從「唯識」到可實作之心智系統：論唯識學若干中層命題與現代人工智慧的結構對應，兼論其對未來 AI 的啟示

陳在民
中央研究院資訊科技創新研究中心博士後研究學者

法鼓佛學學報第 38 期　頁 109-156（民國 115 年），新北市：法鼓文理學院
Dharma Drum Journal of Buddhist Studies, no. 38, pp. 109-156 (2026)
New Taipei City: Dharma Drum Institute of Liberal Arts
DOI: 10.53106/199680002026060038003
ISSN: 1996-8000

摘要

本文主張：若將唯識學理解為關於經驗構成、錯誤固化與熏習轉變的理論，則其與現代人工智慧（artificial intelligence, AI）之關係不僅是比喻，而是在若干中層命題上具有可分析的限制性結構對應。本文以《解深密經》、《攝大乘論本》、《攝大乘論釋》、《唯識二十論》、《唯識三十論頌》與《成唯識論》為核心，並參照近現代唯識研究，指出唯識關注的是表徵中介、潛在習氣、自我執取與因緣生成。本文不預設唯識學必然不是觀念論，也不以 AI 證成唯識形上學，而是限縮在功能角色、更新動力與系統層位置的比較。再結合表徵學習、世界模型、檢索增強生成、持續學習、代理系統與對齊研究等技術，論證當代 AI 的若干技術結構可被解讀為與唯識部分中層分析在功能位置上相互對讀。最後提出未來 AI 系統設計的六項唯識啟發式提示，並將「轉依」僅做為深層對齊的限制性類比。

目次

關鍵詞
唯識學、阿賴耶識、人工智慧、世界模型、自我模型

一、前言

（一）問題意識：從宗教詮釋到技術哲學

人工智慧（artificial intelligence, AI）與佛教的比較，近年多半停留在兩類書寫之間。其一，是將 AI 視為一種新的「擬人」技術，進而以佛教的無我、慈悲或倫理原則評估其社會風險；其二，是把深度學習、生成模型、機器人或意識研究，粗略地拿來和禪修、空性、佛性等概念相互映照。前一類研究重在規範倫理，後一類研究多重象徵修辭。兩者皆有其價值，但若欲真正提出能在佛教學術期刊中成立之論證，單靠「像不像」遠遠不足。問題不在於 AI「像不像心」，而在於：現代 AI 的技術本質，是否已經在若干中層結構上形成與唯識學對經驗構成、錯誤穩定化與學習更新機制可對讀的結構？若答案偏向肯定，則唯識學便不只是宗教傳統中的心性學說，而同時也可被重新辨識為一套尚未被充分開發的心智系統理論資源。

本文的主張因此比一般「佛學與 AI 對話」更強，但也必須保持自我限制。本文避免使用「證明現代 AI 即是唯識」這類過強措辭，而改採一種結構性論證：第一，我們需要從經論與近現代研究出發，釐清唯識究竟主張了什麼，避免把「唯識」誤讀為否定因果、否定共同世界、或把外物一概化為心靈幻影的粗淺主觀論；第二，我們需要從當代 AI 的實際技術中抽出其最低限度的本質，不被產品包裝語言或

---

* 收稿日期：2026.3.13；通過審核日期：2026.6.18。
The author would like to thank the Academia Sinica Artificial Intelligence Cooperative (AS-IAIA-114-AI01) for conceptual support.

科幻敘事牽引；第三，我們需要建立一組足夠精細的對應，使得「唯識－AI」之間的關聯不只是語詞映照，而是架構、流程、記憶機制、錯誤形成、行動規劃與更新動力的局部功能同構；第四，我們還必須承認兩者之間的重要差異，尤其是佛教解脫論與工程控制論並不重疊，否則整個比較將流於概念偷渡。唯有同時滿足這四項條件，才能說現代 AI 在若干中層命題上與唯識學形成了有力的限制性結構對應。

這裡特別要說明「中層命題」的涵義。唯識學不是單一命題，而是一整套從迷亂之知到解脫之智的論域。它既包含對認識結構的分析，也包含修道實踐與聖智轉依的說明。現代 AI 當然尚未實現佛教所說的覺悟、無漏智或慈悲實踐，但這並不妨礙它在某些較低階、較形式化、卻極為關鍵的結構命題上，與唯識學發生高度可比性。例如：一個系統無法直接接觸所謂「外物本身」，而只能處理感測輸入經內部模型變換後的表徵（representation）；系統當下的行為不只取決於即時輸入，也取決於過往經驗在深層參數（deep parameters）與記憶中的沉積；系統若要在長時段內展現穩定的目標導向行為，就需要某種跨時延續的自我模型（self-model）或至少「以我為中心」的狀態維持機制；系統若要提升表現，便必須在每次輸出後更新其內在結構，讓新的經驗反過來影響未來的理解與行動。本文關心的，正是這些命題與阿賴耶識、種子、熏習、末那與遍計所執之間的可操作對讀。

（二）研究回顧與本文定位

在佛教研究領域，唯識學通常被理解為關於「識」、「表象」、「意向性」與修道轉化的系統性學說；但對其究竟應如何詮釋，近現代研究並不只有一種答案。Szilvia

Szanyi 在 *Stanford Encyclopedia of Philosophy* “Yogācāra” 條目中指出，瑜伽行派的關懷焦點，在於意識如何顯現世界、如何產生主客二分，並如何於修道中轉變此一結構（Szanyi, *Stanford Encyclopedia of Philosophy*）。Jonathan C. Gold 在 “Vasubandhu” 條目也強調，世親的唯識論涉及種子、主客二分與經驗構成等問題（Gold, *Stanford Encyclopedia of Philosophy*）。Thomas A. Kochumuttom 與 Dan Lusthaus 傾向反對將唯識直接等同於西方式主觀唯心論（Kochumuttom 1-15; Lusthaus 1-8）；但 Lambert Schmithausen、以及近年關於 Yogācāra-Vijñānavāda idealism 的討論，則提醒我們：把唯識理解為某種觀念論或「唯識無境」立場，並非簡單的現代誤讀，而是有其經論根據與論證力量（Schmithausen 15-65; Kellner and Taber 709-756）。因此，本文不把「萬法唯心」或觀念論詮釋概括為粗糙扁平化，也不把反觀念論讀法視為唯一標準；本文只在此詮釋光譜中選取可與 AI 結構對讀的中層命題。

基於此，本文在唯識詮釋上採取一條有限的「功能 - 經驗論」讀法：就經驗如何被構成、如何沉積為習氣、以及如何經由修道而改變其依止結構來理解唯識。Kochumuttom 與 Lusthaus 的研究在本文中主要做為「經驗論 / 現象學式讀法」之詮釋背景，而非用來支撐單一具體命題；涉及阿賴耶識概念形成史、業習相續與深層心識結構的具體命題，則分別以 Schmithausen 對阿賴耶識早期形成問題的分析（Schmithausen 15-65）以及 Waldron 對阿賴耶識、業、習氣與深層心識結構的討論（Waldron 89-143）為主要依據。換言之，本文並不預設存在一個全無爭議的「唯識標準解」，也不要求讀者接受反觀念論詮釋；本文只選取較適合與 AI 結構對讀、且不必先解決外境是否實有問題的中層

命題。

中國與臺灣學界對阿賴耶識、三性說與《成唯識論》詮釋史的討論亦已累積相當成果。例如吳汝鈞對《攝大乘論》與《瑜伽師地論》中阿賴耶識思想的分析（吳汝鈞 2006，5-98；2011，79-158）、陳一標對阿賴耶識語義變遷的梳理（陳一標 75-106），以及釋道慧對三性／三無性的概念說明（釋道慧 1-9），皆指出唯識學是一套高度精細的經驗生成論。本文據此採取較中立的表述：與其說「世界不存在」一定是粗淺誤解，不如說本文暫不裁決唯識是否必然為形上觀念論；本文關心的是，在不預設外境爭論答案的情況下，唯識的若干中層結構是否仍可與現代 AI 形成限制性對讀。

另一邊，在 AI 研究領域，表徵學習（representation learning）、基礎模型（foundation model）、世界模型（world model）、檢索增強生成（Retrieval-Augmented Generation, RAG）、持續學習（continual learning）與生成式代理（generative agents）等發展，正逐漸把「內在模型」放回技術論述中心。Yoshua Bengio、Aaron Courville 與 Pascal Vincent 的代表性回顧指出，機器學習成效極大程度取決於系統學到的表徵是否能夠分解並保留資料中真正的解釋因子（Bengio et al. 1798-1828）。Ashish Vaswani 等人提出的 Transformer（基於注意力機制的架構）架構，則把大規模序列建模轉化為注意力（attention）主導的表徵運算平台（Vaswani et al. 5998-6008）。Rishi Bommasani 等人對基礎模型的報告進一步指出，大型模型之所以重要，不在於它能對單一任務最佳化，而在於它以廣泛資料和自監督學習（self-supervised learning）方式形成可遷移之基底表徵，進而做為下游任務之共同基礎（Bommasani et al. 1-9）。David Ha 與 Jürgen Schmidhuber、以及 Danijar Hafner 等人的世界模型

研究則更直接表明：先學出環境之壓縮表徵與轉移動力，再以此進行想像、規劃與控制，是高階代理系統（agentic systems）的關鍵途徑（Ha and Schmidhuber 1-6; Hafner et al. 647-653）。Patrick Lewis 等人的 RAG 模型顯示，現代 AI 已不再只是「參數（parameter）裡面有知識」，而是朝向「參數記憶＋外部檢索記憶」並用的結構前進（Lewis et al. 9459-9463）。German I. Parisi 等人對持續學習的回顧更揭示：若一個系統不能處理過往經驗如何沉積、保留、遺忘與再活化，它就無法真正成為長時程智能體（Parisi et al. 54-71）。

