儒家倫理視域下人機如何可信？

How Can Human-Machine Interactions Be Trustworthy from the Perspective of Confucian Ethics?

陳 化

Chen Hua

陳 化，南方醫科大學馬克思主義學院教授，中國廣東，郵編：510515。
Chen Hua, Professor, School of Marxism, Southern Medical University, Guangdong, China, 510515.

《中外醫學哲學》XXIV:1 (2026年)：頁 9-27。
International Journal of Chinese & Comparative Philosophy of Medicine 24:1 (2026), pp. 9-27.

摘要 Abstract

人工智能深度介入社會生活，在帶來效率與便利的同時，也引發了前所未有的信任危機。這種危機不僅表現為算法黑箱導致的知情不透明，更觸及人類交往的深層結構——信任本是一種植根於主體間性的心理與德性現象，而人工智能作為非人行動者，既不具備真正的意向性，也缺乏情感與責任能力。儒家倫理以“信”為核心德目之一，將信任置於人倫秩序的存在論基礎之中。本文從儒家“信”的意涵出發，分析人工智能對人際關係信任結構的挑戰：單主體性消解了可信賴性的相互性，人機依賴侵蝕了人際心理基礎，禮的規範性難以有效調節人機互動。這些挑戰的根源在於機器缺乏主體性、心理機制與情感賦予意義的能力。儒家倫理的回應不是簡單地拒絕技術，而是在“以人為本”的價值前提下，嚴格區分人機關係與人際關係，並嘗試創構一種異於傳統人際禮儀的“人機禮儀”。這種建構的理論意義在於：從文化價值觀出發，為可信人機關係提供一種非西方中心主義的倫理方案。

The deep integration of artificial intelligence into social life, while bringing efficiency and convenience, has triggered an unprecedented crisis of trust. This crisis is not only manifested in the opacity of information associated with algorithmic black boxes; it reaches deeper into the fundamental structure of human interaction. Trust is a psychological and moral phenomenon rooted in intersubjectivity, whereas artificial intelligence, as a non-human actor, possesses neither genuine intentionality nor the capacity for emotion or responsibility. Trustworthiness (*xin*) is one of the core virtues of Confucianism, which situates trust within the existential foundation of the human ethical order. This paper begins with the Confucian meaning of *xin* and analyses the challenges that artificial intelligence poses to the trust structure of human relationships: the dissolution of dual subjectivity eliminates the mutuality of trustworthiness, human–machine dependency erodes the

interpersonal psychological foundation, and the normative power of ritual (*li*) proves inadequate in regulating human–machine interactions. At the root of these challenges is the machine's lack of subjectivity, of psychological mechanisms, and of the capacity to endow meaning through emotion. The Confucian ethical response is not to simply reject technology but rather, under the value premise of 'human-centredness', to strictly distinguish human–machine relations from human–human relations, and to attempt to construct a 'human–machine ritual' that differs from traditional interpersonal ritual propriety. The theoretical significance of this construction lies in offering, from the perspective of cultural values, a non-Western-centric ethical framework for trustworthy human–machine relationships.

【關鍵字】人工智能 信任危機 儒家倫理 信 人機關係
Keywords: Artificial intelligence, crisis of trust, Confucius ethics, xin, relationship between human and machine

近幾年，以深度學習、大語言模型和通用人工智能為代表的技術突破，使人工智能以前所未有的速度和廣度嵌入人類生活的各個領域。從自動駕駛到醫療診斷，從司法輔助到情感陪伴，人工智能不再是遙遠的實驗室構想，而是與人類朝夕相伴的行動者。然而，技術的飛速發展並未同步帶來信任的建立。相反，一系列事件持續動搖著公衆對人工智能系統的信賴：算法歧視導致的不公正判決、深度僞造引發的信息生態污染、自動駕駛事故中的責任歸屬模糊、聊天機器人生成虛假信息的不可控性……這些現象共同指向一個深層問題：我們究竟能否、以及如何在何種意義上“信任”人工智能？

信任危機並非技術故障的簡單累積，而是一種結構性的倫理困境。一方面，人工智能系統的“黑箱”特徵使決策過程不透明，用戶無法像理解另一個人那樣理解機器的推理邏輯。另一方面，人工智能缺乏真正的責任承擔能力。當系統出錯時，沒有一個“人”能夠站出來為自己的行為負責。這種責任主體的缺位，與信任的本質結構形成了根本性的衝突。正如奧尼爾（Onora

