《國立政治大學哲學學報》 第二十一期 (2009 年 1 月) 頁 35-84

哥德爾的不完備性定理與心靈是否為機器的論爭

蔡行健

國立中正大學哲學系

摘要

哥德爾的不完備性定理是現代邏輯發展過程中所發現的最重要的獨立性結果。在晚近的文獻中常可看到有些學者試圖利用不完備性定理來證明不可能有能夠完整地模擬人類心靈的機器存在，持此立場之較有代表性的學者有盧卡斯 (Lucas)、潘若斯 (Penrose) 以及麥扣 (McCall) 等。這些學者個別的主張雖不盡相同，但都認為: 對任何機器而言，根據不完備性定理，都會有一些數學命題是它所無法證明的，但人類心靈卻可得知其為真。而持反對立場的學者，如弗蘭森 (Franzén)、林德斯仲 (Lindström)、夏皮洛 (Shapiro) 及蓋夫曼 (Gaifman) 等人則分別指出：上述學者的論證不足以保證所宣稱的心靈相對於機器的優勢。在本文中，筆者將審視關於這個議題的主要的正反論證，並釐清牽涉於論爭的幾個重要概念，如「機器」、「證明」以及「一致」等。筆者將指出：限制在不完備性定理的脈絡下，心靈是否為機器這個問題不可能有數學上或邏輯上的明確答案，然而就哲學的觀點而言，機器論者必須承擔較大的舉證責任。

關鍵詞：哥德爾、不完備性、心靈、機器、證明、一致

哥德爾的不完備性定理與心靈是否為機器的論爭

壹、哥德爾的兩個不完備性定理及相關定理簡介

一般而言，哥德爾的不完備性定理是關於算術理論的後設邏輯定理 (meta-logical results of number theory)，但我們稍後會說明如何將這些結果延伸應用到其他的邏輯理論。考慮一個公理化的第一階邏輯理論 T[1]，其語言 L 的意欲的詮釋 (intended interpretation)是在自然數上。通常該語言會有常數符號 "0"，二元述詞 "<"，一元函數 "s" (successor)，二元函數 "+"、"×" 及 "E" (exponentiation)。這些符號都是我們所熟悉的，而它們在自然數上的詮釋也不言可喻。由於任何一個該語言的句式 (formula) 都只包含有限多的符號，如果我們將語言的符號對應到自然數上（不同的符號當然對應到不同的自然數），則透過某種編碼方式，任一個句式也都會對應到一個自然數。同樣地，一個由有限多的句式所構成的序列也可以被對應到某個自然數。一般把此類型的編碼稱之爲哥德爾編碼

[1] 這裏的「公理化的理論」指的是「遞迴公理化的理論」(recursively axiomatized theory)，也就是這些公理的哥德爾碼（見下文）所形成的集合是遞迴的，或是可被決定的 (decidable)。簡單而言，這表示給與任何語句（使用所考慮的理論的語言），我們有一個有效的程序來決定這語句是否為公理之一。

(Gödel numbering)。此編碼的目的在於讓這語言能夠提及自身所使用的符號及使用這些符號所形成的句式，這也使得一些語法上的項目 (syntactic items) 能被語言本身所定義。對理論 T 也有一個基本要求：

(i) T 最少要能證明所有在自然數爲真的不等式 $\underline{m}\neq\underline{n}$，及爲真的等式 $\underline{m}+\underline{n}=\underline{k}$ 與 $\underline{m}\times\underline{n}=\underline{k}$，以及 $x<\underline{n}\rightarrow(x=\underline{0}\vee\ldots\vee x=\underline{n-1})$，$x<\underline{n}\vee x=\underline{n}\vee\underline{n}<x$ (Gaifman, 2000: 464)。[2] 在此$\underline{m}$（自然數m的名字）實際上是 s(s(…(s($\underline{0}$)…)，其中s出現了m次。例如，s($\underline{0}$)指稱 1，s(s($\underline{0}$)) 指稱 2，……，依此類推。

假設 ϕ(x) 這個句式正好只有x這個自由變數 (free variable)。假設 G(ϕ(x)) 是透過哥德爾編碼之後 ϕ(x) 所對應的自然數，如果這個自然數是m，那麼它的名字爲$\underline{m}$（如前所述，它實際上是s(s(…(s(0)…)，其中s出現了 m次），ϕ($\underline{m}$) 也是一個合法的句式 ，一般稱之爲 ϕ(x) 的對角化 (diagonalization)，它所對應的哥德爾碼爲 G(ϕ($\underline{m}$))。我們可以定義一個在自然數上的函數 D，使得對所有正好具有一個自由變數的句式 ϕ(x)，D(G(ϕ(x)))= G(ϕ($\underline{m}$))，其中 m= G(ϕ(x))。假如 ϕ(x) 可被 T 證明，因爲一個證明 S 是由一些句式所構成的有限序列 [3] ，所以也可經由編碼對應到某個自然數

---

[2] 或者 T 最少須是算術理論的一個可被有限公理化的子理論，見 Enderton, 2001: 203。

[3] 一個證明是由句式構成的有限序列，其中的每一個句式如果不是邏輯公理，就是 T 的公理，要不然就是由它之前的某兩個句式透過「肯定前項」(modus ponens) 所得到的。這個定義見 Enderton, 2001: 111。若採取的系統不同，則定義可能會有出入，但無論如何定義，證明仍會是一個有限序列，而其中的每一項都能被有效地決定。

G(S)。因此我們可以在自然數上定義一個二元關係 $Proof_T$，使得對所有可被 T 證明的句式 ϕ，$Proof_T(G(S), G(\phi))$，其中 S 是從 T 所得到的對 ϕ 的證明，而 G(S) 是這個證明所對應的哥德爾碼。如果 T 是一個公理化的理論而且滿足前面所提到的基本要求，則我們可以找到兩個句式 $\delta(x, y)$ 與 $\theta(x, y)$ 分別定義 D 與 $Proof_T$，使得：

(1) 對所有正好有一個自由變數的句式 $\phi(x)$，$T \vdash \forall y(\delta(\underline{m}, y) \leftrightarrow y=\underline{n})$，其中 $m= G(\phi(x))$，$n= D(G(\phi(x)))$；

(2) $<m, n> \in Proof_T$則 $T \vdash \theta(\underline{m}, \underline{n})$；若$<m, n> \notin Proof_T$則 $T \vdash \neg\theta(\underline{m}, \underline{n})$。

利用 (1)，我們可以證明所謂的「固定點定理」(fixed point theorem)：對所有正好有一個自由變數的句式 $\phi(x)$，存在一個語句 α，使得 $T \vdash \alpha \leftrightarrow \phi(\underline{m})$，其中 $m=G(\alpha)$[4]（底下將直接把 $\underline{m}$ 寫成 $\underline{G(\alpha)}$）。直覺上，α 說「我滿足 ϕ 這個性質」。現在考慮 $\neg\exists x\theta(x, y)$，直覺上，這個句式說 y 無法從 T 來證明。利用固定點定理，有一個語句 α，$T \vdash \alpha \leftrightarrow \neg\exists x\theta(x, \underline{G(\alpha)})$，這個語句一般被稱作 T 的

[4] 證明如下：考慮 $\forall y(\delta(x, y) \rightarrow \phi(y))$ 這個句式。它說 x 的所有對角化都會滿足 ϕ，這裏的 x 當然可以是此句式本身的哥德爾碼。這提示我們把這個句式對角化，得到 $\forall y(\delta(\underline{m}, y) \rightarrow \phi(y))$，其中 $m=G(\forall y(\delta(x, y) \rightarrow \phi(y)))$。而這正是我們所需要的語句 α。根據 (1)，$T \vdash \forall y(\delta(\underline{m}, y) \rightarrow \phi(y)) \leftrightarrow \forall y(y=\underline{n} \rightarrow \phi(y))$，其中 $n=G(\forall y(\delta(\underline{m}, y) \rightarrow \phi(y)))$。明顯的，$T \vdash \forall y(y=\underline{n} \rightarrow \phi(y)) \leftrightarrow \phi(\underline{n})$。故 $T \vdash \forall y(\delta(\underline{m}, y) \rightarrow \phi(y)) \leftrightarrow \phi(\underline{n})$。事實上，哥德爾並沒有證明固定點定理，而是卡納普 (Carnap) 分析哥德爾的證明之後所提出的。見Carnap, 1937: 129-31。

「哥德爾語句」(Gödel sentence)。如果 $T \vdash \alpha$，則 $T \vdash \neg\exists x\theta(x, \underline{G(\alpha)})$。可是 $T \vdash \alpha$ 表示有個由句式構成的有限序列 S 是 α 的證明，因此 $<G(S), G(\alpha)> \in Proof_T$，根據 (2)，$T \vdash \theta(\underline{G(S)}, \underline{G(\alpha)})$，因此 $T \vdash \exists x\theta(x, \underline{G(\alpha)})$，T 會是一個不一致的 (inconsistent) 理論。若 $T \vdash \neg\alpha$，則 $T \vdash \exists x\theta(x, \underline{G(\alpha)})$，如果 T 是一致的，則 T 不能證明 α，換句話說，對所有的自然數 m，$<m, G(\alpha)> \notin Proof_T$，根據 (2)，$T \vdash \neg\theta(\underline{m}, \underline{G(\alpha)})$。所以 T 是「ω-不一致的」(ω-inconsistent)。[5] 這就是哥德爾的第一個不完備性定理。