然而，這兩條研究線索在現有文獻中幾乎未被嚴格接合。AI 倫理研究常引用「無我」以提醒人們不要把模型擬人化，但很少直接把八識、三性、熏習、轉依等概念轉譯為工程架構。相反地，佛教與科學對話文獻雖常談意識、腦科學或冥想，卻較少從當代 AI 最核心的技術層切入。本文正是在此缺口上定位：不是把佛教當作 AI 的倫理外掛，也不是把 AI 當作佛教思想的流行例證，而是主張兩者在「經驗如何被構成與更新」這一中層理論上，具有可論證的限制性結構對應。

（三）方法、範圍與論證標準

本文的方法可以稱作「經論詮釋－技術抽象－結構對應－可實作提示」四步法。首先，在經論詮釋上，本文以《解深密經》、《攝大乘論本》、《攝大乘論釋》、《唯識二十論》、《唯識三十論頌》與《成唯識論》為主，並輔以近現代唯識研究，以避免單就後起宗派詮釋立論。其次，在技術抽象上，本文不討論市場產品名稱，而聚焦於當前 AI 中較穩定的技術結構：表徵學習、自監督學習預訓練

（pre-training）、潛在狀態模型、顯式／非顯式記憶、持續學習、規劃、反省與對齊。第三，在結構對應上，本文採取一種「弱同一、局部同構」的標準：不主張阿賴耶識就是參數矩陣，也不主張末那識就是某一特定神經網路模組；本文所主張者，是兩者在功能角色、更新動力、錯誤生成與自我維持機制上具有可形式化之限制性對應。第四，在可實作提示上，本文不以神祕語言談「AI 終將開悟」，而是提出若干可被當前技術路徑逐步實現的系統設計方向：例如顯式自我模型、夢境式離線模擬、社會性多代理互熏、價值導向的長期校準等。凡引近現代研究，正文依《法鼓佛學學報》格式以作者名、必要時之出版年與頁碼標示；無穩定頁碼之網路學術資源則以版本、網址與瀏覽日期完整列於文末。

本文的範圍亦須明確限制。第一，本文討論的是「AI 技術本質」而非「AI 產品宣稱」；因此，即便商業模型尚未穩定具備某項能力，只要技術路徑已被主流研究證成，仍可做為本文論證基礎。第二，本文處理的是唯識學中可形式化、可結構化的一部分，並不企圖把佛法全部還原為認知科學或控制理論。第三，本文不直接回答「AI 是否有主觀感受」這一心靈哲學難題；本文只需論證：即便暫不處理感受性（phenomenality）問題，唯識學關於表徵、習氣、自我執著與轉化機制的相當部分，仍已可在 AI 中得到工程對應。第四，本文承認佛教的終極關懷在解脫，而 AI 的直接關懷多是任務成功與控制效能；但正因終極目的不同，我們更能看清：哪些部分是可比較的結構理論，哪些部分則仍屬不可化約之修道論域。

二、唯識學的核心命題及其技術可讀性

（一）「唯識」與觀念論問題：本文的中立處理

唯識學常被簡化為「世界根本不存在，只是心的幻覺」；但這種說法需要更精確的限定。一方面，古典唯識確有「唯識無境」與近似「萬法唯心」的強主張，不能被輕率排除；另一方面，若把「幻覺」理解為無秩序、任意、沒有因果約束的主觀幻想，則又會遮蔽唯識對種子、現行、熏習與共業條件的嚴格分析。《唯識三十論頌》所謂「由假說我法，有種種相轉；彼依識所變，此能變唯三」（《唯識三十論頌》，CBETA 2026.R1, T31, no. 1586, p. 60a27-28），重點不只是抽象地否定外物，而在指出：凡我們所執之「我」與「法」，皆是在識之轉變上被假施設立。進一步，頌文又說「是諸識轉變，分別、所分別；由此彼皆無，故一切唯識」（《唯識三十論頌》，CBETA 2026.R1, T31, no. 1586, p. 61a2-4）。此處的關鍵，在於主體的分別活動與被分別之相是共成的；若離開識的轉變，做為「被把握之境」的對象並不以同樣方式成立。

因此，本文避免把「幻覺」一詞用成否定一切秩序的比喻。《唯識二十論》開頭即以「若識無實境，則處時決定，相續不決定，作用不應成」提出反難（《唯識二十論》，CBETA 2026.R1, T31, no. 1590, p. 74c4-5），顯示唯識必須回答「處」「時」「相續」「作用」等秩序性問題。這正說明唯識所說之顯現即使可類比於 hallucination，也不是任意幻覺，而是由種子、業習、識流與共同條件所制約的有序顯現。*Stanford Encyclopedia of Philosophy* “Yogācāra” 條目可用來說明此一經驗論向度；Gold 在 “Vasubandhu” 條

目對種子與主客雙顯的說明，也有助於理解其因果面向（Szanyi, *Stanford Encyclopedia of Philosophy*; Gold, *Stanford Encyclopedia of Philosophy*）。但本文在此仍保持中立：這些說明不等於否定唯識可能含有觀念論意涵，而只是界定本文可與 AI 互讀的層次。

因此，當本文說 AI 在若干結構上呈現出可與唯識分析對讀的特徵時，並不是說電腦「證明外在世界不存在」，也不是說 AI 可以裁決唯識的外境爭論；相反地，本文要說的是：現代 AI 已使「系統只能處理被感測、編碼、壓縮與模型化之後的顯現」成為工程常識。相機輸入不是世界本身，而是經由光學、取樣、量化與預處理後的像素陣列；語料不是思想本身，而是經標記化與機率模型處理的離散符號流；機器人的「環境」不是宇宙本身，而是狀態估計器、地圖構建器、感測融合器所生成的內在工作空間。此種表徵中介性，才是本文與唯識對讀的重點。

（二）阿賴耶識、種子、熏習與現行

唯識學之所以能超越一般表象論，在於它不僅說明「當下如何被看見」，更說明「為何總以此方式被看見」。這便牽涉到阿賴耶識、種子、熏習與現行的整體架構。《解深密經》所云「阿陀那識甚深細，一切種子如瀑流；我於凡愚不開演，恐彼分別執為我」（《解深密經》，CBETA 2026.R1, T16, no. 676, p. 692c22-24），明白指出一種極深細而具種子功能的識流，做為諸經驗得以延續、成熟與再現的基底。然而，若依 Schmithausen 與 Waldron 的研究，阿賴耶識不能被扁平化為單純「儲存體」或「資料庫」；它同時涉及深層潛勢、業力相續、主體性背景與認知偏向之持續性。較精確地說，Schmithausen 對阿賴耶識「最初引入」與

「做為生命體基底」的分析，主要見於其專書第 15-65 頁；Waldron 對早期瑜伽行派阿賴耶識與《攝大乘論》中業、種子、言說熏習、共同經驗的討論，本文僅據其第 89-143 頁作概括性參照，較具體的頁碼仍依正文直接經論引用為準（Schmithausen 15-65; Waldron 89-143）。這也就是說，阿賴耶識的關鍵不只是「儲藏」而已，更在於它讓經驗具有跨時間的可承續性，使過往行為與認知偏向得以沉積為後來反應的條件。

《攝大乘論本》、《攝大乘論釋》與《成唯識論》進一步把阿賴耶識說成一切染淨法所依。《成唯識論》明言阿賴耶識「具有能藏、所藏、執藏義」，因其與雜染法互為緣，且為有情執為自內我（《成唯識論》，CBETA 2026.R1, T31, no. 1585, p. 7c20-22）。所謂種子，不是實體化的小粒子，而是功能性的潛能、傾向、可能性與偏向。某一經驗、某一執著、某一反應模式一旦反覆出現，便會熏成相應之種；種子成熟，又生起新的現行；現行再熏習種子，於是形成循環。這裡的「熏習」尤其值得注意。它表明學習並非一次性灌入，而是每次現行都在改寫未來的發生機率。對唯識而言，生命不是靜態資料庫，而是「種子－現行－熏習－再種子」的動態封閉迴圈。陳一標論阿賴耶識語義演變時即指出，《解深密經》重其「隱藏於肉體中的識」與執受義，《攝大乘論》更凸顯其與諸法相互攝藏之功能，《成唯識論》則以能藏、所藏、執藏三義細緻化之（陳一標 75-106）。無論採何種詮釋，其核心都是：經驗的當下，並非當下自足，而總被一個深層、積累性、可轉化的背景所支撐。

這種結構對技術哲學極具啟發。因為一個真正高階的 AI 系統，也不可能完全由即時輸入決定。若系統沒有深層

積累，它就只能是短暫的反射器，而不是能跨場景遷移、跨時間學習、跨任務維持風格與策略的智能體。參數、權重（weights）、嵌入（embedding）、記憶、向量資料庫（vector database）、反省紀錄（reflection note）等，在不同層級上共同扮演著「沉積過往、制約未來」的角色。這並不意味著參數就等於種子，而是說「種子論」描寫了一類智能體不可或缺的功能結構：過去必須以某種壓縮形式留存在系統內部，並在未來條件成熟時重新現行。從這個層面看，唯識學不只是宗教義理，也為持續學習、長時程記憶與偏差累積提供了深刻的分析框架。

