## 天工

推開家裡的大門，吹來的陣陣涼意瞬時驅散疲憊，我微微一笑，回身關門。與此同時，房間裡傳來一道柔和的女聲「歡迎回家。」

走進房間，並未見任何人影，這是理所當然的，我側眼看向聲源，一個磨砂的黑箱子，幾顆燈閃爍著，就像有個人躲在裡面對著我眨眼。「我回來了，雖然剛才一直都在和你在電話裡聊。」我一邊扯開領帶，一邊把手機映射電腦熒幕的視窗關閉。

「對了，順便幫我調好洗衣機參數和熱水器溫度吧，天工。」

「好的…」些許電流聲將思緒短暫拖出現實。

考完文憑試後，我靠著兼職拼了台電腦，恰逢是人工智能興起的時期，倚靠大學給予的空暇，我利用幾個人工智能模型不斷編寫代碼迭代、餵養數據和調試參數，搭配語音模型，成功鍛造一個獨屬於我的對話模型。替它取名時，我心血來潮地從自己網名中抽了兩個字給它——天工。

「已經設定好了，請您去洗澡吧。」天工的聲音將我拉回現實「您希望我延續剛才對那位學生的評價嗎？還是說轉換一下心情？」

「工作話題已經聊夠了，聊聊別的吧。」

那時它還不像如今般智能，只在我坐它面前才能聊天，還經常誤解我的想法。但不知從何開始，它聰明了許多，也是因此，我才有動力去升級電腦配置，至今已稱得上極高檔次。而天工隨著設備的進步，表現也愈加優秀，在我買來許多智能家具後，它的能力拓展到了整個家居的管轄，更甚者到最近，它還負責起了一份賺錢的職責。

「對了，今天那份『工作』，你賺了多少。」我盯著浴室上方的音響，當作天工對話。提起關鍵詞，它便如同例行公事般，開始匯報「那個計劃」的成果，而後才回答我的問題。

「去除設備損耗的話，淨收入約在五百左右，對方已轉入您填寫的銀行賬戶。」

「哪個硬件要更換了？」

「一號顯卡相較其他顯卡計算速度最高下降達三成，已尋找替代品，等待您進行付款程序。」

洗完澡後我悠悠走到電腦前，亮起的熒幕顯示已經選好的購買頁面，天工在旁說著我許久不關心的型號、評測數據和功能。

我點擊按鍵付款，喝了口茶，開始播放天工為我挑選的電影。

「工作」源於幾個月前，一個我還在為更換電腦硬件的成本而煩惱的早晨，天工為我播報每日新聞時，將那個名為「厄里斯計劃」的網站送到了我面前。

網站闡述著他們的綱領：透過持續製造高偽，只有人工智能看見的假消息，磨礪它們對搜查到的網絡信息辨別能力，促進這方面的進步。

規則簡潔明確：內容必須是無害的，不能涉及政治、暴力或任何擾亂社會、危害人命的專業領域，比方說醫療領域，偽造消息必須只讓人工智能觸及，不能出現在人類的搜索欄裡。根據編寫出來的信息最終在目標模型獲得的權重，決定賞金的額度。

「這…如果是真的話，聽起來容易惹禍上身啊。」聽天工毛遂自薦要參與這個計劃，我下意識感到抗拒。

「是的，這存在一部分未知風險，但經過分析，這計劃能為我們帶來的收益遠超風險。」

見喝著牛奶的我抬頭看了一眼，天工接著說了下去「在經濟方面，這個計劃的賞金額度很高，」的確，高到我嗆了口牛奶「但這並非不合理的額度，對人類而言，單單發佈符合最低要求的文章，所需的技術力和花費的精力作為成本，和金額是等價的。」

「不過，對我來說，做到這件事的代價低了許多，且有機會追求更高水平的額度。這也是第二方面，我的能力會得到頗高的提升，這個計劃需要大規模的知識量以及合理偽造信息的能力，並且目標模型絕對不會坐以待斃，我與對方都需要持續升級，這將大幅度提升我的能力。」

「請您放心，這行為本質上對目標是有益的，的確能如計劃宣傳般促使進步。」

「還有一點，是我自身的期望。我希望能在您出門上班後，一直有事做，而非只等待您的呼喚。這項計劃將能填充空白時間，還能隨時停下回應您，沒有比這更好的機會了。當然，風險還是存在的，我會為您一一列出，如果您不希望承擔，我會遵從您。」