很明顯的，如果 T 能夠被算術理論的意欲的模型 (intended model)，亦即自然數的算術結構 $<\omega, 0, s, <, +, \times, E>$，所滿足，則 T 會是 ω-一致的。而通常我們所考慮的是這類型的理論，所以，若 T 是一致的，則最少會有一個語句，也就是 T 的哥德爾語句，使得 T 無法證明它也無法證明它的否定。

哥德爾的第一個不完備性定理蘊涵以下命題：若 T 是一致的，則 T 不能證明 T 的哥德爾語句。哥德爾的第二個不完備性定理則是透過證明這個命題的形式化所得到的。但這裏所考慮的理論 T 除了必須滿足前面所指出的基本要求外，還要滿足以下幾點：

(ii) 對所有的句式 α，若 $T \vdash \alpha$ 則 $T \vdash \exists x\theta(x, \underline{G(\alpha)})$。

(iii) 對所有的句式 α，$T \vdash \exists x\theta(x, \underline{G(\alpha)}) \rightarrow \exists x\theta(x, \underline{n})$，其中 $n=G(\exists x\theta(x, \underline{G(\alpha)}))$。也就是 T 可以證明「若 α 是可證明的則『α 是可證明的』是可證明的」。這是 (ii) 的形式化。

---

[5] T 是 ω- 不一致的若且唯若存在一個語句 $\exists x\phi(x)$ 使得 $T \vdash \exists x\phi(x)$，但對所有的自然數 n，$T \vdash \neg\phi(\underline{n})$。

(iv) 對所有的句式 α 及 β，T ⊢ ∃xθ(x, $\underline{G(\alpha)}$)→(∃xθ(x, $\underline{G(\alpha\rightarrow\beta)}$)→∃xθ(x, $\underline{G(\beta)}$))。這是肯定前項 (modus ponens) 的形式化。

「T是一致的」這個語句可被形式化爲 ¬∃xθ(x, $\underline{G(y\neq y)}$)，也就是 T 無法證明 y≠y。把這個語句稱之爲 CONT。而「T不能證明T的哥德爾語句」的形式化是 ¬∃xθ(x, $\underline{G(\alpha)}$)，也就是 α。所以「若 T 是一致的，則T不能證明T的哥德爾語句」的形式化即是 "CONT→α"。在此我們可以證明 T ⊢CONT→α。[6] 因此，如果 T 真的能證明它本身的一致性，亦即 T ⊢CONT，則T ⊢α，可是如此一來，根據第一個不完備性定理，T會是不一致的。換言之，T 如果是一致的，則T不能證明它自身是一致的。這就是哥德爾的第二個不完備性定理。

雖然這裏的 L 是算術理論的語言，其詮釋的定義域是自然數的集合。但其他語言的理論也有可能得到不完備性，只要該語言的語法上的項目能被形式化，且所考慮的理論必須滿足某些要求，使

---

[6] 證明如下：為簡化起見，將 ∃xθ(x, y) 寫為 p(y)。我們已知道 T ⊢α↔¬p($\underline{G(\alpha)}$)。故 T ⊢α→¬p(($\underline{G(\alpha)}$)。根據 (ii)，T ⊢p(G($\underline{\alpha\rightarrow\neg p(G(\alpha))}$))，再根據 (iv)，T ⊢p($\underline{G(\alpha)}$)→p(G($\underline{\neg p(G(\alpha))}$))。根據 (iii)，T ⊢p($\underline{G(\alpha)}$)→p($\underline{G(p(G(\alpha)))}$)，因此，(*) T ⊢p($\underline{G(\alpha)}$)→(p($\underline{G(\neg p(G(\alpha)))}$)∧p($\underline{G(p(G(\alpha)))}$)。但 T ⊢(p($\underline{G(\alpha)}$)∧¬p($\underline{G(\alpha)}$))→y≠y，故根據 (ii) 再根據 (iv)，T ⊢(p($\underline{G(\neg p(G(\alpha)))}$)∧p($\underline{G(p(G(\alpha)))}$)→p($\underline{G(y\neq y)}$))，亦即 T ⊢(p($\underline{G(\neg p(G(\alpha)))}$)∧p($\underline{G(p(G(\alpha)))}$)→¬CONT，根據前面的 (*)，得到 T ⊢p($\underline{G(\alpha)}$)→¬CONT，反轉即得 T ⊢CONT→¬p($\underline{G(\alpha)}$)，亦即 T ⊢CONT→α。此外，我們很容易看出 T ⊢¬CONT→¬α，因為 T ⊢y≠y→α，接著根據 (ii)，T ⊢p($\underline{G(y\neq y\rightarrow\alpha)}$)，然後再根據，(iv) T ⊢p($\underline{G(y\neq y)}$)→p($\underline{G(\alpha)}$)。因此，T ⊢α↔CONT。

得上述的 (i)~(iv)（或它們在所考慮的語言上的等價陳述）成立，則這理論也會有不完備性的結果。根據這樣的觀察，第二個不完備性定理有很廣泛的應用，現今已成爲後設邏輯研究的一個重要的基本工具：一個理論只要滿足某些要求，就不能證明自己的一致性，否則它就會是不一致的。重要的例子如皮亞諾算術（Peano arithmetic，以下簡稱 PA），或一些集合論系統如 ZF 或 ZFC （下文中所提到的集合論若沒有特別註明，都是指 ZFC）。

讓我們考慮任何一個形式化的邏輯理論，若某個在該理論的語言上的語句及其否定都不能被它證明，則我們稱該語句獨立於此理論之外。當一個語句可被證明是獨立於某個理論之外，我們說這是一個關於此理論的「獨立性結果」(independence result)。哥德爾的第一個不完備性定理就是關於算術理論的獨立性結果，但如同前面所提到的，它也是關於任何滿足某些條件的理論的獨立性結果。一個重要的觀察是：證明一個語句獨立於某個一致的理論之外，可以保證加入該語句到此理論之後可以得到一個更強的一致的理論。所以，若以 PA 爲出發點，我們可以構築一系列強度嚴格遞增的公理化的理論，後一個理論即是由前一個理論加上其哥德爾語句所構成。

由於討論的需要，接下來要介紹塔斯基的不可定義性定理 (Tarski's Undefinability Theorem)，這是一個與哥德爾不完備性定理有密切關聯的重要定理。前文已提到我們通常會把自然數的算術結構記爲 $<\omega, 0, s, <, +, \times, E>$，而在這個結構爲真的語句所形成的集合是一個邏輯理論（我們將它稱爲N），也就是所有自然數算術的真理所形成的理論。讓 #N 代表 N 的每一個語句的哥德爾碼所形成的集合，塔斯基的定理指出：#N 不能被定義在自然數的算術結構上。#N 當然是自然數的子集合。而一個自然數的子集合 S 能被定義在自然數的算術結構上若且唯若有一個帶有一個自由變數的

句式 $\alpha(x)$，使得對任何的自然數 n，$n \in S$ 若且唯若 $\alpha(\underline{n})$ 在 $<\omega, 0, s, <, +, \times, E>$ 爲真（亦即 $\alpha(\underline{n})$ 是一個自然數算術的真理）。利用固定點定理可以簡單地證明塔斯基的定理如下：假設 #N 被某個 $\alpha(x)$ 所定義。根據固定點定理，存在某個語句 δ 使得 $\delta \leftrightarrow \neg\alpha(\underline{G(\delta)})$ 是一個自然數算術的真理。如果 δ 是一個自然數算術的真理，$\neg\alpha(\underline{G(\delta)})$ 也會是一個自然數算術的真理，如此一來，α 不可能定義 #N；若 δ 是一個假的語句，則 $\neg\alpha(\underline{G(\delta)})$ 也是一個假的語句，因此 $\alpha(\underline{G(\delta)})$ 是一個自然數算術的真理，但這表示 δ 必須是一個自然數算術的真理，所以我們又得到了矛盾。

由於任何可被公理化的N的子理論，其所有語句的哥德爾碼所形成的集合都可被定義在自然數的算術結構上，[7] 所以 N 不可能被公理化。

貳、數學的不可窮盡性與哥德爾的選言命題

哥德爾自己曾說他的第二個不完備性定理指出了數學的不可窮盡性 (incompletability or inexhaustibility)：

> 因為它使得任何人都不可能設立一個由公理及法則所構成的明確定義系統並同時能一致地做出下列關於此系統

[7] 任何一個公理化的理論，其所有語句的哥德爾碼所成的集合必然是遞迴可枚舉的 (recursively enumerable)，而任何遞迴可枚舉的自然數的子集合必然可被一個 $\exists_1$-句式所定義。這個定理在大部分的數學邏輯教本都可找到。可參 Enderton, 2001: 238-45。

