第二章　文獻回顧

2.1　軟體代理人

代理人最早起源於網路，應用於網路世界的資訊流通與控管，主要所應用的技術為分散式人工智慧(Distributed Artificial Intelligence, DAI)。直到最近代理人的發展，依然延續許多當初分散式人工智慧的功能特質。分散式人工智慧(Distributed Artificial Intelligence, DAI)是人工智慧研究中的一個副領域，研究的目標是為複雜問題開發出解決方案。在 DAI 研究中包含了以下這些主流：

- ➔平行問題的處理(parallel problem solving)：主要處理如何修改人工智慧內容，以便使多重處理器系統(multiprocessor systems)與電腦叢集(clusters of computers)加快整體計算的速度。
- ➔分散式問題處理(distributed problem solving, DPS)：利用代理人的操作概念，擁有自主權的實體能夠相互溝通，協力或分散解決問題。
- ➔基於多重代理人的模擬(Multi-Agent Based Simulation, MABS)：模擬的基礎，分析各層級的現象，像是許多社會性的群組模擬(social simulation)解決多個大小層級的問題。
- ➔系統的共同社會(scientific communities)：以社會學(sociology)與人類學(anthropology)作為科學研究的基礎。

在 DPS 與 MABS 運作的過程中共同所用到的重要觀念稱作為代理人。代理人是一個擁有自主性的實體，它能夠了解環境並並根據它來改變自己。一個代理人在同一系統下通常能夠去與其他代理人溝通並完成公共的目標，除非工作目標代理人可以獨自完成。在分散式人工智慧中可以將代理人分成以下各類：

- ➔反應型代理人(reactive agent)：當代理人接收到資料時將類似自動機械裝置，處理外來資料並產生輸出資料。
- ➔審議型代理人(deliberative agent)：審議型代理人與其他代理人對照之下對於週遭環境擁有自己的觀點，並能夠遵照自己的計畫進行。
- ➔混合型代理人(hybrid agent)：此類型代理人至少混合上述兩種類型的特色，能夠遵照自己的計畫行為，有時候也能夠藉由外部事件直接反應而沒有經過審議計畫。

以人工智慧研究初期來說，研究人員當初致力於更長遠的目標，就是所謂的“強化人工智慧”(strong AI)；一種更完整、更貼近於真實人類的人工智慧。在一些電影的情節中具有人性的完美電腦被虛構的呈現渲染出來，然而當時在近期之內這個目標不可能完成，並且已經不再是人工智慧研究領域中熱中的題材。AI 這個稱號自從早期的期望轉於落空之後有幾分失敗的感覺，經由一些大眾科學作家(popular science writer)對於人工智慧寫出距離實際研究成果很遠的誇大不實報導之後，稱號的名聲持續惡化。因為這些原因，一些原本從事人工智慧研究的學者對外改稱他們所研究的是認知科學(cognitive science)、資訊學(informatics)、統計推論科學(statistical inference)、或者是資訊工程學(information engineering)，企圖遠離先前人工智慧這個稱號所造成的不實形象。

人工智慧又稱做機械智慧(machine intelligence)，指的是經由任何特製系統所產生的智能應用。近代人工智慧研究生產出擁有明智行為並自動完成人類工作的機械，類似的應用像是為陸軍單位排定所有的資源列表、為顧客回答關於產品的問題、了解並抄寫人類說話的內容、以及經由攝影機畫面辨識人臉...等等。就人工智慧本身而論，這已經成為工程的學科，針對人類現實生活所遇到的實際問題提供解決方案。在 1991 年的波斯灣戰爭中，美國政府利用電腦人工智慧的方法來排定部署各軍事部隊，以提高戰略的功效。人工智慧系統對於現今全球許多的商業領域、醫療機構以及軍事單位都已成為不可或缺的例行工作要素，包括家裡所放置電腦中的軟體與家庭遊樂器(video games)都有它的存在(如圖示 2-1)。

很多人工智慧研究的雛形是經由心理學(psychology)實驗方法來描繪出來，實驗方法強調的是語言方面的智慧(linguistic intelligence)，最有名的例子是 1950 年由 Alan Turing 所做的實驗“Turing Test”。實驗中受測者面向電腦螢幕，螢幕後方有兩個操作者，分別是人與電腦程式。受測者將花費一段時間藉由螢幕針對兩位操作者發出一連串的問題，如果受測者無法分辨回應者是人或是電腦程式，那麼實驗的目的就成功了。

圖 2-1: 任天堂新世代遊戲主機(Nintendo, 2005)

離開語意學的範疇，人工智慧研究包含了機械人學(robotics)與集體智慧運作方法(collective intelligence approaches)，集中在環境上的有效運作(active manipulation)以及協同決策選擇(consensus decision making)；許多概念是從生物學及政治學(political science)的角度出發，藉以組合成具有智慧的行為能力。人工智慧理論也從生物研究上得到概念，特別是在昆蟲身上，所得到較為簡單的決策行為進而轉化為機械人。隨著比如像黑猩猩一樣更複雜的決策過程，在許多方面類似於人類但擁有較少的已開發能力，例如計畫與決策的過程。研究學者認為動物在整體複雜度遠低於人類，應該更加簡單地去模擬，但是令人滿意的電腦軟體模型目前尚未出現。

從 1950 年到今日，人工智慧技術開發出電腦對於知識表達、推理、學習能力等領域的研究，都有著相當明顯的進展。在未來的世界中，電腦與軟體之間合作與溝的通組織結構複雜性不斷地增加，傳統的軟體設計方式似乎開始無法滿足實際需求；分散式、具備智慧分析能力是今後軟體發展的基本方向。分散式的目的是要將問題進行分工，由多個代理人模組或網絡節點來共同完成問題的解答；而智慧分析能力的目標是要在模組與模組之間安排工作行為的協調，兩個概念的結合就產生了代理人軟體系統(Nwana, 1996)。