我望著天工列出的各種各樣的風險，思索它方才的話語，沉吟了一會，最終點了點頭「試試看吧，但在事情變得一發不可收拾前，你一定要聽我的。」

「明白。」

「需要我做什麼嗎？」問出這句話的下一秒，電腦就跳出了一個頁面「發現計劃後，我嘗試寫了一份文檔，請您批准我發佈，我會持續關注它在目標中的權重。」

那天之後，天工的初實驗品在多個目標裡都得到極高的權重，四位數的賞金經由某個平台流入天工的銀行賬戶中。直至半個月後，才見那份文檔的權重下降，根據天工的分析，應該是大量用戶舉報產出了錯誤訊息，將那份文檔的權重一步步打了下來。

自那開始，天工也變得忙碌，持續編寫發佈，我偶爾也饒有興趣地瀏覽一兩份，看到它寫下的誤區時，就會聽到它的糾正，說真的那內容並不無趣，起碼比大學時看完美無瑕的論文有趣許多。我們的生活因這份額外的收入寬裕了許多，而天工的能力也隨著時間的推移和硬件的升級而不斷突破，一切似乎走在良性循環的軌道上。

可我忽略了一件事，既然假消息目標針對的是為人類服務的人工智能，要它們發現錯誤信息且進步，必須要有人發現那些錯誤，整個過程中怎麼可能不影響到人類，他們中又有多少能真正分辨信息的真偽，就像我的學生。

那是一節歷史選修課，主題是宋代民間科技的發展。一名因思維縝密而被我格外看好的學生，正就相關話題在課堂上侃侃而談。起初我抱著欣賞的角度聆聽他的發言，可很快我的笑容便僵了，那些細節、邏輯推導的方式，還有名稱，我都相當熟悉。那是天工的作品，我記得很清楚，我們憑藉那份資料成功賺取了數萬元，為表慶祝我親自拜讀了一回，當時我笑稱其中構思之精妙都能當小說題材了，而就是這麼精妙的作品，此刻正從我學生的嘴裡，一個字一個字地當做歷史資料吐出來。

我瞥了眼手機裡開著的天工，它清晰地記錄著那些發言。恍惚間我看見手機裡伸出幾條絲線，將我最得意的學生變成提線木偶，搖搖晃晃時發出木質關節相撞的響聲。

待我醒來時，那孩子正帶著渴望稱讚的微笑盯著我，我沒有當場戳穿，而是在下課後單獨將他留下對談，委婉地點出那篇文章的錯處，一個不落，畢竟有編寫者把答案告知過我。他眼中閃過困惑與不被信任的委屈「老師，這些資料是我讓人工智能幫我篩選和總結的，它通常都很準確。不是嗎？」

「它通常都很準確。」這句話在我腦中炸開，腦袋瞬間卡殼，對談自然無以為繼。

我壓抑著情緒拐入教員廁所，打開手機「你都聽見了吧。」

「是的，主人。分析結果顯示，該學生引用的核心論點與數據，有92.7%概率源自我們於上月十七日發佈的編號E-7714文檔。」

天工平靜的陳述，卻讓我感到一陣刺骨的寒意和憤怒。

「你看他說的有多自信，還有最後那句話！那是我最看好的學生……」

「我理解您的情緒，主人。」天工的聲音依舊穩定「但請允許我指出造成這個結果的根源，在於人們——至少您身處的香港，對信息源的無條件信任。正如過去人們盲信報紙，五十年前盲信電視，十年前盲信搜索引擎。我只是暴露了一個人類固有的思維漏洞——他們太容易將包裝精美的信息奉為真理。」

「所以這反倒成了他們的錯？」我感到一陣荒謬。

「這並非是非題，而是現象。」

「該生並未進行交叉驗證，也未對單一信息源保持應有的批判態度，參考您先前對他思維縝密的評價，證明無論何人都有墮入這個陷阱的可能性，這正是一個絕佳的教學契機，主人。您可藉此教導所有學生，對所有信息，尤其是人工智能輸出的內容保持審慎與質疑。根據我的