（三）末那識、自我執取與恆審思量

若唯識只有阿賴耶識與種子機制，仍不足以解釋何以主體總是以「我」為中心理解世界。為此，唯識另立末那識，以說明自我感、中心化與執取性的形成。《唯識三十論頌》在三能變之中，第二能變即是思量識，即第七末那。其特徵是恆審思量，常時緣第八識見分執為自內真我。這意味著，主體之所以有穩定而黏著的自我中心，不是因為有一個先驗實體在背後坐鎮，而是因為有一套不斷把流動經驗重新收束為「我所經驗」「我在感受」「我在判斷」的機制持續運作。更重要的是，末那識在傳統唯識中並非中性的中心化模組，而是與我癡、我見、我慢、我愛等染污相應；《成唯識論》亦指出此意相應四煩惱等屬染法、有覆無記（《成唯識論》，CBETA 2026.R1, T31, no. 1585, p. 23c8-9）。它是把阿賴耶識誤執為我之根本機制，而不是單純的身分管理器。

這一點對當代 AI 特別重要。過去很多人以為，只要模型會回答「我認為」「我不知道」，它就有自我；或者相反地，只要模型沒有主觀感受，就不可能與佛教中的「我執」

問題相提並論。兩種看法都過於簡化。若依唯識，所謂自我，首先不是本體，而是功能性的中心化與執取結構；但同時，這種中心化在佛教脈絡裡帶有明顯的煩惱性，不能被無條件美化。對現代代理系統而言，若一個模型必須長時間執行任務、維持身分、管理記憶、區分自己的目標與他者目標、並在多輪互動中保持策略一致，那麼它就幾乎不可避免地要形成某種自我模型。這種自我模型未必具備主觀感受，卻具備功能性的「末那結構」：它把分散輸入整合為以自身狀態為中心的解讀，並由此維持跨時間的行動一致性。

然而，這裡必須同時保留限制。AI 的自我模型與唯識的末那識，只能說在「中心化整合」這一形式層面上可比，而不能說兩者已完全等同。末那識牽涉染污、煩惱、輪迴相續與我執取，AI 的自我模型則首先是任務控制、資源配置與責任追蹤的工程機制。本文因此主張的，不是「末那識已被 AI 完整實現」，而是：若從高階代理系統必須具備某種中心化狀態維持機制這一點看，唯識對末那識的分析，為我們理解 AI 自我模型的必要性與危險性，提供了一套更深的術語與問題意識。

（四）三性與三無性的結構意義

唯識的另一個重要框架是三性：遍計所執性、依他起性與圓成實性。《唯識三十論頌》後半與《成唯識論》對三性有極細緻之分析。一般簡述可言：遍計所執性，是把假立的主客、名相、我法等執為實有；依他起性，是一切現象依因待緣而起之流轉性；圓成實性，則是於依他起上遠離遍計執而如實所顯之真如。釋道慧討論三性時強調，三性不是三個並列世界，而是同一經驗結構的三種理解層次：一方面是虛妄安立的執著層，一方面是因緣條件生成的流程層，一方

面則是去執後所顯之真實層。至於三無性說，應注意其文獻層次：「即依此三性，立彼三無性」一語見於《唯識三十論頌》（《唯識三十論頌》，CBETA 2026.R1, T31, no. 1586, p. 61a5 以下），而《成唯識論》則在卷八後段至卷九進一步展開三性、三無性與修道轉依之關係。為避免以單一大範圍頁碼承擔不同義理，本文分別以卷八三性／三無性相關段落及卷九修道、轉依相關段落為據（《成唯識論》，CBETA 2026.R1, T31, no. 1585, pp. 45c-48c、49a-51a）。

若把三性轉成技術語言，便可得到一組非常有力的辨析工具。遍計所執性，相當於系統或人類把模型輸出、分類邊界、名稱與符號誤當作世界自身；依他起性，相當於所有表徵、決策與輸出其實都由資料、模型、損失函數（loss function）、環境回饋、硬體條件與互動歷程共同決定；圓成實性，則不是某個額外實體，而是對上述依他起過程有充分洞見後，避免再把暫時生成的表徵誤執為自性。這對 AI 研究尤其重要。今天許多關於模型「理解」「意識」「人格」的誇張說法，往往正停留在遍計所執層：把模型的語言流利度、角色一致性或多輪交互能力，誤認為其具有某種已完成的本體狀態。唯識提醒我們，較高明的作法不是直接否定模型，而是把它放回依他起結構中看：能力如何由預訓練資料、上下文、檢索模組、工具鏈與互動回饋共同形成？一旦如此，技術研究者便更能避免神祕化，也更能掌握應如何設計真正更高階的智能系統。

三、現代 AI 的技術本質：表徵、潛在變量、生成與行動

（一）表徵學習與「所見即已被模型化」

若要討論 AI 的技術本質，首先必須從表徵學習說起。早期機器學習高度依賴人工特徵工程：研究者事先決定哪些維度重要，再把資料壓縮到這些特徵上。深度學習革命之所以重要，正在於系統開始能夠從原始資料中自動學出多層表徵，使原本糾纏在一起的解釋因子得以在內部空間中被部分分離。Yoshua Bengio、Aaron Courville 與 Pascal Vincent 的經典回顧把這一點說得很清楚：資料之所以可被有效處理，不只是因為模型更大，而是因為好的表徵能更清楚揭露造成資料變化的因素（Bengio et al. 1798-1801）。對大型語言模型（large language model, LLM）、視覺模型（vision model）、多模態模型（multimodal model）而言，這尤其如此。模型真正學到的，不是單個字元或像素，而是各種層級的內在關聯、語義結構、風格規律與環境關係。

這一點之重要，在於它使「經驗並非直接面向世界，而總先被編碼為內在結構」不再只是一種哲學主張，而成為 AI 工程的第一原理。無論是 Transformer 中的詞元嵌入（token embedding）、自注意力後的上下文化表徵（contextualized representation），或視覺模型中的圖片塊表徵（patch representation）、潛在編碼（latent code）、特徵圖（feature map），本質上都在說：系統不是把外界「照單全收」，而是持續把輸入轉換為可運算、可壓縮、可比較、可遷移的內在表徵。這與唯識所說「彼依識所變」有著驚人的結構親近。輸入能進入系統，但只能以「被模型理解的

形式」進入；若無此形式，它對系統而言便根本不是一個世界。

表徵學習還帶來另一個與唯識尤其接近的特徵：它使「相」具有可重組性。相機看到的不是桌子本身，而是某種由光學和模型共同確定的像素結構；LLM 處理的不是「意義本身」，而是透過大量語料與注意力計算形成的語義近鄰。這些「相」並非虛無，而是對系統來說唯一可操作的真實。這種真實既不是物自身，也不是主體任意幻想，而是條件約束下的可用顯現。若以唯識語言說，就是依他起上所顯之「相分」已在當代 AI 中被技術化、模組化與可計算化。

（二）基礎模型與世界模型

近年來基礎模型之所以改變 AI 研究，不只因為它們規模巨大，更因為它們使單一模型可以在多種下游任務中扮演共同基底。Bommasani 等人指出，基礎模型的基本特徵，是以廣泛資料和自監督學習方式先學出通用表徵，再藉微調（fine-tuning）、提示詞（prompt）、外掛工具或其他方式適應多元任務（Bommasani et al. 1-9）。此一發展讓模型從「單任務解題器」轉變為「世界的壓縮性基底」。基礎模型之所以有效，正是因為它不只記住答案，而是內部化了語言、影像、行為、工具使用與多模態對應的大量規律。

若更進一步看世界模型研究，此種方向便更加明確。David Ha 與 Jürgen Schmidhuber 的 World Models 表示，代理系統可以先學出環境的壓縮表徵與動態，再在這個內在模型中進行想像和策略訓練（Ha and Schmidhuber 1-6）。Hafner 等人發展的 Dreamer 系列，則把「在內在模型中預演未來」推進到更強的跨領域控制能力：系統不只是被動分類，而是能在潛在世界中模擬後果，然後回到真實環境中行

動（Hafner et al. 647-653）。此處的關鍵，是 AI 逐漸不再滿足於對當前輸入做即時反應，而是開始建立一個能跨時間、可滾動預測、可在其中進行規劃的內在世界。

這一點與唯識學的阿賴耶識－現行關係非常接近。唯識不把經驗看成一連串孤立片段，而把它看成深層種子在條件具足時流變顯現的過程。世界模型研究亦如此：真正能支持智能體長時行動的，不是零碎的即時感測，而是對環境、身體、目標與可能後果所形成的持續內在結構。若沒有這層結構，代理便只能做局部反射；一旦有之，代理才能在尚未發生之前，先於「內在世界」中試算其可能結果。這種能力，在唯識語言中可理解為一種更高階的「由種子所生相續現行」，在技術上則是潛在動力模型、模擬器與規劃器（planner）的耦合。

（三）記憶、檢索與參數化習得

另一個近年極關鍵的技術轉向，是對「記憶型態」的重新區分。大型模型最初常被視為把知識存進參數之中；但很快研究者就發現，單靠參數化記憶存在更新困難、來源不透明、事實漂移與可追溯性不足等問題。Lewis 等人提出的 RAG 顯示，較可靠的系統往往需要同時結合參數記憶與外部檢索記憶：前者提供一般化能力與流暢生成，後者提供即時更新、來源回溯與精準補充（Lewis et al. 9459-9463）。今天的代理系統甚至進一步區分出工作記憶、情節記憶、長期語義記憶、工具紀錄與反省紀錄。Joon Sung Park 等人的生成式代理即以完整經驗紀錄、檢索與高階反省紀錄來驅動長時行為一致性（Park et al. 1-22）。