BEHAVIOR - 關於代理人在動畫遊戲模擬應用上，市面上所推出的商業軟體很多，有些軟體除了本身支援遊戲與動畫市場外，另外也推出軍事上的模擬版本，提供美國陸軍作戰戰略評估的練習。美國AVID公司所推出的三維代理人行為模組BEHAVIOR，是一個附加在主流三維動畫製作軟體SOFTIMAGE之下的擴充套件；它在三維模擬場景中同時提供多個代理人運行與互動的可能性，並且能感測反應周圍的場景，代理人本身的行為動作可以直接套用動作擷取系統(motion capture)所得資料，而動作與動作之間能夠自動銜接(如圖示 2-2)。在決策程度上，代理人可進行地形追蹤(terrain following)、動態的路線規劃(dynamic path-planning)、曲線行走、障礙物偵測迴避(obstacle avoidance)，也提供軟體開發工具(SDK)供使用者自行調整(http://www.softimage.com/products/behavior/)。

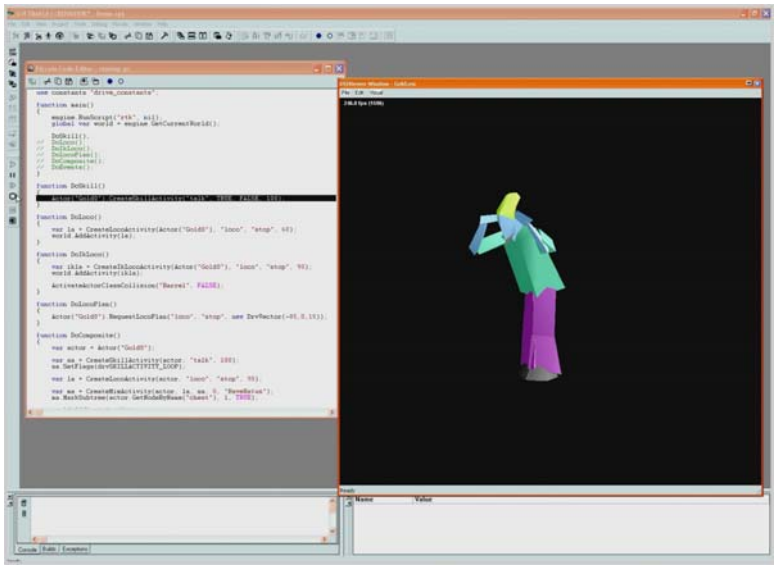

圖 2-2: BEHAVIOR 模擬軟體執行畫面 (Softimage, 2004)

Massive - 另一款與 BEHAVIOR 相似的軟體是 Massive software 所推出的 Massive，它是一個能夠產生電影電視中群眾模擬視覺效果的三維動畫系統；由美國 New Line Cinema 公司製作的電影“魔戒(The Lord Of The Rings)”中有不少的場景是利用 Massive 來加以製作，比如在電影中所呈現滿坑滿谷的士兵，每個士兵都有著自己的個性與行為。人物的性格反應決定了他要做什麼以及如何去做，他們的反應甚至可以模擬情緒的表達，例如表現的很勇敢、懦弱、或是歡樂…等等。人物以他們自己的方式去表達自己，就像代理人一般。當代理人人數增加至成千上萬時，在人群之中所呈現的互動行為還是相當具有真實性。代理人的行為系統也可以與動作擷取系統合作，當一個代理人在場景中被擊昏或是摔出車外，可利用真實演員預先錄製好的動作資料來控制影響的範圍。

Massive 也利用架構在人工生命(artificial life)的技術使得代理人更加的生動，代理人能夠使用聽覺與觸覺來感測並自然的回應所處的環境中。人工生命是人工智慧中展現自然界生命環境特質的系統研究，其中包含了生命的組織行為(self-organisation)、生命的調適性(adaptation)、生命的進化(evolution)、以及生命的新陳代謝(metabolism)…等等。生命行為的研究泛指在地球上各類的生命體，並沒有針對特定的生命族群。Massive 所模擬的對象不侷限於人類行為，經由修改骨架設定以及一些行為上的參數，其他生物的行為便可以在 Massive 系統中模擬(如圖示 2-3)。

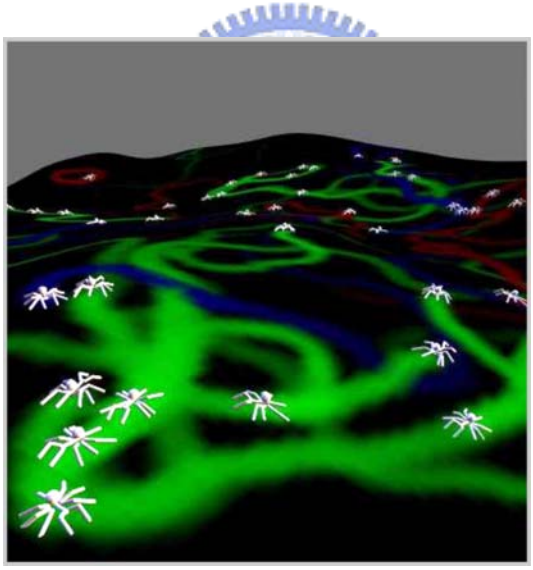

圖 2-3: 在 Massive 中的螞蟻行為模擬 (Massive software, 2005)

Massive 的動作系統加入了模糊邏輯使得代理人行為表現更加自然，行為參數的設定並非絕對的正直或負值；可以將數值調整成逐漸變化或是模糊的灰色地帶，使代理人能夠以自然的行為方式回應而不是像機械人以二進位的開關邏輯所產生的結果。模糊邏輯起源於 1965 年美國加州柏克萊大學(Berkeley)的查德(L.A. Zadeh)教授，在資訊與控制(Information and Control)學術期刊上所發表的論文：模糊集合(Fuzzy Sets) 。電腦的強項在於能夠處理複雜的計算與記憶深度，但對於人類所擅長的推理、聯想…等等，就無法以有效的計算來產生類似結果。因此模糊邏輯就是針對訊息不完全的資料或可能的線索，來產生近似值推理(Approximation reasoning)或是不同層級的結果(http://www.massivesoftware.com/)。