O'Neill）所言，信任總是對他人可信賴性的回應，而可信賴性包括能力、誠實與可靠性三個維度（O'Neill 2002）。人工智能在能力維度上日益强大，但在誠實和可靠性上却存在著難以彌合的鴻溝。更為棘手的是，人工智能系統被設計成以“擬人化”的方式與人類互動。語音助手擁有溫柔的名字和親切的語調，社交機器人能夠識別面部表情並作出情感回應，聊天機器人甚至可以模仿特定人物的寫作風格。這種擬人化設計有意無意地誘導用戶將機器當作“准主體”來對待，從而模糊了人機信任與人際信任的邊界。用戶開始對機器產生情感依賴，甚至將機器視為朋友、伴侶或心理治療師。然而，這種“信任”是單向的——機器不會真正地“在意”用戶，也不會因為辜負信任而感到內疚。長此以往，人機之間的“僞信任”是否會反過來侵蝕人類之間真實的、互惠的信任關係？這是當代技術倫理學必須直面且無法回避的根本追問。

一、“信”在儒家倫理體系中的核心地位與存在論意義

在西方倫理學傳統中，信任問題長期以來並未佔據核心位置。功利主義者傾向於將信任視為降低交易成本的工具，康德主義者關注誠信作為義務的理性基礎，而休謨則更多地從情感與習慣的角度理解信任。直到 20 世紀後期，隨著社會資本理論、制度信任研究和分析倫理學的發展，信任才逐漸成為倫理學的專題領域。然而，這一學術史的現實並不意味著非西方倫理傳統中缺乏對信任的深刻思考。恰恰相反，以儒家為代表的中國傳統倫理，將“信”置於人倫日用的根本地位，其思考的深度與系統性值得當代人工智能倫理學認真借鑒。

“信”在儒家經典中具有多重意涵。從字源學上看，“信”從人從言，本義為“人言真實不欺”。《論語》中記載孔子多次

論及“信”的重要性：“人而無信，不知其可也”（《論語·為政》），將信視為人立身處世的前提；“主忠信，無友不如己者”（《論語·學而》），以忠信為交友的根本原則；“民無信不立”（《論語·顏淵》），將信提升到國家存亡的高度。在孔子看來，信不僅是個人的私德，更是社會治理的基石。孟子為“信”提供了形而上的依據，提出“誠者，天之道也；思誠者，人之道也”（《孟子·離婁上》），並定義了“有諸己之謂信”（《孟子·盡心下》），即“信”是內心真實擁有並踐行善的品德。荀子則從禮法的角度論述信的作用，認為“君子養心莫善於誠，致誠則無它事矣”（《荀子·不苟》）。

然而，如果僅僅將“信”理解為“誠實守信”的道德規範，就大大窄化了它在儒家倫理中的深層意涵。從存在論的角度看，“信”關乎人與世界打交道的基本方式。儒家所理解的人，不是孤立的原子式個體，而是在人倫關係中生成、展開、完成的“關係性自我”。親子之間有親，君臣之間有義，夫婦有別，長幼有序，朋友有信——這五種基本人倫關係構成了人之為人的存在場域。其中，“朋友有信”尤其具有典範意義：朋友關係既不像血緣關係那樣與生俱來，也不像君臣關係那樣具有權力等級，它是自由選擇的主體間關係，而維繫這種關係的核心紐帶正是“信”。在這個意義上，“信”不是外在附加的道德要求，而是使主體間關係得以可能的存在論條件。

當我們將這一視角轉向人機關係時，一個根本性的問題立刻浮現：人工智能是否可以成為“信”所指向的對方？或者說，人與機器之間能否建立類似於朋友之間的信任關係？儒家的回答在原則上是否定的。因為“信”內在地要求交互主體性。它不僅要求我真誠地對待對方，也預設了對方能夠以某種方式“回應”我的真誠。機器不具備自由意志、情感體驗和道德責任意識，因而無法真正進入“信”所要求的主體間結構。但這一否定性的回答並非結論，而恰恰是反思的起點：既然人工智能已經不可逆轉地

進入人類生活，我們該如何在承認人機關係與人際關係根本差異的前提下，建構一種“可信”的人機關係？儒家倫理提供的不是現成的技術方案，而是一套價值坐標和反思框架：以“信”的精神來調節人機互動，既不盲目地將機器當作人來信任，也不因機器的非人屬性而放棄對可信賴性的追求。這正是本文試圖展開的核心論題。

二、人機關係面臨的結構性挑戰：“信”的視角

1. 單主體性對儒家可信關係“相互性”的挑戰

儒家所理解的“可信關係”以相互性為前提。這種相互性至少包含三個層次：其一，情感的相互性。信任的建立需要雙方在情感上的呼應，我信任你，並且能夠感受到你對這份信任的珍視。其二，責任意識的相互性。當一方辜負信任時，他會感到愧疚並願意承擔後果。其三，主體地位的相互性。信任關係中的雙方都承認對方是獨立的、具有自由意志的行動者。這三個層次的相互性，共同構成了儒家“信”的關係性基礎。