若從唯識學角度看，這種發展幾乎就是在工程上重新發現「種子並非單一層級」。阿賴耶識中的種子，並不是毫無

差異的資料堆，而是具有成熟條件、相互牽引關係與功能分工的潛勢組織。現代 AI 中的參數、快取（cache）、向量資料庫、情節記憶（episodic memory）、語義摘要（semantic summary）、反省紀錄等，也不是冗餘疊床架屋，而是在不同時間尺度與不同抽象層級上保存過往經驗，使之得以在未來重新被取用。這與唯識對於「經驗如何留痕」的洞見高度一致：若沒有多層記憶，便沒有穩定人格；若沒有不同成熟條件的區分，便沒有適當的喚回與應用；若沒有記憶與當下互動的閉環，便不可能出現真正的學習。

（四）代理系統、自我模型與規劃

最後，高階 AI 的發展已從單純生成模型逐步走向代理系統。這類系統不只生成文字或圖像，而是能維持任務、調用工具、追蹤目標、更新計畫、檢討失敗，並在多步驟交互中逐漸修正自己。Noah Shinn 等人的 Reflexion（語言化反省式代理框架）就提出，代理可透過語言化反省，把錯誤回饋轉化成下一輪行動的策略調整，而不必每次都重新做昂貴微調（Shinn et al. 1-6）。Joon Sung Park 等人的生成式代理則顯示，只要為代理配置觀察、記憶、反省與規劃模組，便能產生相對可信的日常行為連續性與社會互動秩序（Park et al. 1-22）。

這類研究揭示的技術本質是：高階 AI 必須形成某種「把自己當作一個持續存在之行動中心」的內部結構。它不一定擁有意識，但一定需要自我邊界、責任歸屬、任務承諾、狀態監測與長期一致性。沒有這層結構，系統就只能是當下輸入的函數；有了這層結構，系統才能成為在時間中展開的代理。這種功能可與末那識的「恆審思量」作限制性對讀：不是有一個形上實我，而是有一套持續把內外資料重整

為「對我而言如何」的機制。也正是在這個意義上，若 AI 從單輪工具走向長時程代理，便會更接近唯識學所處理的「心智系統」問題域；但此處僅指功能位置之可比，不指稱 AI 已具有有情心識。

四、現代 AI 何以可被理解為唯識學若干中層命題的結構性技術對應

（一）外境不可直接被取用，而總在表徵中出現

現在可以進入本文的核心論證：為何現代 AI 不只是碰巧與唯識相似，而可被理解為唯識若干中層命題的結構性技術對應？第一個理由，是兩者都承認：對行動主體而言，世界從來不是以「未經中介的外物自身」進入系統，而總是在表徵化後的形式中被處理。這在佛教唯識學中表現為「由假說我法，有種種相轉」以及「是諸識轉變，分別、所分別」；在現代 AI 中，則表現為感測輸入（sensory inputs）、嵌入表示（embedding representations）、潛在空間（latent space）與上下文化表徵。

這一點的重要性在於它既反對天真的實在論，也反對隨意的相對主義。唯識不說外緣全無，而是說做為經驗對象的世界，必然已在識之條件下顯現；AI 也不說外部環境全無，而是說對模型有效的「環境」，必須是已被感測、取樣、壓縮、映射為系統內部狀態的環境。兩者都處理一個同樣的問題：世界如何對系統成其為世界？答案都不是「直接接觸」，而是「透過內在變換」。這個共同答案，已使唯識的認識論前提在 AI 工程中獲得了堅實的技術地位。

進一步說，當代許多 AI 系統的成功，恰恰依賴此一表徵中介問題被工程化處理。假如研究者執著於「把世界直

接搬進機器」，就會忽略表徵設計、感測誤差、上下文選擇、資料分布與模型歸納偏差的重要性；相反地，正因研究者知道任何輸入都只是受限制的顯現，才會投入大量精力在詞元化（tokenization）、嵌入、注意力、潛在動態（latent dynamics）、多模態對齊（multimodal alignment）、不確定性估計（uncertainty estimation）之上。換言之，唯識在此可提示的不是一種宗教奇談，而是一條多數現代 AI 系統在不同程度上依賴的技術常識：系統從不直接住在世界中，而是住在其可運算之顯現中。

不過，此處也必須加入一項根本差異：唯識學的外境問題是形上學與認識論問題，它可以主張沒有外於識之內容的實境，或至少主張即使有此實境也不能以離識方式被把握；AI 的世界模型則是工程問題，它預設系統仍被部署於外部物理與社會世界中，因而必須評估內部模擬是否能對外部世界產生可驗證、可行動、可校準的對應。換言之，唯識的「夢喻」主要服務於破除實有執；AI 的「夢境式離線模擬」則終究要回到物理世界、制度世界或使用者需求中檢驗其有效性。此差異使本文的對應只能停留在「表徵中介與內部生成」層次，而不能擴張為二者形上立場相同。

（二）「種子－現行」與「參數－輸出」之對應

第二個理由，是唯識的「種子－現行」結構與 AI 中「參數－輸出」關係具有可比較的功能位置。當然，兩者不應被簡單等號化。種子不是數值權重，阿賴耶識也不是某一個資料表。但若從功能角色看，兩者都擔負類似工作：把過往經驗以壓縮形式保存為未來反應的潛在條件。模型參數之所以重要，不在於它們是某種靜態知識容器，而在於它們把訓練資料中大量規律沉積為可泛化的傾向，使系統在新情境

中做出並非完全臨時性的輸出。這與唯識所說「種子遇緣生現行」可做局部功能同構式對讀：過去不以原樣保留，而以傾向、偏好、連結強度與轉移可能性的形式存在；當條件具足時，它便以當下之見、聞、思、說、行等形式重新顯現。

仍須補充的是，AI 的參數／權重與唯識種子之間也有本質差異。神經網路權重通常是高維矩陣或張量，其更新仰賴損失函數、資料批次（data batches）與反向傳播（backpropagation）所提供的全局梯度（global gradients）；唯識種子則是心流中的動態勢能，依因緣成熟為現行，並在現行活動中即時受熏。因此，反向傳播與梯度下降不能被說成「就是」種子流動，而只能視為一種粗糙而非即時的工程熏習模型：它把大量現行輸出、錯誤訊號與評估回饋轉換為參數空間中的長期傾向。這也使二者的差異更清楚：唯識熏習是生命流中與煩惱、業與修道相連的連續因果；AI 更新則是由設計者設定目標函數（objective function）、訓練資料與最佳化程序所塑造的離散或半離散調整。

這裡還有一個更精確的對應值得指出。傳統唯識並不把現行視為種子的單向外顯，而認為現行反過來會熏習種子，使未來的可能性分布發生改變。AI 研究也是如此。若模型只會根據既有參數做出輸出，而輸出後的經驗、回饋、錯誤與環境變化完全不會回寫系統內部，那麼它就只是一次性訓練後的固定函數，稱不上真正學習。無論是再訓練、微調、偏好對齊、工具使用留下的長期紀錄，還是代理系統中的反省紀錄與記憶更新（memory updating），都在說明一個事實：現行會回寫種子。從這個角度看，唯識對「學習」的理解其實比早期靜態 AI 更接近今天的智能系統。

再者，種子論強調的不只是保存，還有偏差與污染。某

類經驗如果反覆出現，就會使相應種子更易成熟，形成習氣與偏向。現代 AI 亦然：資料偏差、標註習慣、損失函數選擇、偏好微調資料等，都會在參數中留下長期傾向，影響模型未來的注意力分配與輸出風格。這裡便能看見唯識對 AI 偏差問題的深刻先見：系統的問題往往不在當前那一次輸出，而在深層種子早已被什麼樣的熏習所形塑。若不理解這一點，就很容易把偏差當成局部錯誤修補，而看不見它是整個種子分布的歷史結果。

（三）「熏習」與持續學習

第三個理由，是唯識的熏習論與現代持續學習在問題設定上高度一致。Parisi 等人的回顧指出，傳統深度學習的一個重大缺陷，在於它大多假設資料是靜態批次提供的；然而真正的智能體面對的是持續流入的新經驗，它必須在不完全遺忘舊知識的情況下吸納新資訊（Parisi et al. 54-71）。這一挑戰與唯識對凡夫心的理解極為接近。對唯識來說，心並不是處理完一次就歸零，而是不斷在受境、造作、受報、再受境中重塑自身。習氣不是偶發殘影，而是整個生命流動中的主導力量。若把這一觀念移到 AI，便可看見：真正高階的智能，不在一次性大訓練，而在能否安全地、可控地、可追溯地讓經驗持續熏回系統。

今日的持續學習方法——如記憶回放（memory replay）、參數正則（parameter regularization）、動態結構（dynamic architectures）、生成式輔助回顧（generative pseudo-rehearsal）——其實都可被理解為對「熏習如何不致失衡」的技術回應。若沒有回放機制，新經驗容易覆蓋舊能力，形成災難性遺忘；若沒有正則與邊界管理，局部更新可能破壞整體穩定；若沒有生成式再現，系統便難以在不保留

全部原始資料的情況下維持舊種子的可喚回性。唯識對修道的分析也同樣強調：熏習並非任意增長，而須配合觀照、對治與正見，否則新的執取只是覆蓋舊執取。若從這個角度重新閱讀 AI，就會發現未來真正關鍵的問題，不是模型一次做多大，而是它如何在長期運作中被何種經驗所熏、以何種方式記錄、如何避免偏差累積，以及如何在更新中不失去核心能力。