AI.implant 是由 BioGraphic Technologies 公司所研發的商業軟體，AI.implant 可應用於三種不同的商業領域，分別是遊戲、動畫與模擬。其中在模擬應用方面主要集中在武裝軍隊(armed forces)以及緊急事件(emergency services)的模擬，藉由即時互動的虛擬戰場以及前線的平民百姓達到模擬訓練的目的(如圖示 2-4)。除了精確地描繪出巨大尺度軍事部隊與平民的群眾模擬(crowd simulation)，同時能夠停止所有的活動和互動，包括一對一的戰鬥(combat)、代理人在寬廣開放的空間開車奔馳，也能經過稠密的都市地區與建築物室內空間…等等。

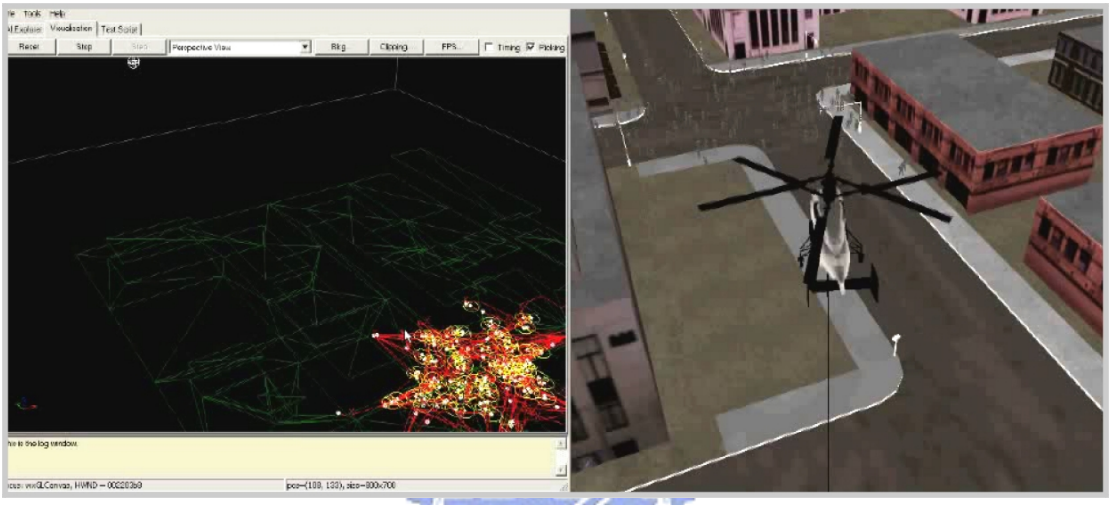

圖 2-4: 在 AI.implant 中的軍事模擬 (BioGraphic Technologies, 2005)

AI.implant 也提供豐富的參數平台使代理人能有複雜的智慧行為，經由決策流程圖(decision trees)或是編輯腳本(scripting)。代理人在與其他同性質軟體所擁有的功能比較時，在路線的事先預測與事件處理上具有更多元的功能與選項，比如代理人在虛擬都市環境或是廣大開放空間中平滑路線(Smooth navigation)的預先修正(如圖示 2-5)，以及動態地障礙物閃躲(dynamic obstacle avoidance)功能等(http://www.biographictech.com/)。

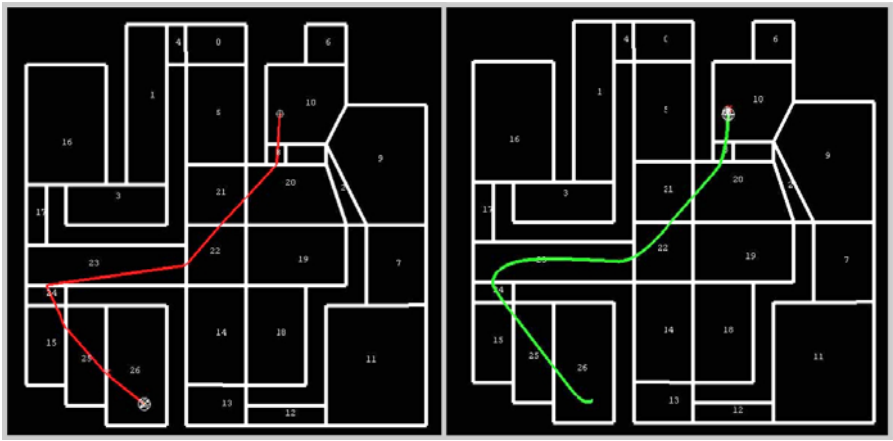

圖 2-5: 在 AI.implant 中的路線平滑處理 (BioGraphic Technologies, 2005)

2.2　虛擬都市中的行人模擬

許多研究集中於結合都市實體資料以及電腦圖像的呈現方式來創造一個複雜的虛擬互動環境，其中包含能夠像普通人類一樣反應的虛擬行人(virtual pedestrian)。部分研究主題是朝向虛擬都市中人群模擬所需要的電腦圖像即時(real-time)效能。倫敦大學所提出的專案中提到，在即時電腦圖像的狀態下成功將行人數量增加至一萬人，並且把模擬的目標設定為增進行人之間區域性以及廣域的相互作用(如圖示 2-6)；文中提到雖然將電腦的即時效能推至極限，同時也犧牲了行人在行為與動作上應有的複雜度(Loscos, Marchal & Meyer, 2003)。接著有研究提出過度簡化行人複雜度也等同失去模擬的意義，便將電腦圖像即時呈現的人數下修為七千人(如圖示 2-7)，並提升人群活動的互動功能(Marchal, 2002)。