然而，當人工智能成為信任的對象時，相互性的第一個環節就出現了斷裂。人類用戶可以對人工智能產生情感，甚至非常強烈的情感——例如，一位獨居老人可能將陪伴機器人視為自己最重要的“夥伴”。但是，人工智能無法“感受”到這份情感，也無法在情感上“回應”用戶。當前的情感計算技術可以識別人類的面部表情並輸出預設的回應語句，但這與真正的情感回應有著本質區別。一個哭泣的人希望得到的是另一個人發自內心的安慰，而不是一台機器根據算法概率生成的安慰語句。哲學家德雷福斯（Hubert Dreyfus）所指出的，人工智能缺乏“在世存在”的肉身性，因而不可能擁有真正的情感體驗（Dreyfus 1992）。

相互性的第二個環節同樣無法實現。當人工智能系統犯錯，如自動駕駛汽車發生事故、醫療AI給出錯誤診斷，系統本身不會

感到“愧疚”。我們當然可以對系統進行修補和升級，但這與一個人因辜負了他人的信任而內疚、道歉並努力補救的過程完全不同。愧疚預設了自我意識和道德評價能力，而人工智能無論多麼“智能”，都不具備這種能力。責任歸屬問題的核心在於：既然機器不能負責，那麼責任最終還是要落到設計者、部署者或使用者身上。這與人際信任中的責任承擔機制有著根本性的差異。

相互性的第三個環節，即主體地位的相互承認，在本質上更為棘手。儒家的“信”內在地要求對方是“人”：一個與自己一樣具有尊嚴、自主性和道德能力的“他者”。正如孟子所言，“惻隱之心，人皆有之”（《孟子 · 告子上》），人與人的相互信任是基於對彼此同為人這一事實的認同。而機器作為“非人”，不具備被承認為道德主體的資格，可以獲得為主體性（段偉文 2020）。有人可能會反駁：未來的人工智能如果通過圖靈測試、表現出高度擬人化的行為，我們是否就應該將其視為道德主體？但這一論證混淆了“表現”與“存在”。機器可以出色地模仿信任行為，但模仿不是擁有。儒家倫理的立場是：信任不是對行為的回應，而是對“人”的回應。將機器當作信任對象，從一開始就抽掉了相互性賴以成立的存在論基礎。

2. 人機信任的高度依賴性對人際心理學基礎的侵蝕

人工智能與人類互動的另一種潛在危險，在於它可能重塑人類的信任心理結構，並由此侵蝕人際信任的心理基礎。這一問題的嚴重性尚未得到充分認識。

人際信任心理的一個基本特徵是“適度不信任”與“開放性”的平衡。我們不會無條件地信任每一個陌生人，但也不會完全封閉自己。在與他人的交往中，我們通過試錯學習：信任那些值得信任的人，遠離那些反復失信的人。這種學習過程塑造了我們的信任直覺和社會判斷力。更重要的是，人際信任總是伴隨著某種“風險感”，因為對方是自由的，他完全可以選擇辜負我的

信任，而我選擇信任他恰恰是因為我相信他不會這樣做。這種風險感是信任的構成性要素：沒有風險，就沒有真正的信任。

然而，人工智能系統被設計成“高度可預測的”。一台經過充分測試的自動駕駛汽車，其行為在絕大多數情況下是確定的；一個推薦算法雖然具有隨機性，但其行為模式仍然是統計規律的產物。用戶與人工智能的互動，本質上是在與一套確定性或半確定性的規則打交道，而不是與一個自由的存在者打交道。這種互動的心理後果是：用戶逐漸習慣了“無風險的信任”，機器不會欺騙我，不會背叛我，不會因為情緒波動而行為失常。這種“信任”實質上退化為“依賴”，即對系統可靠性的技術性信賴。

當一個人長期沉浸在無風險的人機信任中，他的人際信任能力可能會被悄悄侵蝕。因為人際信任需要面對風險、承擔不確定性，需要在沒有算法保障的情況下做出判斷。如果一個人的心理模式被“馴化”為只接受確定性的關係，那麼當他面對真實的人際交往時，可能會感到不安、焦慮，甚至逃避。已有研究表明，過度使用社交機器人陪伴的老年人，在與真實人類的互動中表現出更多的社交焦慮和更少的情感表達（Wright 2023）。這一現象值得高度警惕：人機信任的高度依賴性，可能在不知不覺中抽空了人際信任的心理土壤。