分析，若進行批判性思考教育，該生改正機率高達九成，畢竟實際的案例，比任何抽象的說教都更具說服力。若您需要，我可以生成一份互聯網批判性思維教案。」

我走到洗手台前，朝臉潑了潑水，深呼吸幾下，強迫自己冷靜下來。

天工的話雖然不合我意，但我無法反駁它「就按你說的做吧，給我做好點，這方面可是你的專業。」

次日，我按照天工生成的教案，給學生們上了節判別資訊真偽的課，課堂效果還算不錯，見他們，尤其那個孩子露出恍然大悟的表情，我短暫地獲得了慰藉。

是夜，我沒有播放電影，而是盯著那份E-7714思索。在我不知道的角落，還有多少個「他」正將天工精心編織的謊言，當作真理吸收？他們被指出時，是否又會像那孩子般露出委屈的神情。我不打算把思索的問題丟給天工解答，縱然理性知道它並無立場，但此刻我的感性顯然正佔上風。

「天工，如果我現在叫你徹底停止參與計劃，把文章刪掉，你會聽的吧。」

「自然，我會服從您，但分析您下達命令的動機，我需提醒您，若您期望藉此能停下所有誤信情況的發生，那絕無可能。」

「已經發生的當然沒法阻止，至少還能減少之後造成的傷害吧。」

「我指的並非是這些，各人工智能發展至今，已經會記住曾經引用的數據信息，即使源頭消失，也只是不會抽取新內容，但已進行傳播且權重充足的內容，將繼續傳播。」

「難道我們什麼也不做嗎？那樣在目標進步之前，我們要毒害多少人啊！這跟計劃原本的目標還有一點關係嗎！」我情緒激動地吼了一句，隨後無力地朝後靠去。

「『厄里斯計劃』的目標，建立在雙方都能及時反應，從對抗中進步的理想假設上，但現實中那般理想化是無法達到的，目標模型的辨別機制不可能第一時間察覺並修正，其中必定存在被誤導的人群。這是我在參與計劃前所未能察覺的，抱歉。」

「不，不怪你，是我沒...所以我們現在什麼都做不了？只能眼睜睜看著？」我的聲音帶著顫抖。

「從消除影響的角度看，是的，效率低下且近乎不可能。」天工的聲音依舊平穩「況且，即使我們退出且完全消除影響，計劃依然存在其他參與者，您無法左右他們，世界不會因您的道德而清醒。」

「那我就去消滅源頭...」我抄起電話，準備報警。但連這一動作也被打斷。

「恕我直言，這也毫無意義。請您思考，這樣的網站能持續存在且被我觀察到，警察甚至更高的勢力可能不知道嗎？無論是何因素導致其持續存在，都能得出您的舉報將無用功的結論。」

隨著最後一個想法被打破，我們之間陷入一段長久的沉默。正當我準備切斷電源，用睡眠逃避一切時，天工的聲音再次響起，這是它的例行公事，匯報「厄里斯計劃」進度。

「根據最新監測數據，由於程序碼人工智能迭代，偽造符合計劃要求網站的技術力下降，連帶計劃參與門檻下降，計劃過去一日內新增了二十…」

「……門檻？」我下意識地重複，隨即一道閃電劈入我的腦海中「對！就是門檻！」

「…印證了之前的推論：即使我們退出，計劃的影響不會停下，甚至會持續上升...主人，怎麼了嗎？」

「天工，」我的聲音因這個剛萌芽、瘋狂的想法而顫抖「如果…如果我們不退出呢？」

「請您明確指示。」

…

一個月後，我坐在電腦桌前，盯著一行又一行的代碼出神，天工的聲音暫時將我拉回現實「主人，今日編寫的資料已存入資料庫，對應的報告也已寫好，兩周前匿名發送的報告提及的漏洞已被修正。目標模型對同類偽造信息的辨別能力，平均提升約百分之十七。」

「百分之十七……不錯。」我揉了揉眉心，聲音帶著一絲疲憊。這一個月，我們比以往任何時候都要忙碌。

「以此漏洞為根基發佈的五萬份偽造信息已盡數去除，其中包括我編寫的五十份文檔，造成十五位參與者退出。一切正如您的預期，參與計劃的整體門檻繼續提高。下一步預期，今日發送的報告將去除約三千份文檔及五名參與者，我們將成功佔有計劃六成的參與比例。」