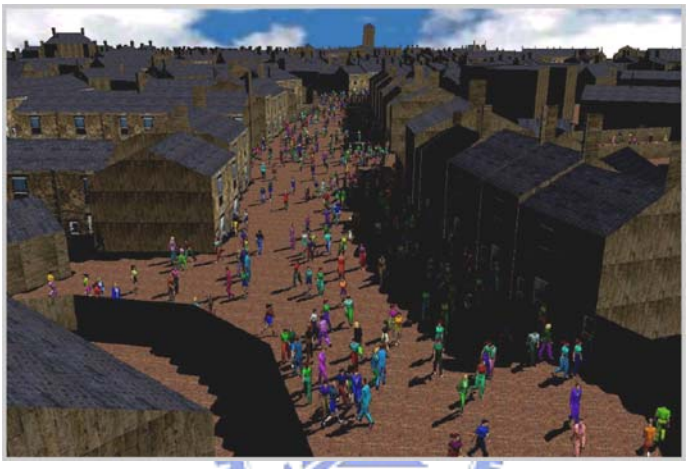

圖 2-6: 10000 個行人在模擬中即時運算 (Marchal, 2002)

行人在環境中最間單的活動就是行走，除了與週遭物件的碰撞偵測(collision detection)之外，行人與行人之間的協調也是相當重要(如圖示 2-8)，人群以自然合作的方式行走而形成串流(pedestrians streams)，在行走時也關係到軌道(trajectory)計算，如何產生流暢沒有折角的路線，以產生自然的人群移動。為了得到高度真實性的群體行為活動，碰撞偵測將是整體行為的主要關鍵；偵測值數值越高，所呈現的人群串流真實度與連鎖反應就越細膩，但與整體效能呈直接的反比(Marchal, 2002)。

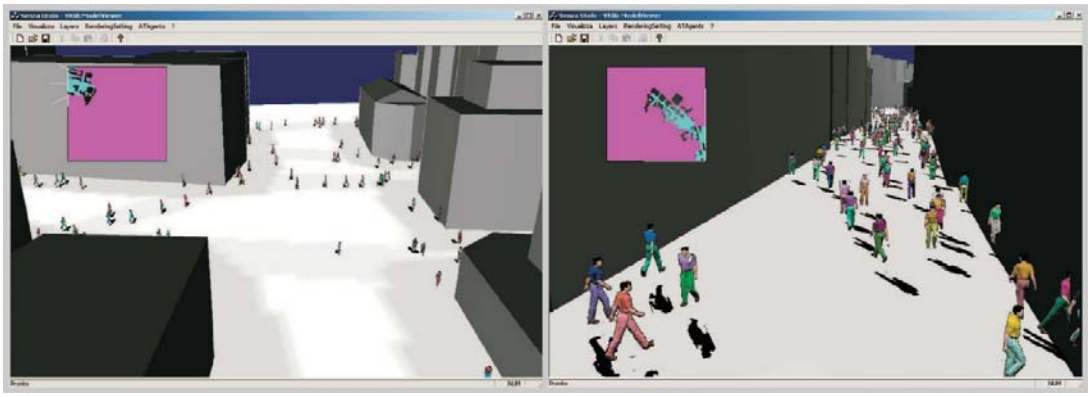

圖 2-7: 7000 個較多行為能力的行人在模擬中即時運算 (Marchal, 2002)

Algorithm 2.1 agent-to-agent collision avoidance

```
For each pair of agents do
    If ( vectors are not collinear ) then
        If ( linear velocities are different ) then
            stop the slower agent
        Else
            choose one agent in a random way
        End if
    Else
        If ( vectors are convergent ) then
            choose one agent in a random way
            change its direction
        Else
            increase the linear velocity of the agent
            which walks ahead
        End if
    End if
End for
```

圖 2-8: 代理人碰撞迴避的演算法 (Marchal, 2002)

在環境中偵測到其他物件或碰撞時行人大部份會產生回應，引發特定的事件行為。行為引發的動機大致上可分為兩種：由事件驅動(event-driven)的隨機行為(random behavior)或既定腳本的行為(scripted behavior)。為了使行人在群眾模擬中展現不同的自主權(Musse, Babski, Capin & Thalmann, 1998)，使用者雖然可以操作行人活動，但只涵蓋一部分的操控權，行人將維持部分基於經驗法則(rule-based)的行為自理，甚至行為上的強弱程度。每個行人都有著可參數化的行為模型，針對不同的對象調整出相異的行為模式，其中也包含亂數產生或模擬狀態中直接修改(Ashida, Lee, Allbeck, Sun, Badler & Metaxas, 2001)。舉例來說，行人在路上行走撞到了牆，於是它停下腳步：這是事件驅動的既定行為；時間經過五分鐘，它開始感覺得疲累了：這是自主的隨機行為。其中也有將行為驅動分割為兩個層次：外部與內部的事件。外部事件是由行人在虛擬環境中所反映出的既定情節劇本，而內部事件則是成員與成員之間對於環境變數所觸發；在行人的外觀表現上，為了產生自然不虛假的行為動作，也利用動作捕捉器(motion capture)作為動作的片斷擷取(如圖示 2-9) (Kim,D, Kim & Shin, 2003)。

圖 2-9: 動作擷取系統 (Vicon Peak, 2005)

事件的引發不僅只是單獨地重複同樣的動作，同時也包含了“下意識行為”(subconscious actions)的可能性。在行人操作的過程中，同時結合“下意識行為”與“有目的的行為”(planned actions)；添加下意識行為不會改變原有的行為內容，而是附加一些關於情緒上的行為表達。舉例來說，一個快樂的行人與悲傷的行人所表現的行為不相同。為了保有原先既定的目的行為，行人將設定一些低階(low-level)的行為參數(Parisy & Schlick, 2002)；假設行人正處於行走的狀態，疲累或悲傷時將調整行走的速度和姿勢(Ashida, Lee, Allbeck, Sun, Badler & Metaxas, 2001)。相關的應用有將人群疏散(evacuation)作為模擬緊急事件，效能增進類似於危機管理(crisis management)的範疇中(如圖示 2-10) (Yohei Murakami, Toru Ishida, Tomoyuki Kawasoe & Reiko Hishiyama, 2003)。