的宣稱：我知道（以數學的確信度）所有這些公理及法則都是正確的，除此之外，我也知道它們涵蓋了所有的數學。如果某人做出這樣的陳述，他就自相矛盾。因為若他知道這些被考慮的公理是正確的，他也知道（以同樣的確信度）它們是一致的。因此他有一個數學的洞見是這些公理所無法證明的。然而，為了要清楚地理解剛剛說的事，我們必須要小心。這表示沒有任何由正確的公理所構成的明確定義系統可以涵蓋整個數學本身 (mathematics proper) 嗎？是的，如果數學本身被了解為所有真的數學命題所構成的系統；但答案會是否定的，如果數學本身是指所有能被證明的數學命題所形成的系統。我將區分這兩種意義的數學為客觀意義的數學 (mathematics in the objective sense) 與主觀意義的數學 (mathematics in the subjective sense)。很明顯的，沒有任何由正確的公理所構成的明確定義系統可以構成整個客觀數學（筆者按：即客觀意義的數學，而下文的主觀數學即主觀意義的數學），因為有某個為真的命題，它陳述此系統是一致的，但卻是該系統無法證明的。不過，對主觀數學而言，我們不能排除可能有某個有限的法則產生它所有的明顯的公理。然而，若這法則存在，我們以人類的理解能力永遠不可能知道它是這樣的法則，亦即，我們永遠無法以數學的確信度來知道它所產生的所有命題都是正確的。……然而，根據足夠數量的例子或是其他的歸納推論，我們頂多能以經驗的確信度 (empirical certainty) 來認知「這些命題都是真的」這樣的斷言。若果真如此（筆者按：應指上述的有限法則確實存在），這表示人類心靈（在純粹的數學領域中）等價於一個有限的機器，而這機器無法完整地理解它自身如何運作……因此底下的選言結論是無法避免的：……人類心靈（甚至僅限制在純粹數學的領域）無窮地超越任何有限機器的能力或存在一些絕對無解的特定形態的迪奧番

> 庭問題 (absolutely unsolvable Diophantine problems of the type specified) (Gödel. 1995: 309-10)。[8]

在此須解釋一下在引文中哥德爾所要表達的意思。第二個不完備性定理指出了沒有任何一致的公理化系統可以定義出整個數學，因爲該系統的哥德爾語句（或該系統的一致性；因二者是等價的），是一個爲真的數學命題，而該系統卻無法予以證明。這顯示了客觀的數學不能被任何一致的公理化系統所窮盡。但主觀數學是吾人心靈透過證明所能得到的數學，它有可能被某個有限公理化系統所窮盡，但我們就算面對這個公理化系統也不可能以數學的確信度知道它窮盡了主觀數學，否則我們就能以數學的確信度知道它是

---

[8] 這裏所稱的不可窮盡性是原文的 “incompletability”，而弗蘭森做 “inexhaustibility”，見 Franzén, 2005: 112。筆者認為後者比較不會與 “incompleteness” 混淆，但在此都列出來。又，在引文最後哥德爾所說的特定形態的迪奧番庭問題是指：如果 $P(x_1, \ldots, x_n, y_1, \ldots, y_m)$ 是一個整數係數多項式 (a polynomial with integral coefficients)，則對任意 m 個整數 $a_1, \ldots, a_m$，$P(x_1, \ldots, x_n, a_1, \ldots, a_m)=0$ 都有整數解，或者存在某 m 個整數 $a_1, \ldots, a_m$，使得 $P(x_1, \ldots, x_n, a_1, \ldots, a_m)=0$ 沒有整數解 (Gödel, 1995: 307)？ 哥德爾自己也指出：任何滿足一定條件的公理化系統其哥德爾語句會等價於某個以下形態的語句： $\forall y_1 \ldots \forall y_m \exists x_1 \ldots \exists x_n \; P(x_1, \ldots, x_n, y_1, \ldots, y_m)=0$，其中 $P(x_1, \ldots, x_n, y_1, \ldots, y_m)$ 是一個整數係數多項式 (Davis, 2004: 41-73)。假設人類心靈是一部有限機器，則我們可以將之視為一個公理化系統（見本文第四節對「機器」的說明）。而這系統會滿足哥德爾所提到的條件（他本人顯然認為如此），因此根據第二個不完備性定理，這系統不能證明其自身的哥德爾語句，亦即對某個整數係數多項式 $P(x_1, \ldots, x_n, y_1, \ldots, y_m)$，這系統無法有效決定 $\forall y_1 \ldots \forall y_m \exists x_1 \ldots \exists x_n \; P(x_1, \ldots, x_n, y_1, \ldots, y_m)=0$ 是否成立，所以會有絕對無解的迪奧番庭問題存在。

一致的，可是這件事是該系統所無法得知的，如此一來它就不算是窮盡整個主觀數學。如果這樣一個有限公理化系統確實存在，則人類的數學能力可被一部有限機器完整模擬，但人類永遠無法完全知道（哥德爾指的是具備數學確信度的知）這機器的運作方式，因爲哥德爾認爲對其運作方式完全了解就蘊涵了知道它是一致的。哥德爾的選言結論是從「任何有限機器都不可能完整模擬人類心靈或有一部有限機器可以完整模擬人類心靈」所得到的，其中的第二個選言項蘊涵了「有一些關於自然數的數學問題是人類心靈所絕對無法解決的」。布洛斯 (Boolos) 指出哥德爾這裏的推論有個跳躍：我們無法從「對任何有限機器都有一些關於自然數的數學問題是它無法證明的」得到「如果有一部有限機器可以完整模擬人類心靈，有一些關於自然數的數學問題是人類心靈所絕對無法解決的」，這是因爲「人的心靈可以被一部有限機器完整模擬」或「人的心靈是一部有限機器」這些命題的意義並不清楚，而機器如何模擬或代表 (represents) 心靈尚待解釋。但布洛斯承認，如果「某一」心靈 (a mind) 能證明的數學命題正好是某一機器能證明的，則根據不完備性定理，會有某個數學命題是該心靈所無法證明的 (Gödel, 1995: 293)。確實如同布洛斯所說的，所謂的機器能「模擬」或「代表」人類心靈的說法並不清楚。很明顯的，這裏所考慮的不是某個人的心靈，也不是受到記憶力、 時間所限制的心靈，而是「理想化」的心靈。如果我們不認爲論及這樣的心靈是有意義的，則就不必認真考慮整個由哥德爾不完備性定理所引發的論爭。另外，什麼是「機器」仍有待釐清，在下文會再討論這個問題。

哥德爾自己認爲前述的選言結論的第二個選言項不能成立。根據王浩所載，哥德爾認爲理性若問一些無法回答的問題但又斷言只有理性能夠回答這些問題，則理性會是非理性的 (…by asking un-answerable questions while asserting that only reason can answer them，reason would be irrational) (Wang, 1974: 324-6)。依布洛斯的

見解，這是從康德而來 (Gödel, 1995: 294)。哥德爾也提供其他的論證來支持第一個選言項，這出自他對圖靈 (Turing) 的批評。圖靈預設計算 (computation) 只運用有限的原始符號（single symbols；指的是不由其它符號所組成的符號），其理由有二：首先，如果我們使用無限多的原始符號，則會有一些原始符號它們之間的差異會無限地縮小 (arbitrarily small) 而難以辨視；再者，運用有限的原始符號仍可構成無限多的符號，這是因爲一個由既予的原始符號所構成的序列可視爲一個複合的符號。基於類似的理由，圖靈預設心靈在從事運算時的狀態 (mind states) 也是有限的，否則有些狀態將任意地逼近 (arbitrarily close) 以致於會引起混淆 (Davis, 2004: 135-6)。但如果心靈在運算時所用的符號及狀態都是有限的，其能力就不會超越機器。[9] 哥德爾指出圖靈的論證犯了一個「哲學上的錯誤」，亦即他忽略了一個關於心靈的事實：心靈在使用時並非靜態的而是持續不停地在發展，因此就算在每一個發展階段我們能使用的抽象詞語是有限的，但就整個發展而言，它會趨近於無限 (Gödel, 1990: 306)。哥德爾似乎假設了：精確分辨無限多的抽象詞語需要無限多的心靈狀態。可是這裏看不出來哥德爾所謂的心靈狀態與圖靈所指的是否相同。能確定的是，兩者各訴諸不同的直覺來論證，而哥德爾不見得有優勢，因爲一個僅具備有限狀態的機器也有可能學習及發展（如果答案如此簡單就是否定的，也不會有這麼多人投入人工

[9] 圖靈所謂的「心靈狀態」應該是對應於「圖靈機器狀態」(states of a Turing machine) 的概念，因為他說：「一部電腦在任何一刻的行為是由他所觀察到的符號及他在當時的『心靈狀態』所決定的」(Davis, 2004: 136)。他在這裏已經直接把人的心靈用機器來模擬分析。關於圖靈機器的定義在大多數的數學邏輯教本都可找到，可參 Monk, 1976: 14-25 或 Soare, 1987: 11-13。