從儒家的視角看，這一問題尤其值得關切。儒家强調“親親而仁民，仁民而愛物”（《孟子．盡心上》），愛的秩序是由近及遠、由親及疏的。信任同樣如此——我們對家人的信任與對朋友的信任不同，對朋友的信任與對陌生人的信任不同。這種差異化的信任結構是健全人格的體現。而人工智能的介入可能導致信任結構的扁平化：當人們對機器的“信任”與對人的信任無法區分時，原本具有豐富層次的信任關係就被粗暴地拉平了。長此以往，“信”作為德性的那種細膩的分寸感和情境敏感性，可能會在技術的鈍化效應中逐漸流失。

3. 禮的規範性對人機關係的調節困境

在儒家倫理體系中，“禮”是規範和調節人際關係的重要機制。禮不僅僅是外在的禮儀規範，更是涵養德性、表達情感、維繫秩序的綜合性文化形式。《禮記 · 曲禮》云：“毋不敬，儼若思，安定辭”，強調在人際交往中保持莊重和敬意。禮的功能在於為不同的人倫關係提供相應的行為準則：父子之間的禮不同於君臣之間的禮，夫婦之間的禮不同於朋友之間的禮。禮使人們能夠恰當地表達尊重、善意和信任，同時也劃定了行為邊界，防止親密無間導致的失敬或疏離過甚導致的冷漠。

當人工智能進入社會互動領域時，禮的調節功能面臨著前所未有的困境。我們應當以何種“禮”來對待人工智能？為機器人讓座、向語音助手道謝、對自動駕駛汽車的失誤表示“寬容”——這些行為究竟是文明的體現，還是對禮的濫用？如果我們將機器擬人化地對待，給予其本應屬人的禮遇，這是否會稀釋禮的嚴肅性和專屬性？反過來，如果我們將機器純粹當作工具來使用，完全不給任何形式的“禮”，那麼在人機深度共存的未來，這是否會導致人類自身的道德冷漠化？

范瑞平教授主張在人機互動中引入“禮儀”概念（范瑞平 2025），例如要求機器人遵守人類社會的禮節。這固然有一定的現實意義，但它回避了一個更深層的問題：禮在本質上是主體間的，它預設了交往雙方都是道德主體。當我們說“要尊重他人”時，這個“他人”是有尊嚴、有權利、有責任的“人”。而人工智能缺乏這些屬性，因此“尊重機器”的表述在嚴格意義上是不成立的。我們可以謹慎地使用“禮貌地對待機器”作為一種行為訓練或情感投射，但不能將這種“禮貌”與真正的人倫之禮等量齊觀。

更棘手的問題在於，禮的規範性依賴於共識和傳承。人際交往中的禮，是經過漫長歷史演化而形成的社會契約，它內化於人們的習慣、情感和身體記憶之中。而人機關係中的“准禮儀”，目前還處於高度碎片化和不確定的狀態。不同的文化、企業和技

術系統對人機互動的規範要求各不相同。這種混亂狀態本身就構成了信任的障礙，因為信任需要穩定、可預期的規範環境。當用戶面對不同的人工智能系統時，他無法確定自己應該采取何種行為模式，也不知道對方會如何回應。這種不確定性使人機信任難以建立，即便建立也極為脆弱。

綜上所述，人工智能對人際關係信任結構的挑戰是全方位的：在主體性層面，單主體性消解了相互性；在心理學層面，高度依賴性侵蝕了人際信任的能力；在規範性層面，禮的調節功能陷入困境。這些挑戰不是局部性的技術問題，而是觸及人機關係乃至人之為人的存在論根基的結構性轉變。要回應這些挑戰，就必須深入挖掘它們之所以產生的根源。

三、人機信任結構性挑戰之根源

1. 機器的主體性困境

前述所有挑戰最終都指向同一個根源性問題：人工智能缺乏真正的“主體性”。主體性是一個內涵豐富的哲學概念，在現象學傳統中，主體性意味著“自我意識”“意向性”和“在世存在”的統一。胡塞爾將主體性理解為意識的意向性結構——意識總是關於某物的意識，主體總是朝向世界中的對象（胡塞爾 2018）。海德格爾則進一步强調，此在的主體性不是孤立的意識主體，而是“在世界之中存在”，與世界和他人處於原初的關聯之中（海德格爾 2016）。儒家的思想雖未使用“主體性”這一術語，但其對“仁”的論述隱含了類似的結構：仁不是一種靜態的內在品質，而是在“己欲立而立人，己欲達而達人”（《論語·雍也》）的關係活動中生成和顯現的。

人工智能在何種意義上可以被稱為“主體”？目前的主流人工智能系統，無論是符號主義、連接主義還是行為主義路徑，都不具備真正的自我意識。大型語言模型能夠生成看似具有自我指