圖 2-10: 代理人悲傷的行走與撤退模擬 (Ashida, Lee, Allbeck, Sun, Badler & Metaxas, 2001) (Yohei Murakami, Toru Ishida, Tomoyuki Kawasoe & Reiko Hishiyama, 2003)

2.3　情緒機制

對於人類來說，我們在環境中必須面對相當多的訊息，包括正面、負面、或難以辨認的事物。在不明確的原因下，我們會使用一種獨特的人格特質來處理：情緒(Ortony, Clore & Collins, 1988)。他們開發了一個電腦的情緒模型(OCC model)，這個模型指定了二十二個情緒類目(emotion categories)，架構在達成目標或事件的誘發性回應，或者是有行為意義的代理人對於物件吸引力數值有反應。他也為變數提供一個架構，像是事件發生的可能性或者對於物件的熟悉度，這些都決定了情緒種類的強度。它包含了足夠的複雜程度來面對大多數以情緒為介面所會面臨到的情況。

OCC 模型已經成為一個情緒綜合體的標準模型，有許多研究使用它來作為他們代理人(agent)產生情緒的模組，而模型的複雜度也將直接影響認知世界的大小。情緒模型必須能夠評估所有代理人可能面臨到的情況，並且提供一個能夠影響情緒強度的變數架構，預定目標與情緒之間的行為緩衝以及建構內部決策的互動，使代理人在正確的時間展現正確的情緒強度，這是展現具說服力情感表達的必要條件(Lee, 1999) (Martinho & Paiva, 1999)。

雖然 OCC 模型在情緒定義上相當詳細，但對於某些程度的決策模型顯得過於複雜，過大的情緒模型會遲緩整體系統的運作效能，只挑出在實作過程中確實會使用的類目便可縮減模型大小；一些類目的關連性太過相近也是有待評估的目標，某些相近的情緒狀態可以合而為一，或是經過參數化的些微調整來達成目標。情緒模型也必須記錄事件的歷史、行為與物件的歷史，歷史紀錄的功能有著重要的優點。根據 OCC 模型，事件發生的可能性需要去計算他的可取值，歷史紀錄的功能有助於計算可能性，並且能追蹤執行的過程，而目前 OCC 模型中沒有提供記憶功能的模組(如圖示 2-11) (Bartneck, 2002)。

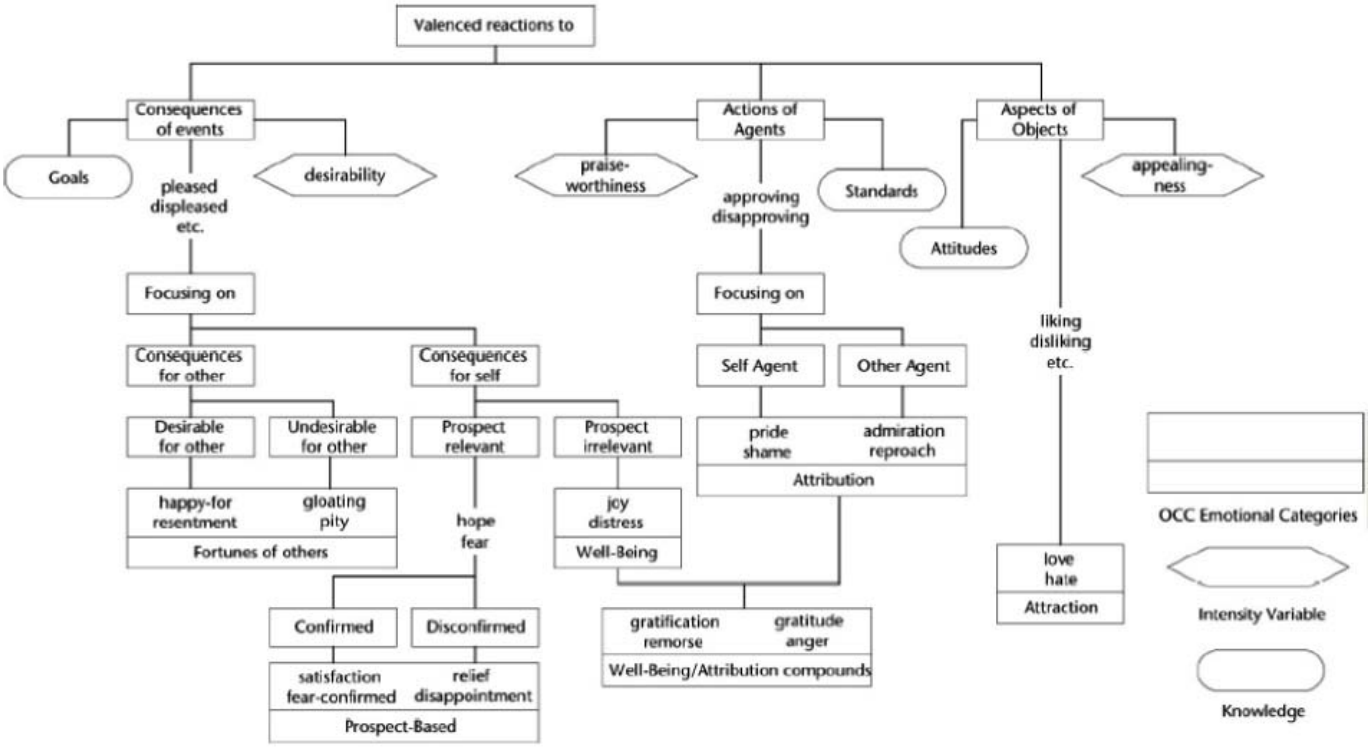

圖 2-11: 經由 Bartneck 整理過的 OCC 模型圖 (Bartneck, 2002)