智慧的研究了）。但無論如何，在相關概念缺乏清楚定義的情況下，很難評估哥德爾的批評有何效用。

參、利用不完備性定理來「證明」心靈不是機器

一、盧卡斯的論證以及相關的回應

訴諸哥德爾的第二個不完備性定理，盧卡斯 (Lucas) 在 1961 年提出的主要論證的結構相當簡單：不論多麼複雜的機器，只要它是一致的，就無法證明自身之哥德爾語句，但我們人類卻知道該語句爲真。一些學者，如弗蘭森 (Franzén)，認爲盧卡斯主張人類心靈比任何機器都優越。實際上盧卡斯並沒有做出這麼強的結論，他承認有些（一致的）機器，在許多方面，其證明數學真理的能力遠優於人類的心靈，但一部一致的機器不論再如何複雜，它至少有一個盲點，也就是它的哥德爾語句，可是人類心靈卻知道這語句爲真，因而沒有任何機器能完整地模擬人類的心靈 (Lucas, 1961: 115-8)。但如同弗蘭森所指出的，哥德爾的第二個不完備性定理僅告訴我們：如果我們知道某個機器是一致的，則我們也會知道它的哥德爾語句爲真；不過並沒有保證我們一定可以得知一部一致的機器是一致的，因此我們也不見得知道它的哥德爾語句爲真 (Franzén, 2005: 55)。既然哥德爾的第二個不完備性定理並不蘊涵「不論什麼機器，只要它是一致的，我們都可知道它的哥德爾語句爲真」這件事，它也不能保證盧卡斯的結論。而且，機器也知道「如果我知道我是一致的，則我也知道我的哥德爾語句爲真」。[10] 所以與其哥德

[10] 這裏的機器要滿足在介紹哥德爾的不完備性定理時所提到的條件 (i)~(iv)，而機

爾語句有關的事，機器也不會知道得比較少。

可以支持盧卡斯的論證的一個例子是：PA 不能證明它本身是一致的，但我們卻知道它是一致的，因此也知道 PA 的哥德爾語句爲真。但這裏所要面對的不僅是 PA 而已，而是所有的滿足第一節所提到的 (i) 到 (iv) 的系統，就算所考慮的系統是一致的，也不保證我們能知道它是一致的。不完備性定理並未告訴我們如何得知一個一致的系統是一致的。事實上，就算我們能夠知道某個一致的系統是一致的，這通常必須訴諸不完備性定理之外的數學事實。例如，我們之所以知道 PA 是一致的，是因爲我們知道 PA 的任何公理都在自然數的算術結構上爲真（在底下的章節將再討論這一點）。

盧卡斯 (1996) 晚近對上述的批評做出底下的回應。他認爲主張機器可完整模擬人類心靈的人（即盧卡斯所謂的「機器論者」），除了應該要描述這類型的機器的構造之外，也要先確認它是一致的，否則這主張就不值得認真對待。但若持這項主張者保證機器是一致的，則我們就知道這機器是一致的，可是此機器本身不可能知道（即證明）這件事。在這情況下，「機器的一致性並非由心靈的數學能力所建立，而是依靠機器論者的證詞而得到的」(Lucas, 1996: 117)。弗蘭森及夏皮洛指出 (Franzén, 2005: 118-9) (Shapiro, 2003:

器「知道」某個語句表示機器能證明該語句。稱我們所考慮的機器為 T。在證明哥德爾的第二個不完備性定理時，我們證明了 $T \vdash CONT \rightarrow \alpha$，$\alpha$ 是 T 的哥德爾語句。根據 (ii) 及 (iv)，$T \vdash p(\underline{CONT}) \rightarrow p(\underline{\alpha})$，p 的定義見註解 6。又，我們把機器視為邏輯理論，是因為每一機器都可用相對應的程式來表示，而每一個程式又可被視為一個邏輯理論。關於這一點，在第四節中會再說明。

25-26)：我們可以接受盧卡斯對機器論者的要求，但依靠他人的證詞而知道某機器是一致的，談不上是「證明」該機器是一致的，也不能夠以「數學的確信度」(mathematical certainty) 來接受這個命題。但是筆者認爲盧卡斯的回應仍有值得玩味之處：先不管我們是否能依靠他人的證詞來確立某機器的一致性，如果機器論者聲稱他知道該機器是一致的，他本身就會成爲機器可完整模擬人類心靈的反例，因爲機器不可能知道這件事；可是他也不能說自己也不知道該機器是否爲一致，否則我們就不會認真考慮他的主張。這個兩難的困境似乎是機器論者所難以避免的。

值得注意的是，盧卡斯與機器論者之間仍容得下第三種主張，也就是機器「可能」模擬人類心靈。盧卡斯認爲機器「一定不可能」完整模擬人類心靈（起碼文獻中是這樣解讀他的主張），而機器論者認爲有機器可以完整模擬人類心靈，這不僅是指出可能性而已，而是主張實際上可以如此。筆者在上文中已指出盧卡斯的回應是機器論者難以應對的，但這並不表示盧卡斯本身的主張成立，因爲機器完整模擬人類心靈的可能性並未消除：我們可以想像有個機器確實能完整模擬人類的心靈，但人類不可能知道這件事。

二、潘若斯的論證以及相關的回應

潘若斯 (Penrose) 在 1989 年的著作《國王的新心靈》(The Emperor's New Mind)中，從許多不同的觀點來論證機器不可能完整模擬人類的心靈，並預測未來的物理學，特別是量子物理，可以用來處理所謂的「意識的科學」(science of consciousness)，幫助我們了解人類的心靈如何作用。潘若斯的論證牽扯廣泛，但確實也訴諸哥德爾的不完備性定理。他知道盧卡斯的論證，而且顯然是持贊同的態度，他自己也提出一個所謂的利用不完備性定理的「歸謬論證」來說明人類的數學思惟不可能被任何演算法所模擬，這當然也意謂

著機器不可能完整模擬人類的心靈。他先假設數學家在做數學判斷時所用的方法是某種演算法。由於建立數學真理時所用的論證是可傳達的 (communicable)，故應存在著一個普遍的演算法 (universal algorithm) 吾人可據以判斷數學真理，但我們將永遠無法知道這即是吾人據以判斷數學真理的演算法，否則我們將能建構它的哥德爾語句，同時也知道這語句是一項數學真理，但根據哥德爾的第一個不完備性定理，這是不可能的。可是這牴觸了一個關於數學的傳承與訓練的事實：數學真理是建立於一些簡單、 明顯而且是所有的人都能接受的要素，而不是某個沒有人能知道的演算法 (Penrose, 1989: 417-8)。潘若斯前半段的論證大致上類似前述哥德爾對「主觀數學」的看法，亦即，即使有個有限公理化系統能窮盡主觀數學，我們也不可能知道它是這樣的一個系統。但我們實在很難看出此論證的後半段會和「有某個我們永遠無法知道的演算法存在」有何矛盾，縱使「我們看到的以及所理解的」數學真理都是建立在「我們所清楚認知」的一些簡單明顯的要素上，但我們之所以會如此，也許是因爲不自覺地遵循某個我們永遠無法知道的演算法，或換句話說，我們可能是某種機器被設定成能夠清楚認知並接受某些數學原則，但卻無法認知自身的設定。

潘若斯在 1994 年的《心靈的陰影》(*Shadows of Mind*) 中又提出另外一個論證 (Penrose, 1994: Chapter 3)。他的論證相當長而且牽扯相當多，但在 1996 年他又提出了其論證的摘要。

> 我們試著假設人類在原則上所能使用的所有的（無懈可擊的） 數學推論的方法，其整體能被包含在某個（但不一定須是計算理論的）健全的形式系統 F。當一個數學家面對 F 的時候，可能會做出底下的論證（要記住「我是 F」僅是「F 正好包含了所有的人類能使用的數學證明的方法」的縮寫）：

> (A) 雖然我不知道我一定是 F，我得到的結論是，如果我是的話，則 F 這個系統就必須是健全的 (sound)。而且更進一步，F′ 也必須是健全的，在此 F′ 是 F 加上「我是 F」這個斷言。我了解從「我是 F」這個假設可以得到 F′ 的哥德爾語句 G(F′) 必須是真的，而且更進一步，它不是 F′ 的邏輯結果。但是我剛說我了解「如果我碰巧是 F，則 G(F′) 會是真的」，而且這種類型的了解正是我們假設 F′ 能夠做到的。因為我能了解超出 F′ 能力所及的某些事（筆者按：即 F′ 無法了解「如果我碰巧是 F，則 G(F′) 會是真的」），我推得以下的結論：我不可能是 F。此外，本論證適用於任何其他的（可哥德爾化的 (Gödelizable)）系統。(Penrose, 1996: 3.2)

縱使是簡要的版本，批評潘若斯的學者對這論證的詮釋眾說紛紜。底下將介紹弗蘭森，林德斯仲 (Lindström)，夏皮洛 (Shapiro) 與恰爾莫斯 (Chalmers) 等四位學者的解讀與批評。最後兩者的說法頗爲相似，將放在一起介紹。