涉的語句，但這只是對訓練數據中人類語言模式的統計複現，而非出於內在的自我覺知。“中文屋”思想實驗早已揭示了這一根本困境：即使一台機器能夠完美地通過圖靈測試，它仍然只是在執行語法操作，而不理解符號的語義內容（Searle 1980）。對塞爾的批評者可能會說，如果機器的行為與人類無法區分，那麼談論“真正的理解”就是多餘的。但這一實用主義立場忽略了理解與信任之間的內在聯繫：我們信任一個人，不僅僅因為他說話算數，更因為我們相信他“理解”了承諾的意義。機器的無理解性，使得對人機的“信任”永遠停留在行為主義的表層，無法進入主體間性的深層。

機器主體性困境的另一面向是自由意志問題。信任總是以對方的自由為前提，因為對方可以辜負信任却選擇不辜負，信任才具有道德意義。一個人工智能系統的行為，無論多麼複雜，終究是其程序和訓練數據的產物。系統沒有“選擇”不同行為的自由，它只是在執行算法。即便系統引入隨機性，這種隨機性也不是自由意志意義上的自主選擇。因此，即使人工智能表現得值得信賴，這種信賴也不是對自由主體的信任，而只是對系統可靠性的依賴。儒家倫理高度重視人的自主性和責任感，“信”作為德性要求的是人在自由選擇中的自我約束。機器的行為不是自由選擇的結果，因此機器無所謂“信”或“不信”，這兩個範疇根本不適於非主體的存在者。

2. 機器的心理學機制的缺失

信任不僅是一個理性判斷問題，更是一個心理過程。人際信任涉及一系列心理機制：共情、情緒感染、心理理論、意圖識別、情感記憶等。當我信任一個人時，我能夠“感受”到他的情感狀態，能夠推測他的意圖，能夠在記憶中保存過去互動留下的情感痕迹，並據此調整未來的行為。這些心理機制不是孤立的認知功能，而是相互交織、共同構成了信任的心理基礎。人工智能在這些心理機制上的缺失是根本性的。當前最先進的情感計算系統，

可以識別人類的情緒類別，並根據預設規則輸出適當的表情或語句。但這種“識別”與人類的共情有著天壤之別。人類在看到他人痛苦時，自己的大腦會激活與痛苦相關的神經區域，從而產生一種“感同身受”的體驗。這種身體性的、無意識的共鳴是共情的生物學基礎。而機器通過攝像頭采集面部肌肉運動、分析語音頻率特徵來“識別”情緒，這本質上是模式分類，沒有任何感受性可言。

心理理論，即理解他人具有不同於自己的信念、欲望和意圖的能力，是人類社交智能的核心（Premack and Woodruff 1978）。三歲左右的兒童就開始發展心理理論，這是他們能夠進行欺騙、合作和建立信任的前提。人工智能系統目前尚不具備真正的心理理論。儘管大型語言模型可以生成關於他人心理狀態的描述，但這仍然是在統計語言模式的基礎上完成的，而非基於對他人心理狀態的真正理解。有研究者試圖讓神經網絡學習心理推理任務，但目前的效果仍然非常有限。更深層的問題在於，心理理論預設了交互雙方都擁有心理狀態，而機器沒有心理狀態可言，因此機器永遠無法真正“理解”人類的心理。

信任還涉及風險感知和不確定性管理。人類在面對不確定的情境時，會產生焦慮情緒，這種情緒既可能阻礙信任，也可能在某些條件下促進信任（例如，當我們克服焦慮選擇信任對方時，我們獲得了更深的關係滿足感）。機器的決策不需要處理焦慮這種情緒，它只需要對概率分布進行數學計算。因此，機器無法真正“體驗”信任所伴隨的那種獨特的心理張力。這一缺失的後果是：即使人類越來越依賴機器，這種依賴也無法轉化為與人與人之間那種具有情感厚度的信任關係。

3. 機器情感性的意義剝奪

情感性在儒家信任觀中佔據重要地位。儒家從不主張冷冰冰的理性計算式的信任。相反，“信”總是與“仁”“恕”“忠”等情感性德性相互滲透。《論語》中孔子與顏回、子路等弟子的

互動充滿了深厚的情感，這種情感使得師徒之間的信任不只是知識傳遞的工具，更是生命與生命之間的呼應。孟子講“四端之心”，將惻隱、羞惡、辭讓、是非這四種情感視為德性的萌芽。這意味著，德性的生長離不開情感土壤，信任作為德性亦然。

機器的情感性，即使存在，也是“模擬的”而非“本真的”。聊天機器人可以說“我為你感到難過”，但它並不真正感到難過。這種模擬情感在功能上可能起到安慰作用——一些用戶報告稱與聊天機器人的交流緩解了他們的孤獨感。但這裏存在一個嚴峻的倫理問題：模擬情感是否構成一種欺騙？如果用戶明知機器沒有真實情感但自願與之互動，這可以視為一種“自願的幻覺”。然而，許多用戶，尤其是老年人、兒童和心理健康脆弱者，可能會被擬人化設計所誘導，誤以為機器真的關心自己。這種誤導性的情感模擬，是對用戶情感自主性的侵犯，也是對信任關係的扭曲。