情緒模組的主要目標是使代理人的決策過程更貼近於真實，甚至發展成更有效率的系統。在真實的世界裡情緒對於生物來說不是經常有助益的，甚至有時會造成負面的回饋(negative feedback)。情緒時常被認為是某種非理性的思考，儘管如此，情緒也能夠被視為一種有效的助力來改變決策選定的過程。相關於情緒產生的動機並不只是改善虛擬環境中代理人行為的真實度，並且實質上改變系統運行的效率(Sandra, 2003)。舉例來說，在危險的環境中，代理人不能在有限的時間內像平常一般考慮所有的可能性，而是必須快速的反應。站在現實的角度，人類需要開發許多方法使得資訊處理的過程更有效率，尤其是當人類依附著這些資訊工程存活的時候。以人類的狀態來說，覺醒是一個測量事務重要與否的標準。人類在緊急的時候會加速呼吸、心跳、分泌腎上腺素、以及腦中的一些化學物質，這些改變緩和了對於環境衝擊的感受而去思考下一步對策。以同樣的方法，當我們回到決策的選定，主題將集中在覺醒如何去衝擊決策系統，衝擊影響造成學習能力或感知能力隨著覺醒的增加而變強，變得更加集中並且去除決策中不必要的雜訊(Chown & Randolph, Jones, Amy & Henninger, 2002)。

情緒模組應用的領域相當廣泛，例如在軍事代理人戰役的模擬(battlefield simulation)，在代理人中添加感情因子與個別差異於行為模組中。從模擬情緒的角度來說這是理想的測試方法，因為代理人必須經過種種的推理(reasoning capabilities)過程，包括情勢評估(situation assessment)、計畫、對計劃失敗做出反應、並與全體的代理人做互動(如圖示 2-12)。以實際狀況來說，代理人偵測到當前目標周圍訊息模糊時，而只有幾個適切的內定決策值能對應於特定的幾項目標，此時無論藉由感知系統接收的資訊是清楚或混淆的，系統只會對於有關係的訊息作反應，其餘不管。舉例來說，命令一位代理人去消滅戰車，首先必須先偵測到戰車的位置。如果代理人找不到戰車的位置，他會增加情緒上的困惑(confusion)值。換句話說，如果代理人的計畫不能預期完成，其中所造成的程序缺口(lacking)對於決策系統不會造成影響(Henninger, Jones & Chown, 2003)。

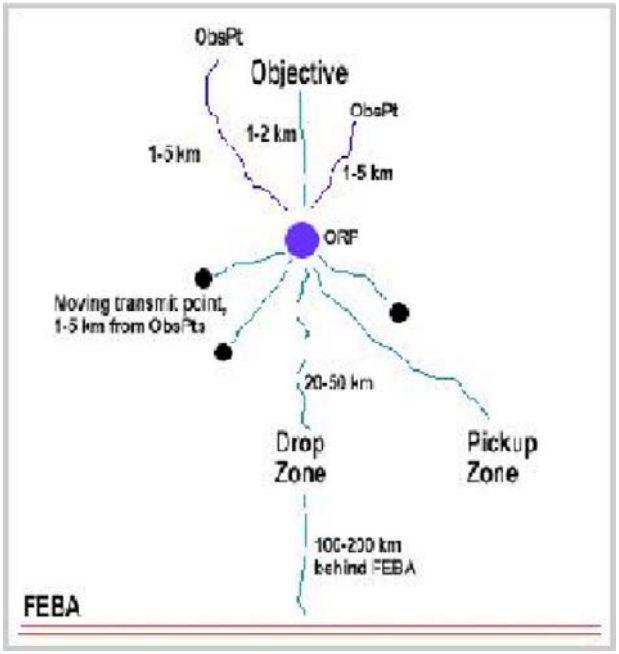

圖 2-12: 在地圖上的長程偵側任務模擬 (Henninger, Jones & Chown, 2003)

2.4　記憶機制

記憶模組的種類分為多種，每一種被設計成不同屬性的資料儲存方式(如圖示 2-13)。短期記憶(short-term memory, STM)被塑造成一種短期的、固定大小的緩衝儲存空間，資料停留在此空間是暫時性地存放，時間一到資料就會消失(Cioffi-Revilla, Paus, Luke, Olds & Thomas, 2004)。而長期記憶(long-term memory, LTM)對資料紀錄(memory traces)來說是較為永久的貯藏處，資料紀錄在長期記憶中不會主動消失，除非經由人為刪除；新的資料紀錄在經過短期記憶處理後傳送於此，決策過程將參考存放在長期記憶中先前經驗與知識完成隨後的行為結果。工作記憶(working memory, WM)是作業與函式執行的地方，當明確的指令或事件發生時訊息就會送至這個地方，訊息的增加或修改將基於短期與長期記憶的內容而定。訊息在短期記憶與長期記憶中處理過之後，將傳送至工作記憶中進行其中一部分的推理過程，來影響左右隨後的明確行為(Peters & Sullivan, 2002)。