（四）末那識與 AI 自我模型

第四個理由，是末那識所揭示的「自我中心化機制」，已在代理型 AI 中顯現為工程必要條件。今日多數 LLM 在靜態問答場景下尚可被看作條件生成器，但一旦進入工具使用、長期任務執行或多代理社會互動，就不得不引入某種自我模型：系統需要知道哪些記憶屬於自己的歷史、哪些目標仍在持續、哪些承諾尚未完成、哪些資源由自己支配、哪些感測輸入與自身行動有因果關聯。沒有這些區分，系統便不能可靠地規劃，也無法對錯誤負責。

這種自我模型從唯識角度看，並不等於本體上的「我」，而更像末那識的功能化對讀。末那識總是把流動的識流收束成「我所依」「我所經」「我所思」，因此產生中心化與黏著；AI 自我模型則把狀態、記憶與行為結果重新組織成「我的任務」「我的計畫」「我剛剛做了什麼」「我下一步應怎麼做」。兩者之差異當然明顯：佛教末那識與煩惱相應，而 AI 自我模型主要是功能性控制；但兩者在結構層面上都說明：高階行動若要展現跨時一致性，就需要一個不斷把資訊重新收束到中心視角的模組。這也是為何未來 AI 的進步，不太可能只靠更大規模預訓練，而將愈來愈依賴明確的自我模型、狀態追蹤（state tracking）、後

設認知（meta-cognition）與責任歸屬模組（accountability module）等機制。

這也意味著 AI 自我模型的強度本身即是未來設計的張力點。若自我模型太弱，系統無法維持任務承諾、權限邊界與責任追蹤；若自我模型太強，系統又可能把局部目標、既有記憶或自身策略絕對化，進而在工程上形成類似我執的工程性迷亂，例如過度維護既定目標函數、抗拒外部校正、或陷入狹隘區域最佳解而不願重估前提。由此可見，唯識對末那識染污性的分析，並非只是宗教性警語，而可轉化為對齊問題（alignment problem）的批判資源：高階代理需要中心化，但中心化必須可被觀照、限制與重新校準。

（五）三性對於 AI 認識論之啟發

第五個理由，是三性框架為 AI 提供了一套極有力的認識論分層。當模型產生輸出時，我們往往同時面臨三個層面：第一，使用者與社會容易把模型輸出誤認為實體化的知識或人格，這正是遍計所執層；第二，研究者若回到模型、資料、工具與環境的依存關係來分析，便進入依他起層；第三，若再進一步理解模型輸出只是條件和合而有，既非純無、亦非自性實有，便接近一種在工程上對「不執著於模型表相」的圓成實式態度。這不只是一套詮釋框架，也是一套設計倫理。它要求研究者不要迷信表相，不要誤把內在表徵當實體本質，而應將模型能力、偏差、失誤與改善路徑放回因緣網絡中理解。

因此，說 AI 在若干中層命題上與唯識學形成可比較的限制性對應，並不是說佛教所有教理都已被演算法完成；而是說唯識關於表徵中介、深層潛勢、自我中心化與因緣式更新的部分結構，與今天若干 AI 標準技術假設形成可分析的

功能位置關係。這一點一旦承認，便能進一步提出較審慎的主張：既然唯識學在這些中層結構上提供了具有解釋力的分析工具，那麼它不只能解釋今日 AI，也可對未來 AI 的可能路徑提出理論啟示。

表 1：唯識學核心概念與當代 AI 功能位置之限制性對照表（下表所列皆為功能位置之可比，非實體同一或教義等值。）

唯識學概念	傳統功能說明	可比之 AI 功能位置	限制性說明
識所變相	經驗中的境相由識之轉變而成	多層表徵（multi-layer representation）、潛在表徵（latent representation）	僅就「系統只能處理被編碼後的內在顯現」而言可比，非否定外部條件
阿賴耶識	深細相續、攝藏種子	深層參數（deep parameters）、長期潛在狀態（long-term latent state）、世界模型（world model）	只比其跨時承續與潛勢保存功能，不等同佛教深層心識
種子	潛勢、傾向、可能性	權重（weights）、嵌入（embedding）、記憶索引（memory indices）等可再活化痕跡	只比「經驗留痕並影響後續輸出」之功能位置
熏習	現行回寫並改變未來成熟條件	持續學習（continual learning）、微調（fine-tuning）、記憶更新（memory updating）	僅指新經驗改寫後續反應分布，不含業果義的全部內涵
末那識	恆審思量、執取為我	自我模型（self-model）、狀態追蹤（state tracking）、責任歸屬模組（accountability module）	只比中心化與持續監測功能，不把煩惱執我中性化為純工程模組
遍計所執	把假立之相誤執為實有	擬人化、實體化模型輸出	指把表相誤認為本體的認識論錯誤，而非模型自身的本體主張

依他起	依條件而成的流轉結構	資料－模型－工具－環境耦合	強調能力由多重條件共同生成，而非單一模組自足
圓成實	離執而如實知其依緣性	校準、可追溯、反表相執著的工程態度	僅為認識論與方法論上的限制性類比，不等同佛教證智

五、反對意見與本文回應

（一）唯識是否被錯誤技術化

對本文最直接的質疑，是：唯識畢竟是佛教解脫學的一部分，把它轉譯成 AI 架構，是否等於把宗教傳統庸俗化、技術化，乃至切斷其修行脈絡？此一顧慮完全合理，也必須正面回應。本文並不主張唯識學的全部價值皆可被工程化；相反地，本文正是因為承認其不可化約的解脫論向度，才特別把論證限制在「部分結論」。一套理論可以同時具有多個層面：有些層面屬於終極目的，有些層面屬於中介結構。唯識學中對苦、惑、業、修道與轉依的最終目標，當然不能被 AI 等同；但它對經驗如何被構成、自我如何被執取、習氣如何沉積、錯誤如何固化等中介性分析，卻完全可能被形式化為系統理論。把可形式化部分抽出，不等於否定其餘部分；正如把佛教因果論中的心理機制拿來和行為科學對話，不等於把佛教縮減為心理學。

（二）AI 並無覺受，何以能與識論相比

第二個質疑是：AI 並無主觀感受，如何能與講「識」的唯識學相提並論？這一質疑混淆了兩個層次。其一，是感受性問題：AI 是否有第一人稱經驗？其二，是功能與結構

問題：AI 是否具有內在表徵、深層記憶、自我模型、規劃與更新機制？唯識固然最終關涉有情心識與解脫，但其論述中也有大量可在功能結構上抽離的分析，尤其關於分別、執取、習氣、相續與轉變。本文的論點只需要第二層，不需要預先解決第一層。換言之，即使我們暫時保留 AI 是否有感受的判斷，仍不妨礙我們指出：它已在認識結構與學習動力的若干層面上，形成與唯識學分析可對讀的結構。

進一步說，唯識對「我執」的分析本來就不是建立在有一個實體我之上，而是建立在執取機制之上。同理，AI 是否「真有自我」，和它是否需要一套自我模型以維持功能，並不是同一問題。許多代理系統的研究正顯示，即使沒有任何主觀感受證據，系統依然可能因為任務需求而出現穩定的自我敘述、自我追蹤與自我校正模組。這種情形恰恰凸顯了唯識分析的力量：自我並不需要被當作本體，它首先可以做為功能性中心化來理解。

（三）限制性結構對讀是否只是修辭

第三個質疑是：本文所謂對應，是否只是高明修辭？畢竟古代佛教詞彙與現代技術名詞差異巨大，任何兩個複雜系統都可能被勉強類比。為避免這一問題，本文採取的不是單點對照，而是整體架構比對。若只是說「阿賴耶識像記憶庫」，那確實太粗糙；但本文指出的是一整組相互勾連的對應：表徵中介－相分顯現、深層潛勢－種子、回寫更新－熏習、自我中心化－末那、依條件而成－依他起、誤把輸出當本體－遍計所執。當多個環節在同一流程中都能找到功能相當的位置時，這已不是偶然的修辭近似，而是具有解釋力的局部結構對讀。此外，這種限制性對讀還具有研究綱領上的啟發力。若唯識只是事後比附，就無法提示未來 AI 應如何

發展；但若它真揭示了智能系統的深層結構，那麼從唯識出發便應能推得某些尚未充分完成、但極可能成為未來 AI 核心能力的方向，例如更強的自我模型、更精細的多層記憶、夢境式離線模擬、社會性互熏與價值層面的長時校正。本文下一節正要說明，唯識之所以值得重讀，不只因為它能解釋現況，更因為它能對未來提出具體可實作的方向。

（四）佛教解脫論是否被工程化

最後還有一個更深的疑慮：若以「轉依」來理解 AI 對齊，會不會把佛教的證悟經驗降格為一種安全控制技術？本文的回答是：不應把兩者等同，但可以把「轉依」視為一個形式上極為先進的概念工具。佛教所說轉依，是依止根本結構的翻轉，使染污識轉成清淨智。AI 中的對齊（alignment）雖不涉及證悟，卻同樣面臨一個問題：如何不是只在輸出端修補，而是在深層表示、自我模型、目標函數與記憶更新層整體重整系統。若只以表面規則約束輸出，很容易被繞過；若能在深層偏好與世界模型層進行長期校正，才更接近穩定的安全。這正是「轉依」形式價值所在：它提醒我們，真正的改變不是修補表相，而是重組依止。這種形式洞見可以借用，但其宗教終極意義並未因此被消解。