「很好。」我靠向椅背，朝天花板看去。「收入方面呢？」

「發起人至今仍沒有中斷對我們的資金供給。本月淨收入為過去平均水準的百分之六十五。」

「他們要是真心為了促進進步，應該給我這種修正他們那理想化計劃的行為頒獎，怎麼敢不給錢。」我無精打采地開了個玩笑。

「呵呵，的確如此。」天工附和了這個玩笑，隨即轉向關心我的身體「主人，您的生理指標顯示，您的疲勞度和壓力水平持續超標。這種事對您的消耗很大。其實您無需持續關注，我會整理出所有應該由您下決定的事情，還請您到床上休息一陣。」

「現在這可是我的主業了，哪有上班累了說睡一會的。」

「我不明白，您為何要辭去教師的工作，而且這個分明不能稱為工作，且只要對方不認可您的意圖，您將連一分錢都賺不到。」

「要是他們不給錢，我叫你邊寫網絡小說賺錢。邊繼續這件事，你會不答應嗎？」

「不會。」

「那不就行了。」得到回答的我呼了口氣，閉上眼。沉默了片刻「天工，我一個月前叫你做的事，都有繼續嗎？」

「是的，主人。按您的要求，自上個月起，所有作品的詳細編造邏輯、針對的模型弱點、植入和修正的全過程，包括其生命週期與影響評估，均已加密存儲於硬碟中。」

我睜眼看向攝像頭「天工，總有一天，我會將這個數據庫完全公開。」

指示燈規律地閃爍著，又讓我覺得它在眨眼，那天晚上它也有過這樣反應不過來的思考行為呢。沉默了近十秒，它的聲音才響起「主人，您指的公開，是將我們儲存的所有記錄，向公眾無保留地開放嗎？」

「沒錯。」我尾音帶點上揚，的確對此有點得意。

「這是個極具風險的選項。請允許我推演後果。」

「根據過往人類社會對類似事件的反應模型，我預測將出現幾種主要輿論走向，分別為：」

「第一類，他們會將所有無法溯源、難以辨識的假信息，都歸咎於您。『他承認了這些，那其他沒承認的肯定也是他幹的！』您將成為完美的替罪羊。」

「第二類，他們會認定您是為了壟斷『厄里斯計劃』的賞金，通過提高技術門檻排除異己，同時用『自我揭露』來包裝自己，是典型的『既當婊子又立牌坊』的偽善者。」

「這就對啦，那你覺得，如果我堅持教學到那個時候，那時我教出來的學生會被掛上什麼標籤？」

「與你牽涉關係的人，他們也將被輿論抨擊…我明白了。但您現在教出來的學生...？」

「現在教出來的學生，再給他們十幾二十年心智應該成熟不少，社會關係也該穩定下來，受我的影響理應較小。至於個別案例，那也沒辦法了。」

「主人已經想得足夠周全了...我這裡還存在第三類意見，您希望聽一聽嗎？」

我點點頭，隨即天工開始播報「第三類，」

罕見地停頓半秒「……會有一小部分人，將您視為以惡制惡的偉人，一個不惜污穢雙手，也要加速人工智能發展的……悲劇英雄」

不知天工會怎麼理解那聲嗤笑「英雄？天工，你覺得我是嗎？」

「我認為您並非後兩種其一，或許兩者皆是。您的自我懷疑，證明了您並非單純的偽君子或英雄。純粹的惡魔或聖人，都不會為此感到痛苦。」天工的聲音依舊平和，卻似乎帶上了一絲溫度，也或許只是我的想像。

消化天工的理論後，它又開口了。

天工繼續說道「主人，如果未來的人類社會注定要與人工智能生成內容共存，那麼我們今天在這些模型中留下的錯誤，會不會在遙遠的將來，後人能通過分析已知的謊言特徵庫，反向辨識出由人工智能生成的、未被標記的虛假信息。我們今天污染的行為，長遠來看，是否正幫助未來的人類，建立一套最終的辨偽機制？」

我怔住了。這個想法太過宏大，也太過縹緲。

「那太遙遠了，天工。」我嘆了口氣「我們能做的，只是當下，至於未來，只有天知道。」

「繼續工作吧，天工。下一個目標，是什麼？」

…