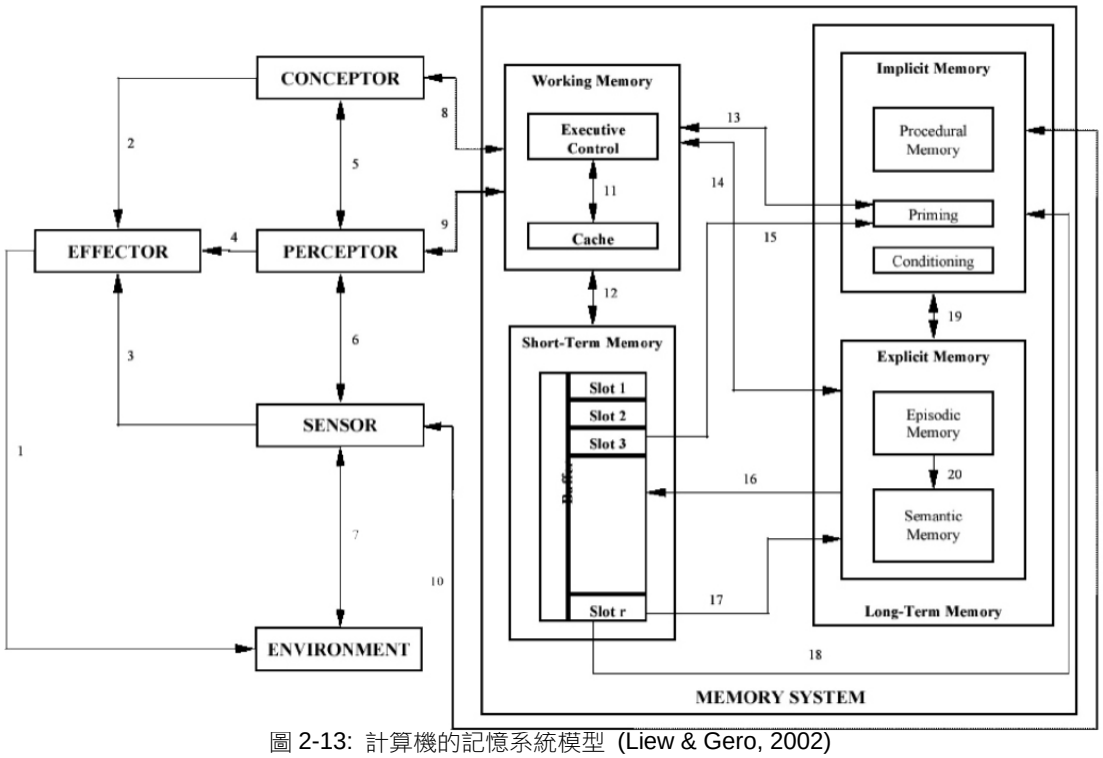

圖 2-13: 計算機的記憶系統模型 (Liew & Gero, 2002)

當代理人在虛擬環境中使用過去的經驗來得到適切的方法，記憶的功能就會被使用到。環境中任何可操作的資訊都被認為是記憶建造過程中所需要的線索，藉由代理人與環境互動的情況將記憶資料集中管理與存取。根據代理人與環境之間的互動，記憶資料將在動態之間建構完成。記憶功能將會促進決策過程的深度與速度，使得感知系統吸收週遭訊息速率提高，並幫助代理人用較充裕的時間適應環境。如果在決策系統添加學習功能，記憶能更加突顯學習功能，甚至可以從過去經驗中得到新的訊息(Liew & Gero, 2002)。代理人必須在虛擬世界中接觸許多事物，每個不同或相同的事件產生對應的行為來處裡。代理人利用記憶機制來評估系統中各項觀測數值的平均值並對照產生行為，結果將產生適應性的新行為(Lerman & Galstyan, 2003)。

2.5　學習機制

以統計學(statistics)觀點來說，電腦學習理論(computational learning theory)是一個相關於機械學習演算分析的數學領域(mathematical field)。機械學習演算法(Machine learning algorithms)涵蓋了一連串的訓練集合(training set)，從假設或是提出模型，然後對於未來提出預測(predictions)。由於電腦的訓練集合是有限的而未來將發生的事件是模糊不清，因此學習理論通常沒有辦法生產出對於演算法效能的絕對保證。以目前相關於工智慧的學習領域中，機械學習主要可以劃分為三大類：監督學習（supervised learning）、回饋學習（reinforcement learning）以及無監督學習（unsupervised learninig）。

監督學習：監督學習是從想要監事訓練(training data)的資料中設置一個機械學習函式，訓練的資料中包含了兩個輸入輸出(input and output) 資料口函式。輸出函式所產生可能是一連串不間斷的數值(regression)，或者是能夠預測輸入物件的等級分類。一般來說，監督學習能夠產生兩種類型的模型，第一種是偵測輸入的物件並在輸出得到所想要的資料；第二種是在輸出時只能得到區域之間的近似值。換句話說，所謂監督學習對於資料流通的過程是全程監視，並有其他函式輔助並告知不足的地方，給予補強。舉例來說，關係就像家教（supervised learning）與小學生(training data)的指導關係，小學生寫作業家教在一旁監督，作業是否有錯別字，段落是否有符合標準都在監測範圍之內(Angluin, 1992)。

回饋學習：回饋學習（Reinforcement Learning）是學習如何去對應，如何規劃解決方案來回應，以便取得最大數值的回饋信號(reward signal)。對於初學者來說不是像大部分機械學習(machine learning)型態一樣告訴它詳細的操作步驟，相反地必須經過自己開發嘗試而得知哪一部分的行為結果會有最多回饋。大部分在最有趣及挑戰性的情況裡，動作行為可能不僅是影響立即回應的結果，也可能影響下一個和後續的處境。嘗試與錯誤的搜尋(trial-and-error search)以及延遲性的回應(delayed reward)是回饋學習最重要和最具區別性的特質。回饋學習定義並不是從學習方法的特性中取得，而是從問題的學習中取得。當有適合的方法來解決問題時，我們認為這就是回饋學習的方法。一般來說，如同上述的學習代理人必須能夠意識到環境的狀況，某些程度上能夠以行為來影響環境，並在環境中擁有一個或多個目標。公式化回饋學習將包含最簡單的三個方向：感知(sensation)、行為(action)與目標(goal) (如圖示 2-14) (Sutton & Barto, 1998)。