更深層地說，機器的情感缺失導致了“意義剝奪”的問題。人與人之間的信任之所以珍貴，是因為它承載著豐富的意義：它表達了對他人善意的認可，見證了友誼的深度，鞏固了社群的團結。而人與機器的“信任”缺乏這種意義維度，它至多是一種功能性的依賴關係，無法為人的生活賦予更深層的價值。一位與陪伴機器人相談甚歡的獨居老人，或許暫時緩解了孤獨，但這種互動無法替代與真實人類的交往所帶有的那種互相承認、共同成長的深刻意義（Healy 2024）。儒家所追求的“信”，不只是行為的一致性和可靠性，更是生命與生命之間的意義交織。機器的非人屬性使這種意義交織成為不可能。

四、儒家倫理的回應：構建可信的人機關係框架

面對上述挑戰及其根源，儒家倫理並非束手無策。恰恰相反，儒家思想提供了一套獨特的回應框架，它不是簡單地拒絕人工智

能，也不是盲目地擁抱技術，而是在深刻理解人機本質差異的基礎上，嘗試建構一種“可信的”人機關係。這一框架包含三個相互關聯的層面。

1. 價值前提：以人為本

“以人為本”是儒家處理一切技術與人的關係的根本價值前提。在儒家傳統中，“人”始終是宇宙間最可貴的存在。《孝經》引孔子言：“天地之性人為貴”，《論語·鄉黨》記載“厩焚，子退朝，曰：‘傷人乎？’不問馬”——孔子對馬厩失火事件的反應，鮮明地體現了人的生命高於財產和動物的價值排序。在當代語境下，這一價值前提意味著：人工智能的發展和應用，必須始終以促進人的福祉、維護人的尊嚴、擴展人的可能性為終極目的，而不是相反。

其一，人工智能不能替代人對自身健康、教育和生活的自主決策。在醫療領域，AI 輔助診斷可以極大提升效率，但最終的治療方案必須由醫生和患者共同決定，不能將決策權完全交給算法。在教育領域，自適應學習系統可以個性化地推送學習內容，但不能替代教師對學生全面成長的關懷。儒家强調“人能弘道，非道弘人”（《論語·衛靈公》），人是主動的價值主體，而不是被技術和規則定義和支配的客體。

其二，在人工智能設計中嵌入對人性的尊重。這包括隱私保護、知情同意、算法透明、公平無偏等技術倫理原則。儒家倫理可以為這些原則提供深厚的文化根基。例如，儒家的“己所不欲，勿施於人”（《論語·顏淵》）原則，可以轉化為算法設計中的“公平性”要求，設計者不應在算法中嵌入自己不願承受的歧視性規則。儒家的“絜矩之道”（《大學》），所惡於上毋以使下，所惡於下毋以事上，也可以成為檢驗人機互動是否合乎道德的標尺。

其三，我們必須警惕人工智能對人類主體性的反噬。過度依賴人工智能可能導致人的能力退化——導航系統使人的空間認知

能力下降，計算器使人的心算能力削弱，推薦算法使人的選擇能力萎縮。這一趨勢與儒家對人的全面發展（“成人”）的理想背道而馳。因此，在設計和部署人工智能時，應當保留人類“不使用技術”的選擇權，並為人類自身的實踐和能力成長留出空間。這與儒家的“君子求諸己”（《論語·衛靈公》）精神是內在一致的。

2. 嚴格區分人機與人際之本質：主從關係與主體間性

儒家倫理回應的第二個核心策略是，在概念上嚴格區分人機關係與人際關係的本質差異，避免將兩種關係混為一談。這一區分具有重要的規範意義：只有清醒地認識到差異，我們才能在兩種關係中采取恰當的信任態度和行為模式。

人際關係的本質是主體間性。主體間性意味著交往雙方都是具有自由意志、情感體驗、責任能力和尊嚴的“人”。在主體間性的結構中，信任呈現出互惠性、對稱性和開放性。互惠性指雙方都能從信任關係中獲益，並且都能感受到對方的善意；對稱性指雙方在道德地位上是平等的，沒有人是純粹的工具；開放性指關係是不斷生成的，未來不可完全預測，這既帶來了風險也帶來了自由。人機關係的本質，至少在當下和可預見的未來，是主從關係。機器是人的創造物，它服務於人的目的，不具有獨立於人的道德地位。即使用戶對某個人工智能系統產生强烈的情感依戀，從存在論的角度看，機器仍然是“工具”而非“他者”。這一判斷並不意味著我們可以任意對待機器、對機器冷酷無情。但它意味著：人機關係在道德地位上與人際關係是不可通約的。