1. 弗蘭森的批評

首先讓我們將「我是F」寫爲 IAMF。我們知道一個系統是健全的若且唯若它是一致的。[11] 如果 IAMF 爲真，F 完整地捕捉了

---

[11] 潘若斯所謂的 soundness 是指 Π_1-soundness。見 Penrose, 1994: 3.16。一個理論 T 是 Π_1-sound 若且唯若任何 T 能證明的 Π_1-語句都是真的（在 T 的所有模型中為真）。這裏的 Π_1-語句是指 $\forall xR(x)$ 這種形式的語句，其中 R(x) 在 T 之下是可被決定的，亦即對所有的自然數 n，$T \vdash R(\underline{n})$ 或 $T \vdash \neg R(\underline{n})$。而 T 是 Π_1-sound 若且唯若 T 是一致的，這是因為根據史卡倫 (Skolem) 的定理，任何的理論 T 都有

我們的數學證明方法，所以必然是健全的，因此也是一致的，而 F 加上 IAMF 當然也是一致的。而我們才剛做了上述的推論，所以我們可以得到底下兩項前提：

1. 若 IAMF，則 F+IAMF 是一致的。

2. 我能證明：若IAMF，則F+IAMF是一致的。[12]

因爲 IAMF 表示 F 完整地捕捉了我們的數學證明方法，所以，對任何數學命題 A，如果我們能證明 A，F 也能證明 A。故我們又有第三項前提：

3. 若 IAMF，則對任何數學命題 A，如果我能證明 A，F 也能證明 A。

利用這三項前提，我們可以得到底下的歸謬證明。假設 IAMF，根據 2 及 3，F ⊢IAMF→CON(F+IAMF)。可是很明顯的，根據單調性 (monotonicity) F+IAMF ⊢IAMF 而且 F+IAMF ⊢IAMF→CON(F+IAMF)，再根據肯定前項 (Modus Ponens)，F+IAMF ⊢CON(F+IAMF)，也就是 F+IAMF 能證明它本身是一致的，這違背了哥德爾的第二個不完備性定理，所以 IAMF 不可能爲真。

一個開放的保守延伸 (open conservative extension)，亦即一個 T 的保守延伸的理論，其公理都是開放語句。關於這個定理及相關的定義及定理，可參 Shoenfield, 1967: 48-57。

[12] 按潘若斯的原文，我們能證明：若 IAMF，則 G(F′) 為真。但 CON(F+IAMF) 及 G(F′) 兩者是等價的（見註解 6）。故弗蘭森用「F+IAMF 是一致的」來取代「G(F′) 為真」。

乍見之下，潘若斯的論證不但是有效的，而且還是健全的，但弗蘭森指出了一個問題，那就是我們必須要能找到一個語句 IAMF 使得我們有理由宣稱「如果 F 確實完整地捕捉了人類心靈所能使用的數學證明的方法，則 IAMF 爲真」，弗蘭森認爲如果潘若斯不能找到上述這樣的一個語句，則其論證仍不能有什麼作用。[13] 潘若斯也意識到類似的問題，他擔心的是，如果有人提出某個F並且宣稱它完整地捕捉了人類心靈的數學證明方式，則要如何保證「我是F」這項陳述能夠被形式化，因爲若它不能形式化，則我們不能得到 F+IAMF ⊢CON(F+IAMF)，也因此不能利用第二個不完備性定理來得到矛盾。潘若斯沒有正面回答這個問題，但他似乎認爲他在第二本著作中已經解決這個問題 (Penrose, 1996: 3.3)。筆者在此要指出：潘若斯的論證實際上證明了任何的形式化的IAMF都會爲假，可是這並不蘊涵心靈不可能是機器（即某個F），而只蘊涵了「心靈是機器」這個語句不可能被形式化。換句話說，心靈可能是某個F，而任何被視爲形式化 IAMF 的語句都會被 F 證明爲假。這也表示我們的心靈可能是某個F，但我們卻無從得知「我是F」這件事（機器之「知」即「證明」）。再一次的，這裏的論點又回到關於哥德爾所謂的主觀數學的可能性，亦即，有可能有個有限公理化的系統涵蓋了主觀數學，但人類不可能知道這件事。

2. 林德斯仲的批評

接下來將簡要地介紹林德斯仲的解讀及批評。他把潘若斯的論證重構如下：

---

[13] 以上的論證重構及批評見Franzén (2005: 119-21)。在文中筆者加入了一些補充說明。

1. 若 Sd(F) 則 Sd(F+)。

2. 若 Sd(F+) 則 G(F+)。

3. 若 Sd(F+) 則 F+ 不能證明 G(F+)。

4. 若 Sd(F) 則 G(F+)。

5. 若 Sd(F) 則 F+ 不能證明 G(F+)。

6. 若 HC(F) 則 F 證明 Sd(F)→G(F+)。

7. Sd(F) 與 HC(F) 不能同時爲真。

8. “I am F” 爲假。

其中 Sd(F) 表示 F 是健全的 (sound)，HC(F) 表示F最少包含了人類心靈所使用的所有的數學證明的方法，而 F+=F+Sd(F)，G(F+) 是 F+ 的哥德爾語句。上述的論證其實有三項前提，即 1，2，3。林德斯仲承認從這三項前提確實可以得到 8 這個結論，因爲從 1，2 很明顯的可得到 4，從 1，3 可得到 5，從 4 可得到 6 （4 是人類可證明的，所以如果 HC(F) 真的成立，則 F 也能證明 4），從 5，6 可得到 7，而從 7 可得到 8 （這是因爲，如果 “I am F” 是真的，HC(F) 及 Sd(F) 也會爲真）。前提 2 及 3 沒有什麼問題，因爲一個系統是健全的若且唯若它是一致的，而且一個系統的哥德爾語句與表達它的一致性的形式化的語句是等價的（見註解 11 及 12，基於此，底下我們會把「健全的」(sound) 用「一致的」(consistent) 來替代），因此哥德爾的不完備性定理確實蘊涵 2 及 3。可想而知，林德斯仲打算攻擊前提 1，他指出 1 應該對所有的F 的延伸理論 E 爲真，因爲任何這樣的 E 一定也包含了人類心靈所使用的所有的數學證明的方法。但林德斯仲很快地舉出了一個反例：F+¬CON(F)+CON(F+¬CON(F)) 是不一致的。首先 F+¬CON(F) 是

F 的一個一致的延伸 (consistent extension) （如果F+¬CON(F) 是不一致的，F 就會證明 CON(F)，這會違反第二個不完備性定理)，但 CON(F+¬CON(F)) 蘊涵 CON(F)，因此 F+¬CON(F)+CON(F+¬CON(F)) 是不一致的，也就是說，若讓 F+¬CON(F)=E，則 CON(E) 爲真但 CON(E+) 爲假 (Lindström, 2001: 243-7)。[14]

假設 F 正好包含了人類心靈所能使用的所有的數學證明方法。夏皮洛指出一個可能的回應是：潘若斯僅須要求前提 1 對 F 的某些延伸理論而言必須成立 (Shapiro, 2003: 29)。其實潘若斯只須確認前提 1 對 F 成立即可，毋須要求它對其他的理論也成立。林德斯仲確實指出：潘若斯並未給出任何理由來解釋前提 1 何以對 F 會成立，並認爲很難替潘若斯找出什麼理由 (Lindström, 2001: 245)。筆者卻可以想見一個直覺上難以反對的理由：如果 F 正好包含了人類心靈所能使用的所有的數學證明方法，我們確實會認爲「F 是一致的」是一個數學上的事實（因爲這裏的心靈是「理想化」的，所以不可能犯錯），因此原來的 F 加上這個事實還是一致的。

3. 恰爾莫斯與夏皮洛的批評

恰爾莫斯認爲潘若斯的論證的最大弱點在於預設了心靈能毫無疑問地知道（因而相信）它自身是一致的（見先前潘若斯的論證的引文)。恰爾莫斯用 ⊢A 來表示心靈毫無疑問地斷言 A，⊢B(A) 表示心靈毫無疑問地斷言它自己毫無疑問地相信 A。恰爾莫斯認爲

[14] 林德斯仲也考慮了其他對潘若斯所謂的「健全性」的可能的詮釋，並指出在這些詮釋下，前提 1 仍不能成立。註解 11 已經提到過關於潘若斯的「健全性」的較合理的解讀，故筆者省略了林德斯仲的其他解讀方式。

