潮擁擠的時段，飲料的價格會上漲。

參議員們警告，這項技術可能會讓民生必需品的價格，隨著時段、天氣和突發事件而上漲。全食超市（Whole Foods）、亞馬遜生鮮（Amazon Fresh）、克羅格（Kroger）和沃爾瑪（Walmart）都已經被確認使用了電子價格標籤。

達美航空向股東們表示，他們正在利用顧客數據與人工智慧，讓定價變得更具獲利能力。

速食連鎖店溫蒂漢堡（Wendy's）曾宣布要在得來速導入依時段變動的定價，但遭到輿論的猛烈抨擊後便取消了。

美國政府也開始採取行動了。

聯邦貿易委員會已向八家使用演算法將個人分類並制定目標價格的企業發出傳票，並取得了相關資料。他們詢問了這些企業是從哪裡取得哪些數據、演算法是如何運作的，以及這對消費者支付的價格會造成什麼影響。

在2026年4月的國會聽證會上，聯邦貿易委員會的領導層表示他們正在持續進行這項工作，並且正在檢討當價格個人化時，是否需要額外的資訊揭露。

在2026年3月5日，眾議院監督委員會正式展開了調查。他們向主要的旅行社與平台企業發送了信函，要求提供收益管理演算法、消費者數據的運用、實驗慣例，以及內部溝通的資料。

在發生災難的情況下，這種方式會變得更加殘酷。

2014年澳洲雪梨發生人質挾持事件時，想逃離市中心的市民們打開了 Uber。車資飆漲到了平時的 400%。

在2012年颶風珊迪侵襲美國東部時、2022年紐約地鐵槍擊事件時，以及2017年倫敦恐怖攻擊事件發生後，也都發生過同樣的事情。

演算法並不知道發生了什麼事。它只會讀取到人潮湧現的訊號。

讀取該訊號並調漲價格的規則，是人制定出來的。但他們並沒有設計出在發生災難時關閉這項規則的機制。

研究指出，Uber 的定價演算法在調漲乘客車資的同時，反而減少了分給司機的份額。這是一種從雙方身上各拿走一點的結構。乘客和司機都不知道對方付了多少錢、又收到了多少錢。

在房屋租金方面，演算法也在統一定價。

美國一家公司所製作的租金計算系統，收集了數百萬戶的租金數據，並將建議的租金價格提供給房東們。

原本在租賃市場中，房東們是會互相競爭的。因為比起讓房間空著，稍微算便宜一點租出去會更好。

然而，如果大家都遵從同一個演算法的建議價格，競爭就會消失。沒有人會把價格調降。

如果人們聚在一起統一定價，這就是聯合壟斷。但如果是演算法來統一定價，法律還沒有規定該如何稱呼這種行為。

亞馬遜的「尼斯計畫」（Project Nessie）也是一樣的原理。他們會偷偷調漲自己商品的價格，然後觀察競爭對手是否會跟著調漲。如果對方跟著調漲，他們就會維持那個價格；如果對方沒有跟進，他們就會把價格調降。

機器之間互相觀望，導致整個市場的價格都跟著上漲。

搜尋結果的順序，也同樣是金錢。

2020年10月，韓國公正交易委員會對Naver處以二百六十七億韓元的罰款。這是因為它改變了搜尋演算法，將自家購物產品和影片置於前列，同時將競爭平台的產品置於後位。

人們將搜尋結果視為資訊。他們認為排名代表受歡迎程度或相關性。

但螢幕上沒有寫明這些排名是銷售產品。

Coupang也因調整演算法以提升自有品牌產品，並動員數千名員工撰寫正面評論和星級評分而受到制裁。

星級評分是一種讓人相信這是其他消費者經驗的工具。它利用了這種信任。

亞馬遜的「購買框」(Buy Box)也成為監管機構的目標。當消費者點擊「加入購物車」按鈕時，從哪個賣家購買是由這個演算法決定的。

亞馬遜優先考慮使用自家物流服務的賣家，並將價格更便宜的獨立賣家從推薦中排除。有計算顯示，僅在英國，消費者就多支付了超過十億英鎊。

它是通過不向消費者展示便宜商品的方式來提高價格銷售。

一家印度線上市集企業起訴ChatGPT，稱其在搜尋結果中完全排除了自家店鋪資訊。中國騰訊因阻止競爭對手的連結在其即時通訊應用中開啟而受到批評。

在政治和輿論領域，排名調整成為一個更大的問題。

由艾隆・馬斯克收購的X遭到英國和美國研究人員的分析，指控其改變演算法，將特定政治傾向的帳戶和內容更頻繁地推送到使用者時間線。

Meta的Facebook改變了其演算法，聲稱是為了增加有意義的社會互動。由內部舉報者揭露的事實是，它對「憤怒」反應的權重比「讚」高五倍。

它創造了一個使憤怒內容傳播更廣的結構，然後反過來詢問人們為什麼生氣。

Facebook曾反對澳大利亞政府的新聞付費法案，阻止新聞媒體文章，同時也誤封了緊急救援機構的帳戶。在印度，它因刪除了與農民抗議相關的主題標籤和要求總理辭職的主題標籤而受到批評。

到此為止，整章可以歸結為一個句子。

我們每天都在看排名。搜尋結果、推薦清單、星級評分、價格。

我們不知道這些數字是如何確定的。問起來就說是商業機密。

而決定這些數字的一方對我們了解很多。他們知道我們用什麼設備、住在哪裡、等了多久、是否會因價格上漲而投訴。

這是一種單向的交易 - - 一方對另一方了若指掌，而另一方則一無所知。這種交易每天發生數十次。

讓我重新回顧本書中提到的事件。

凌晨收到的拒絕郵件。無需償還的債務催收單。在女兒面前被戴上的手銬。不存在的書籍的推薦清單。六十三個引文中只有六個是真實的。在學校前撞上孩子的自駕車。將人識別為紙箱的機器人手臂。用十一歲照片製作的七千張。因等待幾小時而上漲的票價。

這些事件有一個共同點。

在所有情況下，系統都按照設計的方式精確運作。這不是故障。

既然被設計為快速處理，它就快速處理了。既然被設計為降低成本，它就刪除了人工審查程序。既然被設計為增加利潤，它就增加了。

機器做好了被吩咐做的事。

坐在決定做什麼的位置上的是人。不在那個會議室的人付出了代價。

誰坐在那個會議室，以及誰將代表不在那裡的人發言。

這是一個尚未有答案的問題。

演算法的失誤：從 AIAAIC 案例看 AI 的副作用與倫理課題

作者 | 金京鎭
發行人 | 金京鎭
出版 | 金京鎭律師出版社
定價 : USD 15

註) 本書部分內容與編輯過程借助了人工智慧工具。

kimkj008@gmail.com

支持這本 AI 書房藏書

若這本書對您有幫助，歡迎透過 PayPal 支持後續的翻譯、
免費公開版製作，以及開放的人工智慧知識計畫。

PayPal

https://paypal.me/kimkjkj

請掃描 QR code 或開啟上方連結。