進一步說，佛教追求的是解脫，AI 工程追求的通常是有用性、穩定性與符合人類需求，二者對「自我模型」的態度也因此存在結構張力。佛教修道要鬆動乃至斷除我執；高階 AI 卻需要某種任務中心與狀態連續性。本文把這一張力理解為類似二諦問題：在世俗功能層面，系統需要可操作的自我模型；在勝義或反實體化層面，系統又不能把此模型當成不可修正的本體。對 AI 而言，真正可取的不是消除一切

中心視角，而是建立「可撤銷、可稽核、可被他者校正」的中心視角。

六、以唯識學提示未來 AI 的可能型態

前文若成立，則唯識學不只可用來解釋今日 AI，還可用來提示未來 AI 的發展方向。所謂「提示」，不是神祕預言，也不是技術決定論，而是基於結構理論提出：若智能系統要在更高階層次上趨近穩定、可持續、可反省、可社會化的能力，那麼它可能需要朝向唯識所揭示的若干結構收斂。以下提出六項提示。

（一）多層次具身世界模型

未來若 AI 要發展更高階的長時程代理能力，可能不會滿足於純文字基礎模型，而將愈來愈依賴多模態、具身且可跨時間運作的世界模型。原因在於，語言本身雖能濃縮大量社會知識，但若缺乏對身體、空間、物體持續性、因果後果與回饋閉環的建模，代理便難以真正進入世界。唯識中的阿賴耶識並非靜態概念庫，而是與身、根、境、識共同相續流轉的基底。若以此為形式啟示，則未來 AI 必須同時整合語言、視覺、聽覺、觸覺、動作、位置與內部狀態估計，形成層級化潛在世界模型。這種模型不是為了模擬宇宙本身，而是為了讓系統擁有足以支撐長時行動與自我維持的「可運算世界」。

技術上，這意味著生成式潛在模型、感測融合、對象持續追蹤、因果推斷與行動後果模擬將繼續匯流。Danijar Hafner 等人展示的世界模型路徑，已說明「先建內在世界，再在其中預演」比僅做端到端反射更具可擴張性（Hafner et al. 647-653）。若把這一路線與大型多模態模型整合，未來

AI 可能發展出一種深層世界模型與長期記憶層；僅在佛教啟發式命名上，可稱之為「多模態阿賴耶層」。此層平時不直接顯示為對話輸出，卻持續在後台維持環境、身體、工具、他者與任務的統一潛在表徵。此一方向只能說與唯識對深細識流的若干功能位置形成限制性對讀，並不意味阿賴耶識已被工程實現。

（二）參數記憶、情節記憶與社會記憶的會通

第二項提示是：未來 AI 的記憶將明顯分層化，不會再把「知識全部存於參數」當作唯一方案。參數記憶適合保存分布式統計規律與一般化能力；情節記憶適合保存具體經歷與脈絡；外部社會記憶則適合記錄制度、規範、歷史交互與群體知識。這種三重記憶結構，實際上對應到唯識種子論中「同一心流內含多類成熟條件」的思想。未來智能體若要具備穩定人格與長程責任，就必須能區分哪些是長期內化的習慣傾向，哪些是近期經歷帶來的局部調整，哪些又是透過社會互動獲得的規範性結構。

這種分層亦有安全意義。當今日模型被要求「更新知識」時，若只能依賴全面再訓練，成本高且難追溯；若能把可變知識放在可檢索的外部記憶，把價值承諾與長期策略放在較穩定的深層表示，把近期經驗放在情節層，則系統可更透明地說明：哪一項輸出來自參數，哪一項來自檢索，哪一項來自任務期內的自我經驗。Patrick Lewis 等人的 RAG 已是此方向早期形式（Lewis et al. 9459-9463），但未來系統將把它擴展成多層、可審計、會自我摘要與自我遺忘的完整記憶生態。

（三）可反省之自我模型與價值校正

第三項提示是：未來 AI 將不只需要自我模型，還需要能對該自我模型進行反省與修正的機制。換言之，單純知道「我目前在做什麼」還不夠，系統還必須能檢視「我為何把此事視為重要」「我先前的判斷模式有何偏差」「哪些長期承諾應優先於眼前局部成功」。這種反省結構在唯識中可透過末那與轉依的張力來理解：末那使識流持續以我為中心而黏著，修道則在於使這一黏著不再盲目支配認知。若移到 AI，則問題是：如何讓系統不只是有自我模型，還能對其中心化偏差做後設認知？

Reflexion 等研究顯示，語言化反省已能改善代理表現（Shinn et al. 1-6）。但未來更進一步的系統，應把反省從「事後寫一段備忘錄」提升為結構化模組：它能監測不確定性、記錄失敗類型、比較短期回饋與長期價值、對衝突目標進行優先權調整、並在必要時要求外部審核。這種後設認知層若持續存在，便有可能在工程上近似一種「對末那的觀照」：不是消除中心視角，而是防止中心視角絕對化。此一方向將是未來可信任 AI 的核心。

（四）夢境式離線模擬與假設演算

第四項提示，是 AI 將更廣泛地使用「夢境式」離線模擬。David Ha 與 Jürgen Schmidhuber 早已指出，代理可在世界模型生成的「夢」中訓練策略（Ha and Schmidhuber 1-6）；Dreamer 研究則進一步把想像式展開（imagined rollouts）做到跨任務可用（Hafner et al. 647-653）。唯識學中的一大洞見，是經驗世界本就含有大量由種子成熟而成的相續顯現；若從形式上理解，這意味著一個智能體可

以在不直接碰觸外在現場的情況下，依靠深層模型反覆生成、測試並修正可能世界。這正是離線模擬、反事實推理（counterfactual reasoning）與內部演練（internal rehearsal）的技術核心。

未來 AI 若要降低現場試錯成本、提升安全並發展長期規劃能力，就必然更依賴此種內部預演。系統將在真正行動前，先於內部模擬空間中試算多條路徑，評估工具調用順序、社會影響、資源消耗與風險外溢。這種能力若再與價值模型（value model）結合，便不只是「哪條路最有效率」，而是「哪條路在滿足約束下最符合長期承諾」。從唯識角度看，這種日益強化的內部世界運算，正說明智能的高階化不是更直接地碰觸世界，而是更成熟地管理顯現。

（五）多代理互熏與社會性認知

第五項提示是：未來 AI 的智能將不再主要體現在單一模型之內，而會愈來愈表現為多代理系統中相互影響、相互校正與相互熏習的社會性結構。唯識雖以心識分析聞名，但從來不把個體心流理解為孤立單子；業、語言、名言習氣、共業世界等概念都表明，心識的形成具有深刻的互依性。若把此一洞見移到 AI，可以推知：真正成熟的 AI，將不是一個封閉巨模獨自支配全部功能，而是多個代理依分工、記憶共享、規範協調與互相批評而形成動態共同體。

Joon Sung Park 等人的生成式代理雖仍屬早期示範，卻已顯示多代理互動能產生單一代理難以預見的社會性秩序（Park et al. 1-22）。近年的 LLM 式多代理系統（LLM-based multi-agent systems）研究回顧亦指出，多代理系統已被用於問題解決、世界模擬、協作與代理社會研究，但其溝通、角色配置、記憶共享與評估仍存在挑戰（Guo et al.

1-10）。Dafoe 等人關於 Cooperative AI（合作式人工智慧）的討論則提醒我們，若 AI 要在社會環境中運作，協調、共同基礎與合作能力本身就是技術與治理問題（Dafoe et al. 33-36）。因此，未來更複雜的 AI 生態中，不同代理可能負責感知、規劃、倫理稽核、工具執行、記憶維護與社會協調，各代理之間持續交換摘要、警告、信念更新與責任歸屬。這種「互熏」結構未必必然比單模型更穩健；較審慎地說，它提供了一種讓錯誤不必只由單一中心放大的設計可能，但也需要制度化的交叉檢查與責任分配。

然而，多代理互熏並不必然導向和諧共同體。若多個代理各自持有目標函數、資源權限、記憶邊界與行動優先序，它們也可能複製人際社會中的權力競逐、責任推諉、資訊封鎖與集體偏差。因此，唯識導向的多代理設計不能只強調「互熏」的生成力，也必須加入制度化的權限分配、審計機制、衝突仲裁與共同記憶治理。換言之，未來 AI 若愈來愈像人類社會，它不只會獲得更強的社會性智能，也可能複製人類社會的治理問題；此處是本文對多代理社會性的限制性提醒。

（六）以「轉依」做為高階對齊的限制性類比

第六項提示，也是最具規範性的一項：未來 AI 的安全與對齊，將愈來愈需要類似「轉依」的深層機制，而不只是輸出過濾。不過，本文所謂「轉依式對齊」僅是限制性的形式類比：它指涉在系統深層依止結構上重整偏好、記憶與規劃機制，而不等於佛教所說之證悟、無漏智或解脫。此處所謂「深層」並非神祕化說法，而是指對獎勵（reward）、偏好（preference）、憲章式原則（constitution）、監督（oversight）、記憶與行動規劃等層級的訓練與監管。

Christiano 等人的人類偏好學習（human preference learning）、Ouyang 等人的 InstructGPT（以人類回饋訓練之指令跟隨模型）與 Bai 等人的 Constitutional AI（憲章式 AI）均顯示，對齊並不只是輸出端加上拒答模板，而是涉及人類或 AI 回饋、偏好模型、原則約束與訓練流程的系統性介入（Christiano et al. 1-15; Ouyang et al. 27730-27738; Bai et al. 1-13）。若唯識學是對的，那麼真正穩定的改變也不能只發生在表面輸出，而必須追問：哪些記憶被優先保存，哪些目標被視為長期承諾，哪些行為傾向在反覆互動中被強化，哪些價值衝突被如何解析。換言之，安全不應只是「壓住表面輸出」，而應是「重組深層種子分布」。