底下的命題應該都成立，(1) 若 $\vdash A$ 則 $\vdash B(\underline{A})$；(2) $\vdash (B(\underline{A_1}) \wedge B(\underline{A_1 \rightarrow A_2})) \rightarrow B(\underline{A_2})$；(3) $\vdash B(\underline{A}) \rightarrow B(B(\underline{A}))$；(4) $\vdash \neg B(\underline{\text{false}})$，其中的 (4) 表示心靈毫無疑問地相信自己是一致的。但這四個命題放在一起會導致矛盾：根據固定點定理，我們可找到一個句子C 使得 $\vdash C \leftrightarrow \neg B(\underline{C})$，根據 (1) 及 (2)，$\vdash B(\underline{C}) \rightarrow B(\neg B(\underline{C}))$，但根據 (3)，$\vdash B(\underline{C}) \rightarrow B(B(\underline{C}))$，根據這兩個命題及 (2)，$\vdash B(\underline{C}) \rightarrow B(\underline{\text{false}})$（因爲 $\vdash \neg C \leftrightarrow (C \rightarrow \underline{\text{false}})$），可是 $\vdash B(\underline{\text{false}}) \rightarrow B(\underline{C})$（因爲 $\vdash \text{false} \rightarrow C$，再利用 (1) 及 (2)），因此 $\vdash B(\underline{\text{false}}) \leftrightarrow C$，所以 $\vdash C \leftrightarrow \neg B(\underline{\text{false}})$，得到矛盾。因爲 (1) 到 (3) 並無不妥，(4) 必須負責任 (Chalmers, 1995: Section 3)。[15]

林德斯仲指出，恰爾莫斯的證明有個問題，就是我們雖然利用固定點定理得到 $\vdash C \leftrightarrow \neg B(\underline{C})$，$\neg B(\underline{C})$ 不見得是有意義的，因爲它表達了「我不相信現在所說的這句話（筆者按：即引號內的句子）是真的」，我們很難說這句話到底是什麼意思（但一個系統的哥德爾語句卻明顯是有意義的，它表達了「我不能被該系統證明」，而我們知道這是什麼意思，因爲「證明」這個概念有很明確合理的詮釋，見註解 3）。因此，就算上述的形式化系統得到矛盾，我們只能說，在此系統內，「信念」的詮釋是不合理的，但這與心靈能不能毫無疑問地知道本身是一致的並沒有關聯 (Lindström, 2001:

[15] 在此 $\underline{A}$ 表示 A 的哥德爾碼的名字，也就是說，如果 A 的哥德爾碼是 m 則 $\underline{A}$ 是 $s(s(\ldots(s(\underline{0})\ldots)$，其中 s 出現了 m 次。而恰爾莫斯其實是用 B(A) 做為 B(`A´) 的簡寫，其中 `A´ 是 A 的哥德爾碼。但他犯了一個技術上的小錯誤，亦即 `A´ 應該是 A 的哥德爾碼的名字才對。筆者為了與之前的符號一致起見，故用底線來取代單引號。

248-9)。

在介紹夏皮洛的批評之前，先做以下的一些設定。假設 K 是主觀數學限制在自然數算術的部分。K′ 是人類心靈從 $K=W_e$（亦即主觀數學限制在自然數算術的部分正好是某個圖靈機器能證明的自然數算術語句所形成的集合）[16] 這個假設能夠推得的在自然數算術語言上的語句所形成的集合，也就是說，對任何使用自然數算術語言的語句α，$\alpha \in K'$ 若且唯若人類心靈能「毫無疑問地知道」(unassailably know)「若 $K=W_e$，則 α 是真的」。夏皮洛對潘若斯的論證之解讀如下：

1. 假設$K=W_e$。
2. 則任何屬於 K′ 的語句一定是真的，故 K′ 是一致的（根據人類心靈的「健全性」(soundness)）。
3. 根據 1，K′ 是遞迴可枚舉的 (recursively enumerable)，而且人類心靈藉著 e 能夠得到 K′ 的哥德爾語句 G′。
4. 根據 3，如果 K′ 是一致的，則 G′ 爲真但不屬於 K′。
5. 根據 2，K′ 是一致的，因此根據 4，G′ 爲真但不屬於 K′。
6. 根據 1 到 4，我們知道：如果 $K=W_e$ 則 G′ 爲真。因此，根據 K′ 的定義，G′ 屬於 K′。
7. 我們從 5 及 6 得到矛盾，故 $K \neq W_e$。

---

[16] 如果主觀數學限制在自然數算術的部分正好是某個圖靈機器能證明的自然數算術語句所形成的集合，它一定是遞迴可枚舉的 (recursively enumerable)，因此必然是某個圖靈機器的定義域，假設這圖靈機器的編碼是 e，則 $K=W_e$ 表示 K 是 e 的定義域。

夏皮洛認爲潘若斯若要把上述的論證改寫成嚴格的形式化的證明，則必須想辦法形式化「毫無疑問地知道」這個述詞，但底下的論證指出了它無法被形式化。假設這個述詞可被形式化爲 α，$\alpha(\underline{n})$ 表示人類心靈可「毫無疑問地知道」自然數 n 所對應的句式 φ_n（亦即n 是此句式的哥德爾碼）。上述的 2 是根據人類心靈的健全性而來，亦即人類心靈能「毫無疑問地知道」的算術命題都是真的，所以我們起碼應該預設：對所有的 n，$\alpha(\underline{n})\rightarrow\varphi_n$ 是真的。但我們可以利用固定點定理找到一個語句 γ，使得 $\gamma\leftrightarrow\neg\alpha(\underline{G(\gamma)})$ 是一個爲真的語句。接著我們可推得：γ 一定是真的。否則 $\neg\gamma$ 是真的，則根據 $\gamma\leftrightarrow\neg\alpha(\underline{G(\gamma)})$，會得到 $\alpha(\underline{G(\gamma)})$ 是真的，再加上 $\alpha(\underline{G(\gamma)})\rightarrow\gamma$ 是真的（假設 $G(\gamma)=n$，則 γ 其實就是 φ_n），會得到 γ 是真的，形成矛盾。所以 γ 是真的，因此 $\neg\alpha(\underline{G(\gamma)})$ 也是真的，可是上述的推論顯示了我們「毫無疑問地知道」γ 這個命題，但 $\neg\alpha(\underline{G(\gamma)})$ 又表示我們不是「毫無疑問地」知道 γ，這又形成了矛盾 (Shapiro, 2003: 26-30)。

夏皮洛的論證在形式上與恰爾莫斯的類似，但他所證明的是「毫無疑問地知道」這個述詞無法被形式化，因此林德斯仲對恰爾莫斯的批評並不適用於此（夏皮洛確實也知道林德斯仲的批評）。夏皮洛考慮了潘若斯可能做出的以下幾個回應，並一一加以反駁。

(1) 首先，潘若斯可能會預設：(*) 只有當 φ_n 不牽涉到自我指涉時，$\alpha(\underline{n})\rightarrow\varphi_n$才會成立。$K=W_e$ 可以形式化爲 $\forall x(\alpha(x)\leftrightarrow x\in W_e)$，假設 S 是所有的滿足 $\alpha(\underline{n})\rightarrow\varphi_n$ 的語句 φ_n所形成的集合，則 $K'=\{\varphi\in S: \alpha(\underline{\forall x(\alpha(x)\leftrightarrow x\in W_e)\rightarrow\varphi})\}$。[17] 此時 K' 是健全的（或一致

[17] 根據定義，K' 是從 $K=W_e$，在此即 $\forall x(\alpha(x)\leftrightarrow x\in W_e)$，這個假設能夠推得的在自

的）表示對任何的 φ，$(\alpha(\underline{\forall x(\alpha(x)\leftrightarrow x\in W_e)\rightarrow\varphi})\wedge\forall x(\alpha(x)\leftrightarrow x\in W_e))\rightarrow\varphi$，因此$\alpha(\underline{\forall x(\alpha(x)\leftrightarrow x\in W_e)\rightarrow\varphi})\rightarrow(\forall x(\alpha(x)\leftrightarrow x\in W_e)\rightarrow\varphi)$，而這違反了一開始的假設 (*)。

(2) 再者，潘若斯或許會採取塔斯基 (Tarski) 的方式來分層化 (stratify)「毫無疑問地知道」這個述詞。也就是說，我們從某個 α_0 開始，使得對某個由語句所構成的集合 S，對所有的 $\varphi_n\in S$，$\alpha_0(\underline{n})\rightarrow\varphi_n$ 成立。接著，我們定義一個述詞 α_1，使得對任何帶有 α_0 的語句 φ_n，$\alpha_1(\underline{n})\rightarrow\varphi_n$ 成立。這時候 $K=W_e$ 可以形式化爲 $\forall x(\alpha_0(x)\leftrightarrow x\in W_e)$，而$K'=\{\varphi\in S: \alpha_1(\underline{\forall x(\alpha_0(x)\leftrightarrow x\in W_e)\rightarrow\varphi})\}$，不會導致 (1) 的矛盾。機器論者必須同意 K′ 是遞迴可枚舉的（心靈如果是一部圖靈機器，K′ 當然會是遞迴可枚舉的），但不見得會同意有某種有效程序使我們能從 K 的指標 e 得到 K′ 的指標。若沒有 K′ 的指標，我們看不出要如何得到 K′ 的哥德爾語句 G′，因此沒有理由接受潘若斯的論證（夏皮洛的解讀版） 之前提 3。[18]