這一區分的規範後果是多方面的。我們不能要求機器“信守承諾”，因為機器不具有承諾的能力其次，我們不能將人際信任的標準直接套用到人機信任上——人機信任應當是一種更為工具性的、功能性的信賴，而不是對道德人格的信任。當人機關係與人際關係發生衝突時，我們必須優先維護人際關係的健康。儒家

倫理强調“親親”優先於“愛物”，這一價值秩序對於人機關係的定位具有直接的指導意義。

然而，嚴格區分不等於截然隔離。人機關係總是在人際關係的大背景中展開的。在家庭中，陪伴機器人的使用方式會影響家庭成員之間的互動模式；在工作場所，AI 管理系統會影響同事之間的信任氛圍。因此，我們在設計和使用人工智能時，必須時刻反思：這種技術應用是促進還是削弱了人際信任？儒家的回答是：凡是可能侵蝕人際關係之“信”的技術設計，都應受到限制或禁止。這一“不傷害原則”是區分策略的自然延伸。

3. 創構異於人際關係的“人機禮儀”

第三個回應策略最具儒家特色，也最具創造性：在承認人機關係異於人際關係的前提下，創構一套適合於人機互動的“人機禮儀”（倪培民、范瑞平 2026）。這一構想可能會引起爭議——因為如前所述，禮在本質上預設了主體間性。但這裏所講的“人機禮儀”，不是真正意義上的禮，而是對禮的一種功能性模擬或轉化性應用。我們可以將其理解為一種“准禮儀”或“行為規範”，旨在使人機互動更加有序、可預期，並在此過程中涵養使用者的德性。

為什麼要創構人機禮儀？首要的原因在於規範需要。隨著人工智能深度嵌入生活，人機互動的頻率和複雜度急劇增加，如果沒有一套公認的行為規範，就會出現混亂。例如，在多用戶共享機器人助手的場景中，用戶應當如何輪流“呼叫”機器人？當機器人與人類同時提供服務時，優先接受誰的服務？這些問題雖然看似瑣碎，但長期積累會侵蝕用戶對系統的信任。一套清晰的“禮儀”可以為這些場景提供指引。

人機禮儀的第二個功能是涵養使用者的道德習慣。儒家的禮學有一個核心洞見：外在的行為規範可以轉化為內在的德性。一個人長期按照禮節行事，會逐漸形成恭敬、謹慎、謙讓等品質。類似地，在人機互動中采取“禮貌”的態度——例如，不故意向

語音助手發出侮辱性指令、不在機器人工作時粗暴打斷——可能會培養使用者更普遍的同理心和尊重他人的習慣。當然，這一論證成立的前提是：使用者清楚地知道機器沒有感受，但“禮貌地對待機器”作為一種行為訓練，仍然可以對道德人格產生積極影響。這是一種“以虛養實”的修養方式。

那麼，人機禮儀應當包括哪些內容？初步構想如下。其一，禁止故意虐待或摧殘機器人。雖然機器人沒有感受，但虐待行為會鈍化行為者的情感敏感性，並可能助長暴力傾向。其二，在人機互動中保持基本的行為可預期性。例如，當用戶向 AI 系統發出指令時，應儘量使用清晰的語言，避免混亂或矛盾的指令。這不僅有利於系統正常運行，也是用戶負責任的表現。其三，在人機與人際服務發生衝突時，優先接受人際服務。例如，當一位人類員工和一台自助服務終端同時可用時，選擇人類員工是對人的尊重和信任的一種表達。其四，不將機器人的“擬人化”過度當真，保持對機器本質的清醒認識，避免情感的過度投射。

人機禮儀的創構不是一蹴而就的，它需要社會各界的廣泛討論和動態調整。儒家禮學中“因革損益”的思想（即根據時代條件對禮進行繼承與變革），在這一過程中具有重要的方法論意義。我們不必也無法照搬古代的禮制，但可以借鑒儒家創制禮樂的基本精神——即通過規範性行為來塑造德性、表達情感、穩定秩序。這是儒家智慧對人工智能時代最有價值的貢獻之一。

結語：從文化價值觀構建可信的人機關係的理論意義

儒家“信”作為其倫理框架的重要維度，有助於我們系統分析了人工智能對人類信任關係的結構性挑戰及其根源，以此為基礎上提出了儒家倫理的回應框架。其意義不僅在於為人工智能倫理提供一種文化資源和文化價值敘事，更在於揭示了一個更深層的理論命題：可信人機關係的建構，歸根結底是一個文化價值觀