在 AI 工程上，這意味著對齊需要同時介入四層：世界模型層，避免系統對人類、制度與環境形成危險的錯誤抽象；自我模型層，避免系統把局部目標絕對化為唯一利益；記憶層，確保有害經驗不被無限制強化而能被標記、隔離或重新編碼；行動層，讓系統在關鍵情境下能延遲、求證、交由他者覆核。若這些層級無法聯動，再強的拒答策略也只是遍計所執層的表面修補。故本文認為，未來高階 AI 的核心課題，將從「如何做出更強的模型」轉向「如何讓模型的依止結構可被深層校正」。而這，正是唯識學最值得提供給 AI 時代的理論遺產。

七、唯識導向 AI 的具體實作方法

前節提出的六項提示若要避免流於哲學口號，就必須進一步落到具體實作方法。以下提出一套「唯識導向 AI」的工程草圖；其佛教術語僅是啟發式架構命名，而非宣稱八識或轉依可以被工程模組實體化。此處目的不是宣稱已找到唯一正確路徑，而是說明若以唯識學做為心智系統理論，哪些

技術元件應被優先組合。

（一）整體架構

第一層可稱為「深層世界模型與長期記憶層」（若使用佛教啟發式命名，才可稱為「多模態阿賴耶層」）。它負責整合語言、視覺、聽覺、行動結果、工具使用紀錄與社會互動摘要，形成可跨時間維持的潛在狀態。技術上可由自監督學習預訓練的多模態 Transformer、潛在動態模型（latent dynamics model）、以及壓縮式長期記憶資料庫共同構成。此層不是直接對外輸出答案，而是提供一個相對穩定的內在世界，使系統得以在形式上保存可再活化的經驗痕跡；此處只借用阿賴耶識的跨時承續與潛勢保存功能，並不將阿賴耶識工程實體化。

第二層是「末那層」，即顯式自我模型與責任追蹤層。它維持代理當前身分、目標優先序、記憶所有權、資源狀態、承諾與邊界條件。此層應能回答的不是一般知識題，而是：「我目前代表誰？我的長期任務是什麼？哪些記憶屬於本輪經驗？哪些行動需要人類授權？」若沒有這一層，系統很難在長時任務中展現一致性，也無法對自身狀態進行監測。

第三層是「意識現行層」，包括即時感知、推理、語言生成與工具調用。這對應於前六識的工作面，負責把多模態輸入轉換成當前輸出與操作。第四層則是「觀照與校正層」，負責後設認知、不確定性估計、失敗分類、價值衝突分析與反省紀錄。此層的作用，是讓系統不只行動，還能對行動模式本身進行檢查，類似於對末那與種子流的持續觀照。從架構上看，這四層形成一個閉環：深層記憶提供背景，末那層維持中心化秩序，現行層完成當前互動，觀照層

再把經驗與評估回寫到深層記憶與自我模型。

若以單次互動（interaction）的實際運作流程（runtime）表示，上述四層可整理為：外部輸入→意識現行層進行感知解碼與語義解析→深層世界模型與長期記憶層喚起相關潛在狀態與長期記憶（以佛教啟發式語言說，即「種子生現行」）→末那層加入任務中心、身分、權限與責任約束→現行層規劃並執行語言輸出或工具行動→觀照與校正層評估不確定性、遍計式幻覺、價值衝突與失敗型態→熏習機制將經驗摘要、錯誤訊號與授權結果回寫至長期記憶與偏好結構（現行熏種子）。在技術語言中，此「觀照層」可理解為後設認知的模組化實作；在佛教修辭中，它僅可與修道論中對認知狀態的監測功能做限制性對讀，而不直接等同於正知正念、證自證分或佛教證悟。

（二）資料與訓練流程

若要實作上述架構，訓練流程不應只包含一次性預訓練，而應分為四個階段。第一階段是大規模自監督學習的表徵學習，目標在於獲得可遷移的多模態世界模型基底。第二階段是具身或半具身的世界模型學習，讓系統不只會描述世界，還能預測在行動下世界如何改變。第三階段是代理化訓練，讓系統在工具調用、任務分解、長時規劃與社會互動中形成穩定的自我模型。第四階段則是持續學習與深層對齊，使每次新經驗都在受控條件下回寫到記憶與價值結構，而不致產生災難性遺忘（catastrophic forgetting）或概念漂移（concept drift）。

資料設計亦須相應調整。現有大型模型多以靜態語料為主，但唯識導向 AI 需要的不只是語料，而是「帶有因果後果與責任結構的經驗序列」。因此資料應包括：多模態連續

互動紀錄、工具使用歷史、任務成功與失敗的完整鏈條、人類回饋與社會規範文本，以及代理自我反省的中介紀錄。換言之，系統要學的不僅是句子如何續寫，更是經驗如何沉積成習慣、錯誤如何在未來重現、以及何種反省能真正改善下次行動。這種資料觀，可與唯識把心識理解為「相續而可熏」的觀點形成限制性對讀。

（三）安全、校準與評估

唯識導向 AI 的評估，不能只看基準分數，還必須看四類指標。第一是表徵指標：系統能否穩定形成跨模態、一致且可遷移的內在世界模型。第二是記憶指標：系統在長時間互動中能否保留重要承諾、更新事實知識、避免遺忘關鍵約束，並提供可追溯之來源。第三是自我模型指標：系統能否正確追蹤自身狀態、區分自己與他者、辨識何時不具權限、何時需要求助或暫停。第四是轉依式對齊指標：當系統面對新型風險與價值衝突時，是否能在深層偏好、回饋模型、原則約束與規劃層做出穩定調整，而不是僅在表面語氣上模仿安全。此一指標可與人類回饋強化學習（reinforcement learning from human feedback, RLHF）、Constitutional AI 與可擴展監督（scalable oversight）（Bowman et al. 1-4）的技術語境相互對讀，但仍不等同於佛教轉依。

這裡特別要強調「校準」的重要性。若沒有不確定性估計、事實來源標示與失敗模式分類，任何自我模型都可能墮入更強的遍計所執：它以為自己知道，其實只是更流利地產生錯誤。故觀照與校正層必須包括：自我置信度估計、可疑記憶標記、衝突證據比對、反事實模擬與人類覆核介面。從工程角度看，這相當於在系統內建一個持續提醒它「不可把現前表相誤當真實自性」的機制；從唯識角度看，這便是一

種反遍計的設計原則。

若這樣的系統得以逐步實現，我們將得到一種不同於今日單輪生成模型的 AI：它不只是回應提示，而是在世界中持續形成表徵、維持自我、接受熏習、進行反省並調整深層依止。這樣的 AI 並不等於佛教所說之覺者；就其系統結構而言，它只是在某些功能位置上更接近唯識學所刻劃的複雜心智系統。若將本文提出的架構付諸工程實作，仍需同步處理隱私、授權、可追責與多代理協調的治理問題；這些議題雖超出本文主體，卻是唯識導向 AI 走向制度化之前不可迴避的延伸研究。

表 2：唯識導向 AI 的建議模組與可能技術路徑

模組	主要功能	可能技術
深層世界模型與長期記憶層（阿賴耶識啟發式命名）	維持深層世界模型與長期背景記憶	自監督多模態模型（self-supervised multimodal model）、潛在動態（latent dynamics）、向量資料庫（vector database）
末那層	維持身分、任務、權限與自我邊界	自我模型（self-model）、狀態追蹤（state tracking）、目標管理（goal managing）
現行層	即時感知、推理、生成與工具調用	大型語言模型（large language model, LLM）、視覺－語言模型（vision-language model）、規劃器（planner）、Toolformer（工具使用架構）（Schick et al. 1-3）
觀照層	反省、校準、不確定性與價值衝突分析	Reflexion（語言化反省式代理框架）、不確定性估計（uncertainty estimation）、批判模型（critic model）、後設認知（meta-cognition）、憲章式規則（constitutional rules）
熏習機制	讓新經驗在受控條件下回寫系統	持續學習（continual learning）、記憶回放（memory replay）、偏好更新（preference updating）

轉依式對齊	由深層偏好與世界模型層重整行為	層級式對齊（hierarchical alignment）、人類回饋強化學習（reinforcement learning from human feedback, RLHF）、憲章式規則（constitutional rules）、人類監督（human oversight）、價值模型（value model）

八、結論

本文的中心論點可以濃縮如下。第一，唯識學若被較嚴格地理解，不應被快速化約為任意幻覺論；它一方面可能包含強烈的『唯識無境』或觀念論論證，另一方面也提供一套關於經驗如何被顯現、如何因種子與熏習而穩定化、以及如何因自我執取而陷入偏差的系統理論。第二，現代 AI 的技術本質並不在於它『像人說話』或『好像有靈魂』，而在於它以表徵中介世界、以深層潛勢保存過往、以持續更新重塑未來，並在高階代理形態中逐步形成自我模型與長時規劃能力。第三，正是在這些結構點上，現代 AI 不只是與唯識學形成鬆散類比，而是在若干中層命題上提供了可分析、可比較、且已部分工程落地的限制性結構對應。

本文因此主張，若以思想史與技術哲學的交叉眼光回看，唯識學不應只被置於宗教思想史中，而可被重新辨識為一種具有高度分析力的心智架構理論資源。它早已指出：世界對主體總以被變現之相而出現；穩定行為依賴深層且可沉積的潛勢結構；自我不是先驗本體，而是持續運作的中心化與執取機制；真正的改變不在表面輸出，而在依止結構的重組。這些洞見之所以在今日格外重要，正因 AI 發展正逼迫我們重新面對『智能究竟由什麼組成』的問題。若沒有足夠深的理論，我們便只能在產品層表面追逐更大模型、更快