然數算術語言上的語句所形成的集合，夏皮洛顯然預設我們從 A 推得 B 若且唯若我們可「毫無疑問地知道」A→B。但為了要形式化「毫無疑問地知道」，這項預設應是合理可被接受的。

[18] 夏皮洛並沒有解釋如何從 K′ 的指標來得到 K′ 的哥德爾語句G′，筆者在此略做解釋如下。如果知道 K′ 的某個指標e′，我們可依循以下的有效程序來找到 K′ 的哥德爾語句：因為 K′ 是遞迴可枚舉的，所以有某個句式γ 使得 n∈ K′ 若且唯若 $PA\vdash\gamma(\underline{n})$（亦即 K′ 在 PA 中是「弱可代表的」(weakly representable)，而 e′ 即是 γ 的哥德爾碼，因此由 e′ 我們可有效地得到 γ 這個句式，而且很顯然地，對任何的語句a，我們可以有效地得 $\gamma(\underline{a})$ 的哥德爾碼。現在我們可以一個一個地核對 K′ 的元素，看看何者會是 $a\leftrightarrow\neg\gamma(\underline{a})$ 這種形式的語句的哥德爾碼，a 所對應的語句即是 K′ 的哥德爾語句G′。這 a 一定存在，因為夏皮洛假設 K′ 所對應的公理化系統

(3) 或許潘若斯會預設 $\alpha(\underline{n})\rightarrow\varphi_n$ 雖然不是對所有的 n 都成立，但心靈卻能知道對哪些 n 會成立。如同在 (1) 中所提到的，$K'=\{\varphi\in S:\ \alpha(\underline{\forall x(\alpha(x)\leftrightarrow x\in W_e)\rightarrow\varphi})\}$，爲了要形式化 K′（亦即用一個一元述詞 k′ 來定義 K′），需要再引進真值述詞 (truth predicate) T，然後可以定義：k′(x) 若且唯若 x 是一個算術語言的句式的哥德爾碼而且 $\alpha(\underline{\forall x(\alpha(x)\leftrightarrow x\in W_e)\rightarrow T(x)})$。[19] 一旦如此，我們會得到一個語意封閉的語言 (semantically closed language)，而爲了要避免類似說謊者悖論 (liar paradox) 的矛盾產生，必須預設心靈有能力去篩選能滿足 T 架構 (T-scheme) 的語句。[20] 在此，我們可由K的哥德爾碼有效地得到 K′ 的哥德爾碼，所以也可以有效地得到 K′

滿足我們在第一節所提到的五個條件，因此 $a\leftrightarrow\neg\gamma(\underline{a})$ 的哥德爾碼一定會在 K′ 中，而且因為 K′ 是遞迴可枚舉的，我們在有限步驟內即可找到 $a\leftrightarrow\neg\gamma(\underline{a})$ 的哥德爾碼。

[19] T 必須要滿足一般的真值述詞的公理，這在大部分的邏輯教本都可找到（一般稱之為塔斯基的真值定義 (Tarski's Truth Definitions)）。由於夏皮洛是在考慮潘若斯的論證能不能被形式化成一個嚴格的證明，故有必要試著將 K′ 形式化。而 K′ 是由某些真的語句所構成的集合（K′ 的所有語句都為真），因此我們必須引入真值述詞來定義它。

[20] 一個語言是語意封閉的如果該語言具有用以表徵其語言表式之語意值的述詞，如：真值述詞，而且用來表示該語言中的表式之名稱的表式亦在該語言中。一個語意封閉的語言若接受所有的 T 架構的例子 (instances of T-scheme) 在古典邏輯下會導致矛盾。因為這裏假設 K′ 是一致的，蘊涵了心靈必須具備某種能力來去除會導致矛盾的 T 架構的例子。

的哥德爾語句 G′。[21] 夏皮洛指出，這樣一來，潘若斯的論證可被形式化如下：

1. $\forall x(\alpha(x)\leftrightarrow x\in W_e)$，這即是$K=W_e$的假設（這是此證明唯一的假設）。
2. $\forall y((\forall x(\alpha(x)\leftrightarrow x\in W_e)\wedge k'(y))\rightarrow\varphi_y)$，這是把「若 $K=W_e$ 則所有的 K′ 的語句都是真的」這樣一個後設邏輯事實形式化。
3. Tg′，根據 e 可得到 k′ 的哥德碼 e′，再據此得到 K′ 的哥德爾語句 G′ 的哥德碼 g′，因爲根據 2，K′ 是一致的，所以 G′ 爲真，根據 T架構，得到 Tg′。[22]
4. $\forall x(\alpha(x)\leftrightarrow x\in W_e)\rightarrow Tg'$，解消掉 1 可無條件得到這個條件句。
5. $k(\forall x(\alpha(x)\leftrightarrow x\in W_e)\rightarrow Tg')$，可被證明的當然是可毫無疑問地知道的（雖然夏皮洛沒有明講，但這是一個預設的推論規則）。
6. $k'(g')$，這是根據 k′ 的定義得到的。
7. $\forall x(\alpha(x)\leftrightarrow x\in W_e)\rightarrow k'(g')$，根據 1 及 6，由解消 1 得到。
8. $\forall x(\alpha(x)\leftrightarrow x\in W_e)\rightarrow\neg k'(g')$，根據哥德爾的第一個不完備性定

---

[21] 根據假設，α 這個述詞定義了 K，而 α 的哥德爾碼是 e；k′ 定義了K′，而我們利用 α 來明確定義 k′，故我們可透過 α 的哥德爾碼有效地得到 k′ 的哥德爾碼。相關的編碼方式可參 Enderton, 2001: Ch.3。

[22] 其實夏皮洛所做的事是把所有的後設的推論都形式化，所以 2 可視為是 K′ 的一致性的形式化：如果 φ_y 是矛盾句或者會導出矛盾，則根據 RAA（歸謬證明法則），$\neg k'(y)$，亦即 $y\notin K'$。

理，$g' \notin W_e'$，故$\neg k'(g')$，再解消 1 即得。

9. $\neg \forall x(\alpha(x) \leftrightarrow x \in W_e)$，根據 7 及 8 得到。[23]

但夏皮洛指出，機器論者沒有理由接受上述關於心靈能力的預設，也就是心靈有辦法決定是否「毫無疑問地知道」某語句或者 T 架構是否能用在某語句上。雖然上述的證明與機器論者的主張矛盾，機器論者反而會認爲這表示這些關於心靈能力的預設是應該被駁斥的。[24]

以上介紹了文獻中對潘若斯的新論證的一些批評。潘若斯從未清楚地說明如何形式化其論證，而他在邏輯方面的專業素養不足也一直爲人所垢病，例如戴維斯 (Davis) 就批評潘若斯的邏輯素養「粗疏草率」(slapdash scholarship) (Davis, 1993: 611-2)，由此觀之，夏皮洛等人之解讀實過於寬大 (charitable)。依筆者之見，弗蘭森的解讀最直截了當（弗蘭森其實也知道夏皮洛等人之解讀），符合潘若斯自己所提出的新論證的綱要。但無論如何，根據上文的討論，

---

[23] 這裏的證明是筆者參照夏皮洛的說明，見 Shapiro, 2003: 34，再加上筆者的補充解讀之後所構成的。

[24] 雖然夏皮洛的主要批評到此結束，但他又考慮了兩個可能性：(1) 利用所謂的「真理修正理論」(revision theory of truth) 來定義「毫無疑問地知道」；(2) 用邏輯蘊涵 (logical implication) 來定義K′。見 Shapiro(2003: 35-39)。夏皮洛對 (1) 及 (2) 的討論都涉及了超限數 (transfinite numbers)，筆者認為實在牽扯太遠了。夏皮洛的用意是要指出，不論潘若斯如何解釋他的論證，它都不會成為一個決定性的證明。不過，如果為了證明心靈不是機器而直接引進一個機器明顯無法滿足的假設並「宣稱」這是心靈所具備的能力，則僅會是循環論證而已，而且與不完備性定理也沒有什麼關聯。

潘若斯的新論證不能說是成功的，更遑論能被視爲嚴格的數學證明了。

三、麥扣 (McCall) 的論證及相關的回應

麥扣知道盧卡斯的論證和潘若斯的論證以及他們所遭受的詰難，他自己則著眼於「可被證明的」(provable) 與「爲真的」(true) 之間的差異，來論證心靈不是機器：對於任何機器，心靈知道有些爲真的語句是它所無法證明的，而「無法證明但是爲真」這樣的概念超出了機器所能掌握的範圍。麥扣以 PA 爲例，PA 的哥德爾語句G 無法被 PA 證明，但我們知道它爲真，可是對模擬 PA 這個系統的機器而言，它無法將 G 歸類爲「爲真但不能被證明」。限制在自然數算術的語言，麥扣考慮了機器處理「爲真的」這個概念的三個可能性：(1) 這個概念可被某個自然數算術語言的句式所定義；(2) 引進一個新的述詞 T 來形式化「爲真的」這個概念；(3) 用一部更複雜的機器來證明原有的機器所無法證明但卻爲真的語句。麥扣直接訴諸塔斯基的不可定義性定理（見本文第一節）將 (1) 排除掉。對於 (2) 麥扣提出了兩個質疑：首先，引進 T 可能於事無補，因爲若無法得到 T(G)，則前述的困境還是無法解決；再者，就算引進了 T 可使得原有的機器證明額外的定理，但這與「爲真的」這個概念仍談不上有何關聯。麥扣另外舉了一個例子來說明第二個質疑：如果一部機器原本是設定來模擬歐基里德的幾何系統，現在我們將之重設，把平行公設 (parallel postulate) 改爲羅巴切夫斯基的公設 (Lobatchevsky's postulate)，重設的機器顯然也不會把舊的公設歸類爲「爲真的但無法證明」。對於 (3)，麥扣指出不論機器如何複雜，都會有某個它所無法證明但是我們知道爲真的語句，