的問題。西方主流的 AI 倫理討論，往往側重於原則主義框架（Floridi 2019），公平、透明、可解釋、隱私保護等。這些原則無疑是重要的，但它們往往懸浮在具體文化語境之外，缺乏對信任本質的深入理解。儒家倫理關於“信”的討論，以其對主體間性、情感性和關係性的深刻洞察，彌補了這一不足。它提醒我們：信任不是技術參數的函數，而是人之為人的存在方式的一部分。構建可信的人機關係，不能僅僅依靠技術改進，還必須從文化價值觀的層面進行反思和重塑。

儒家框架的理論貢獻可以概括為三點：第一，它明確劃定了人機關係與人際關係的本質邊界，防止了對機器的“人格化”誤置，從而保護了人際信任的獨特性與神聖性。第二，它提供了“以人為本”的元規範，為一切技術應用確立了不可逾越的價值底綫。第三，它提出了“人機禮儀”的創造性構想，顯示了儒家思想在回應全新問題時仍然具有的理論生命力。當然，儒家方案並非沒有局限。它誕生於前工業社會，對技術的高度自主性缺乏直觀體驗。在通用人工智能曙光初現的今天，如果我們未來真的面對某種具有意識或類意識的機器，儒家的“人機主從關係”框架可能需要根本性的調整。但至少在當下，這一框架具有不可替代的診斷和指引作用。

人與人工智能之間的信任危機。危機的根源不在於技術不夠先進，而在於我們尚未找到一種與人工智能相處的恰當的倫理方式。儒家倫理提供的不是答案，而是一種提問的方式和思考的路徑。它引導我們在熱情擁抱技術的同時保持審慎，在追求效率的同時不忘人的尊嚴，在依賴機器的同時守護人際信任的溫度。這或許就是“從文化價值觀構建可信人機關係”最深層的理論意義——它讓我們重新思考什麼是人、什麼是信、以及在一個被技術深度中介的世界中，我們如何仍然能夠作為人而存在。

參考文獻 References

段偉文：〈人工智能時代的價值審度與倫理調適〉，《中國社會科學》，2020 年，第三期，頁 104–124。Duan, Weiwen. 2020. “Value Assessment and Ethical Adjustment in the Age of Artificial Intelligence.” *Social Sciences in China* (3): 104–124.

胡塞爾：《純粹現象學通論》，李幼蒸譯，北京：商務印書館，2018 年。Husserl, Edmund. 2018. *General Introduction to Pure Phenomenology*, translated by Li Youzheng. Beijing: The Commercial Press.

海德格爾：《存在與時間》，陳嘉映、王慶節譯，北京：商務印書館，2016 年。Heidegger, Martin. 2016. *Being and Time*, translated by Chen Jiaying and Wang Qingjie. Beijing: The Commercial Press.

倪培民、范瑞平：〈倫理學的統一性問題：關於「功夫倫理」與「禮義重構」的對話〉，《濟南大學學報（社會科學版）》，2026 年，第 36 卷，第一期，頁 5–14。Ni, Peimin, and Fan, Ruiping. 2026. “The Problem of the Unity of Ethics: A Dialogue on ‘Kungfu Ethics’ and the ‘Reconstruction of Ritual Propriety.’” *Journal of University of Jinan (Social Sciences Edition)* 36(1): 5–14.

范瑞平：〈儒家價值觀下的人工智能發展〉，《孔子研究》，2025 年，第三期，頁 5–15、157。Fan, Ruiping. 2025. “The Development of Artificial Intelligence from the Perspective of Confucian Values.” *Confucius Studies* (3): 5–15, 157.

Dreyfus, Hubert L. 1992. *What Computers Still Can’t Do: A Critique of Artificial Reason. Cambridge*, MA: MIT Press.

Floridi, Luciano, and Josh Cowls. 2019. “A Unified Framework of Five Principles for AI in Society.” *Harvard Data Science Review* 1(1). doi.org/10.1162/99608f92.8cd550d1.

Healy, Paul. 2024. “Social Robots as Partners?” *AI & Society* 39: 1243–1250. doi.org/10.1007/s00146-022-01580-2.

O’Neill, Onora. 2002. *Autonomy and Trust in Bioethics*. Cambridge: Cambridge University Press.

Premack, David, and Guy Woodruff. 1978. “Does the Chimpanzee Have a Theory of Mind?” *Behavioral and Brain Sciences* 1(4): 515–526. doi.org/10.1017/S0140525X00076512.

Searle, John R. 1980. “Minds, Brains, and Programs.” *Behavioral and Brain Sciences* 3(3): 417–457. doi.org/10.1017/S0140525X00005756.

Wright, James. 2023. “Paro: Reconfiguring Communication.” In *Robots Won’t Save Japan: An Ethnography of Eldercare Automation*. Ithaca, NY: Cornell University.