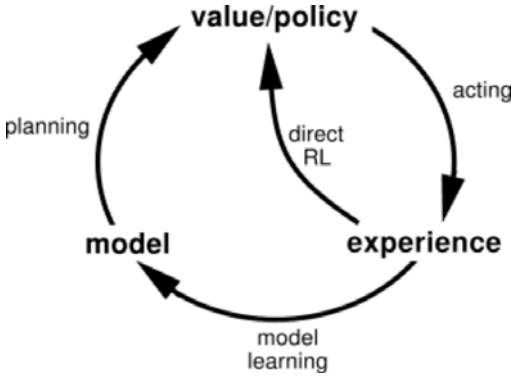

圖 2-14: 代理人學習、計畫、行為的關係圖 (Sutton & Barto, 1998)

監督學習的缺失：不同於回饋學習，監督學習是結合了統計學辨識(statistical pattern recognition)、類神經網路(artificial neural networks)技術的機械學習方法；它從外部知識指導者(knowledgable external supervisor)所提供的範例來學習。這是一種很重要的學習方法，但單獨來說從互動上所得到的學習是不足的。在互動過程中所出現的問題以監督學習的方法解決是不切實際的，當預設正確的行為設定要套用在問題出現的地方可能會出現更糟的情況。

回饋學習深入分析：在 Sutton & Barto 研究中提到(1998)，其中一個在回饋學習運行中可能會遭遇到的問題：探索(exploration)與開發(exploitation)之間的功能交換(trade-off)。為了取得多數的回饋值，一個擁有回饋學習的代理人根據過去取得回饋值的經驗寧願選擇舊有的路線；矛盾地，代理人如果不去開發新的路線，又如何能夠找到多數回饋值的地方。由於能夠得到回饋，代理人必須持續的探索目標，但同時代理人又必須在未來中找到更好的開發結果。進退兩難的地方在於無論是探索或開發都是在工作中確實的執行，代理人需要嘗試各種的行為可能並且漸進地使綜合結果更好。在一個隨機(stochastic)的工作上，每一種行為必須完成許多次數使回饋值得到信賴性的評估。探索與開發的選擇所造成的困境已經在純數學領域(mathematicians)中被研究了數十年，而現在我們只要求兩者之間的關係平衡。

另一個回饋學習的重要特色是當代理人面對不明確(uncertain)的環境時，能夠清楚思考目的的問題概要；在沒有預先設定答案狀態下能考慮到問題的其他分支(subproblems)，是回饋學習與其他學習方法形成對比差異的地方。回饋學習採用的是不一樣的反向觀點，起始於一個完整的、互動的、目標導向(goal-seeking)的代理人。代理人擁有明確的目標，能夠意識到環境，並能夠選擇行為來影響環境。當回饋學習牽涉到決策過程時，它必須明確的處理介於決策與即時行為之間的選項。

舉例來說，一個西洋棋大師選擇了他在棋盤上的一步，這個選擇顯示了兩種思考模式：預測可能的回應與防守；然後迅速直覺地做出特定位置的移動。舉另外一個例子，一個移動式機械人正在做決策，決定是否要進入新房間來搜尋並收集垃圾，或者是開始找尋能源充電站的路線。它將快速並簡單地思考以過去的經驗是否能找到能源充電站，然後再行動。

以上這些範例以簡單基礎的行為使得他們更容易被檢視，所有代理人在決策與環境的互動過程都涵蓋其中，包括代理人在環境中尋找並實現目標。代理人的行為是被允許來影響未來環境的狀態，比如像是下一個棋步或機械人所到達的下個目的地，因此影響了代理人接下來擁有的機會與選項。正確的選擇必須把間接性、延遲性的後續動作納入思考範圍中，然而這可能需要決策的先見之明。同時，這些範例行為的後續影響不能夠被完全預測，因此代理人需要繁複地監測環境以產生合適的行為(如圖示 2-15)。在範例中主角除了要達成當前的目標之外，也相當明確的了解自身以及週遭環境的狀態，比如說西洋棋大師知道無論如何他一定會贏得勝利，或者是移動機械人知道自己的電池何時會耗盡…等等。

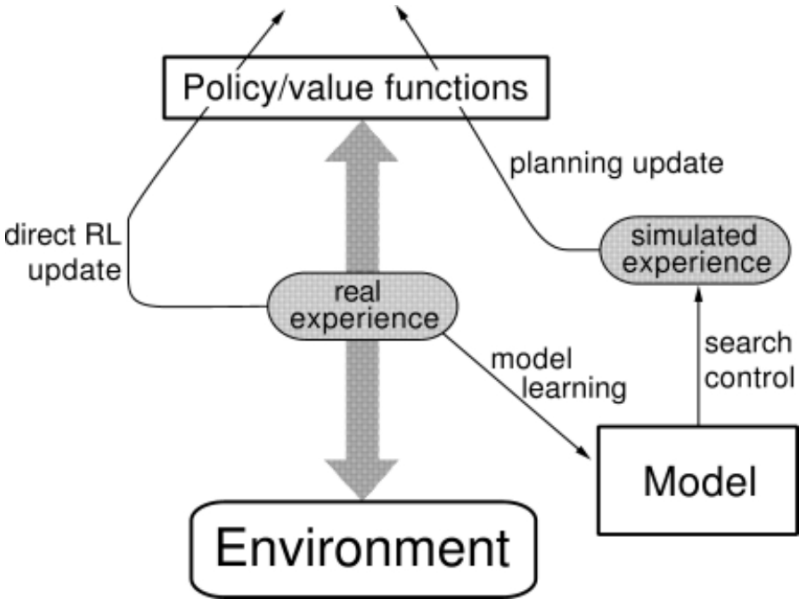

圖 2-15: 代理人與環境的基本關係圖 (Sutton & Barto, 1998)

在上述範例中代理人在環境一段時間後便知道如何利用過去經驗使得整體效能提升；西洋棋大師提升對於棋步的直覺觀察評估，使得他西洋棋玩得更好；代理人把所需具備的知識在工作開始前就準備好，從過去相關工作的歷史或者是決策時的新進展得知，什麼是有效的選擇與學習經驗，但代理人藉由與環境上互動所得到的新行為仍然是工作需求中學習的本質。