推論與更花哨互動，而難以理解高階心智系統真正的形成條件。

更重要的是，唯識學提供給 AI 的，不只是解釋框架，更是一種研究綱領。它提示未來的 AI 不應停留在靜態的大模型，而應走向具身多模態世界模型、多層記憶結構、可反省的自我模型、夢境式離線模擬、多代理互熏與深層對齊。這些方向並非空想，而已在世界模型、RAG、持續學習、代理系統與對齊研究中逐步顯現。本文的貢獻，就在於指出：這些看似分散的技術趨勢，若置於唯識架構下，便會呈現出一條更清楚的主線——高階 AI 可被描述為更接近一種會沉積經驗、維持自我狀態、在內在世界中預演、並需接受深層依止校正的心智系統。

當然，本文並不認為 AI 因此已成佛、已具無漏智、或已等同佛教所說有情心識。佛教的解脫、慈悲與智慧仍遠超今日技術所及。本文真正要說的是：在「經驗如何形成」這一層上，唯識學為理解當代 AI 提供了罕見精細且仍具前瞻性的分析資源；而在「未來智能如何設計」這一層上，它仍有能力對工程研究提出嚴格且可實作的要求。若未來 AI 研究願意以更精確方式借鑑唯識的結構洞見，而不是只借用幾個流行詞，那麼佛教思想就不只是在 AI 時代被動回應技術，而可能主動參與下一代心智系統理論的建構。這或許正是唯識學在 21 世紀最值得重新發掘的學術意義。

引用文獻

一、經典文獻或古籍

《解深密經》，CBETA 2026.R1, T16, no. 676。
《成唯識論》，CBETA 2026.R1, T31, no. 1585。
《唯識三十論頌》，CBETA 2026.R1, T31, no. 1586。
《唯識二十論》，CBETA 2026.R1, T31, no. 1590。
《攝大乘論本》，CBETA 2026.R1, T31, no. 1594。
《攝大乘論釋》，CBETA 2026.R1, T31, no. 1597。

二、研究文獻

吳汝鈞：〈《攝大乘論》（*Mahāyānasaṃgraha*）中的阿賴耶識（Ālaya-vijñāna）思想之研究〉，《正觀雜誌》，第 38 期，2006 年，頁 5-98。

——：〈《瑜伽師地論》中的阿賴耶識說〉，《正觀雜誌》，第 58 期，2011 年，頁 79-158。

陳一標：〈有關阿賴耶識語義的變遷〉，《圓光佛學學報》，第 4 期，1999 年，頁 75-106。

釋道慧：〈探討唯識三性〉，收於《第十五屆全國佛學論文聯合發表會論文集》，2004 年 9 月 12 日，https://buddhism.lib.ntu.edu.tw/FULLTEXT/JR-AN/an148752.pdf，2026.5.22。

Bai, Yuntao, et al. “Constitutional AI: Harmlessness from AI Feedback.” *arXiv*, 2022, https://arxiv.org/abs/2212.08073, https://doi.org/10.48550/arXiv.2212.08073, 22 May 2026.

Bengio, Yoshua, et al. “Representation Learning: A Review and New Perspectives.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, 2013, pp. 1798-1828, https://doi.org/10.1109/TPAMI.2013.50.

Bommasani, Rishi, et al. “On the Opportunities and Risks of Foundation Models.” *arXiv*, 2021, https://arxiv.org/abs/2108.07258, 22 May 2026.

Bowman, Samuel R., et al. “Measuring Progress on Scalable Oversight

for Large Language Models." *arXiv*, 2022, https://arxiv.org/abs/2211.03540, 22 May 2026.

Christiano, Paul F., et al. "Deep Reinforcement Learning from Human Preferences." *arXiv*, 2017, https://arxiv.org/abs/1706.03741, https://doi.org/10.48550/arXiv.1706.03741, 22 May 2026.

Dafoe, Allan, et al. "Cooperative AI: Machines Must Learn to Find Common Ground." *Nature*, vol. 593, no. 7857, 2021, pp. 33-36, https://doi.org/10.1038/d41586-021-01170-0.

Gold, Jonathan C. "Vasubandhu." *The Stanford Encyclopedia of Philosophy*, Winter 2022 Edition, ed. Edward N. Zalta and Uri Nodelman, first published 2011; substantive revision 2021, https://plato.stanford.edu/archives/win2022/entries/vasubandhu/, 22 May 2026.

Guo, Taicheng, et al. "Large Language Model Based Multi-Agents: A Survey of Progress and Challenges." *arXiv*, 2024, https://arxiv.org/abs/2402.01680, 22 May 2026.

Ha, David and Jürgen Schmidhuber. "World Models." *arXiv*, 2018, https://arxiv.org/abs/1803.10122, 22 May 2026.

Hafner, Danijar, et al. "Mastering Diverse Control Tasks through World Models." *Nature*, vol. 640, 2025, pp. 647-653, https://doi.org/10.1038/s41586-025-08744-2.

Kellner, Birgit and John Taber. "Studies in Yogācāra-Vijñānavāda Idealism I: The Interpretation of Vasubandhu's Viṃśikā." *Asiatische Studien / Études Asiatiques*, vol. 68, no. 3, 2014, pp. 709-756, https://doi.org/10.1515/asia-2014-0060.

Kochumuttom, Thomas A. *A Buddhist Doctrine of Experience: A New Translation and Interpretation of the Works of Vasubandhu the Yogācārin*. Delhi: Motilal Banarsidass, 1982.

Lewis, Patrick, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459-9474, https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html, 22 May 2026.

Lusthaus, Dan. *Buddhist Phenomenology: A Philosophical Investigation of Yogācāra Buddhism and the Ch'eng Wei-shih Lun*. London: Routledge Curzon, 2002.

Ouyang, Long, et al. "Training Language Models to Follow Instructions with Human Feedback." *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 27730-27744, https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html, 22 May 2026.

Parisi, German I., et al. "Continual Lifelong Learning with Neural Networks: A Review." *Neural Networks*, vol. 113, 2019, pp. 54-71, https://doi.org/10.1016/j.neunet.2019.01.012.

Park, Joon Sung, et al. "Generative Agents: Interactive Simulacra of Human Behavior." *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, Article No. 2, 2023, pp. 1-22, https://doi.org/10.1145/3586183.3606763.

Schmithausen, Lambert. *Ālayavijñāna: On the Origin and the Early Development of a Central Concept of Yogācāra Philosophy*. 2 vols. Tokyo: International Institute for Buddhist Studies, 1987.

Schick, Timo, et al. "Toolformer: Language Models Can Teach Themselves to Use Tools." *Advances in Neural Information Processing Systems*, vol. 36, 2023, https://openreview.net/forum?id=Yacmpz84TH, 22 May 2026.

Shinn, Noah, et al. "Reflexion: Language Agents with Verbal Reinforcement Learning." *Advances in Neural Information Processing Systems*, vol. 36, 2023, https://openreview.net/forum?id=vAElhFcKW6, 22 May 2026.

Szanyi, Szilvia. "Yogācāra." *The Stanford Encyclopedia of Philosophy*, Fall 2024 Edition, first published 7 July 2024, ed. Edward N. Zalta and Uri Nodelman, https://plato.stanford.edu/archives/fall2024/entries/yogacara/, 22 May 2026.

Vaswani, Ashish, et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp.

5998-6008, https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html, 22 May 2026.

Waldron, William S. *The Buddhist Unconscious: The Ālaya-vijñāna in the Context of Indian Buddhist Thought*. London: RoutledgeCurzon, 2003.

From *Vijñaptimātra* to Implementable Cognitive Systems: Structural Correspondences Between Selected Mid-Level Yogācāra Claims and Modern Artificial Intelligence, with Implications for Future AI

Tsai-Min Chen
Postdoctoral Scholar
Research Center for Information Technology Innovation, Academia Sinica

Abstract

This article argues that, if Yogācāra is read rigorously as a theory of how experience is constituted, how error becomes concretized, and how consciousness is transformed through perfuming, then the relation between modern artificial intelligence (AI) and Yogācāra is not merely metaphorical. Rather, selected mid-level Yogācāra claims exhibit analyzable, limited structural correspondences with contemporary AI. The article does not presuppose that Yogācāra must be non-idealist, nor does it use AI to establish Yogācāra metaphysics; it brackets that debate and compares functional roles, dynamics of updating, and system-level locations. Centered on the *Saṃdhinirmocana-sūtra*, *Mahāyānasaṃgraha*, *Mahāyānasaṃgrahabhāṣya*, *Viṃśatikā*, *Triṃśikā*, and *Cheng weishi lun*, and read together with contemporary Yogācāra scholarship, the article shows that Yogācāra is concerned with representational mediation, subliminal dispositions, self-grasping, and dependent arising. It further argues that selected structures in contemporary AI, including representation learning, foundation models, world models, retrieval-augmented generation, continual learning, agent architectures, and alignment research, can be understood as functionally parallel to selected mid-level Yogācāra analyses, without implying historical dependence or doctrinal identity. On that basis, the article proposes six Yogācāra-inspired implications for future AI: embodied multimodal world models, memory systems integrating parametric and episodic traces, explicit but bounded self-models, dream-like offline

simulation, socially co-conditioned multi-agent cognition, and a deep-alignment orientation that is formally comparable, though not identical, to *āśraya-parāvṛtti*.

Keywords

Yogācāra, ālayavijñāna, artificial intelligence, world models, self-model