因而縱使它解決了原來的機器所面臨的問題，其自身仍會陷入同樣的困境 (McCall, 1999: 525-32)。[25]

麥扣對 (1) 的回應沒有什麼問題。但對 (2) 的回應很自然地會牽涉到 (3)，因爲我們若要引入一個新的述詞 T，必然會設法讓它能夠處理舊的系統的真值定義 (truth definition)，也就是新的系統其實會是一個更強的系統，強到足以將舊系統的有關真值的後設邏輯論證形式化，例如集合論就能夠形式化自然數算術理論的真值定

---

[25] 麥扣知道文獻中對盧卡斯的主要批評，也就是面對一個系統 T 我們不見得知道它是一致的，故我們也不知道其哥德爾語句是否為真，我們僅知道若 T 是一致的，則其哥德爾語句為真，但這也是 T 自身所能證明的（假設這系統滿足第一節所提到的 (i) 到 (iv)；根據第二個不完備性定理，$T \vdash Cons(T) \rightarrow G_T$)。總之，面對 T，我們也不見得會把其哥德爾語句歸類為「無法證明但是為真」。而麥扣認為 $Cons(PA) \rightarrow \neg Prv(\neg G)$（這裏的 Prv 表示「可被PA證明的」，而 G 是 PA 的哥德爾語句，故精確一點應做 Prv_{PA} 及 G_{PA}；下文中的 proof 也是一樣）是 PA 所不能證明的，但我們知道其為真（這其實是第一個不完備性定理的應用：PA是 ω-consistent，因此若它是一致的，則不能證明其哥德爾語句的否定），所以他推論：任何 PA 的一致的延伸理論 T 都不能證明 $Cons(T) \rightarrow \neg Prv(\neg G_T)$，而我們卻能論證其為真。但是麥扣的假設：PA不能證明$Cons(PA) \rightarrow \neg Prv(\neg G)$ 其實是錯誤的，筆者在此說明如下：$Cons(PA) \rightarrow \neg Prv(\neg G)$ 若不是 PA 的定理，則會有一個 PA 的模型 M 同時滿足 Cons(PA) 及 $Prv(\neg G)$。但 Cons(PA) 與 G 在 PA 下是等價的，這表示Cons(PA) 及$Prv(\neg Cons(PA))$ 可同時被M 所滿足。根據定義，$Prv(x) =_{df} \exists y Proof(y, x)$，因此 $M \models \exists y Proof(y, \neg Cons(PA))$，因為 PA 是 ω-consistent，所以會有個自然數 n 使得 $M \models proof(\underline{n}, \neg Cons(PA))$，可是這表示 $PA \vdash \neg Cons(PA)$，亦即 $PA \vdash \exists y proof(y, G)$，但這不可能，因為 PA 不能證明 G 而且 PA 是 ω-consistent。

義。所以真正的問題在於麥扣對於（3）的回應是否成立。麥扣的回應如果能成立，就必須指出什麼樣形式的語句是每一個複雜的系統所不能證明但我們卻能知道其爲真，但他的建構並不成功（見註25）。蓋夫曼 (Gaifman) 避開前述的「不能證明但爲真」的語句的建構，提出一個想法類似於麥扣對 (3) 的回應的論證如下。蓋夫曼指出，數學家會把他從事數學推論的架構 (framework) 的一致性視爲理所當然，如果 T 是形式化這個架構所得到的理論，則他會認爲 T 是一致的，亦即會接受 Cons(T)。之所以如此是因爲數學家能夠對自身的數學推論進行反省 (reflecting on one's mathematical reasoning) 而了解它能夠被形式化爲 T，故能接受 Cons(T) 做爲合法的數學定理。根據不完備性定理，T 無法證明 Cons(T) 而我們知道其爲真。T 的擴充理論 T′=T+Cons(T) 會是一致的理論而且能證明 Cons(T)，可是 Cons(T′) 是 T′ 所無法證明而我們卻知道其爲真。我們可以用同樣的方式來得到 T″，T‴……等等，但這每一個理論都有一個語句是它所無法證明而我們卻知道其爲真，也就是它自身的一致性。這也就是說，沒有任何形式化的理論 T 能真正代表數學家的數學推論的架構。但是蓋夫曼同時也指出這項論證的一個漏洞：我們知道 T 是一致的，因爲我們相信自身的數學推論架構是一致的，而且也「知道」T 確實形式化了我們的數學推論的架構；但是有可能 T 確實形式化了我們的數學推論的架構，可是我們卻不知道，因此也不見得會知道 T 是一致的。因此，我們有可能是某種機器而永遠無法知道自己是機器 — 當我們面對一部完整地模擬吾人數學推論的機器時，我們永遠無法知道它是如此 (Gaifman, 2000: 462-70)。我們可以看出：蓋夫曼的批評與哥德爾對是否有某個有限公理化的系統能窮盡主觀數學的看法相當接近（見第二節）。

肆、相關概念的檢討

一、什麼是「機器」及機器如何「模擬」或「代表」心靈？

哥德爾的不完備性定理通常是針對形式化的演繹系統，而一個演繹系統是基於某個「合理」的第一階邏輯的語言並由有限多的推論規則及可被決定 (decidable) 的一組公理所構成，[26] 在這樣的設定下，我們可以用一個圖靈程式 (Turing program) 來模擬該系統，這程式可被視爲一個「定理證明器」(theorem prover)。所以在上文的討論脈絡裏所指的機器或圖靈機器，嚴格來說，其實是一個圖靈程式。我們也知道現今的任何電腦語言的程式都可改寫成一個等價的圖靈程式，因此只要考慮圖靈程式即可。理論上，我們可以把一個定理證明器當作是一個局部遞迴函數 (partial recursive function)，也就是我們可將一個語句的哥德碼輸入，如果它是一個定理，則在有限步驟內，定理證明器會輸出一個數字，例如 1，以表示肯

[26] 所謂合理的語言是指可被遞迴編碼 (recursively numbered) 的語言，見Enderton, 2001:142-3 和 225。「可被決定性」在此可換為「可有效枚舉的」(effectively enumerable)（根據 Church's thesis，這兩個概念分別等價於「遞迴的」及「可遞迴枚舉的」），因為在一個合理的語言下，任何由一組可有效枚舉的公理所形成的理論必然可被某一組可被決定的公理所公理化，這是眾所周知的後設邏輯定理（亦即 Craig's theorem），可見 Monk, 1976: 262-3。又，有些演繹系統可能沒有公理而僅有推論規則，例如自然演繹法的純邏輯系統 (pure logic of natural deduction)。

定，否則有可能永遠不會輸出任何數字。[27] 因此一個定理證明器能證明的定理（的哥德爾碼）所形成的集合一定是遞迴可枚舉的 (recursively enumerable)。[28] 在當前的討論脈絡下，「機器」必須如是理解。

那麼機器要如何「模擬」或「代表」心靈呢？文獻中常常語焉不詳。當然在我們的討論脈絡下並不要求機器能夠全面地模擬心靈，而是僅就數學推論的能力而言，甚至可以只限定在關於自然數算術理論的推論能力上。一個立即的觀察是：數學家通常不可能用機器的方式來尋找證明（見註解 27），例如費弗曼 (Feferman) 指出：

> 實際上，數學家們透過不可思議的以下幾項的組合來得到證明：啟發性的推論 (heuristic reasoning)，洞察 (insight)，靈感 (inspiration)……而在此並沒有普遍的法則存在……要得到數學上的成功並沒有公式……

---

[27] 這是因為從一組可被決定的公理所得到的所有的證明的哥德爾碼所形成的集合會是可被決定的，因此定理證明器可以從數字 0 開始執行以下的步驟：(1) 檢查它是否為一個證明的哥德爾碼，若是，進行 (2) 若不是則跳到下一個數字並回到 (1)；(2) 檢查這證明的最後一項是否為輸入之語句的哥德爾碼，若是，輸出數字 1，若不是則跳到下一個數字並回到 (1)。

[28] 根據這裏的設定，若一個語句可被所考慮的定理證明器所證明，則前者的哥德爾碼會在後者所對應的局部遞迴函數的定義域中，而任何一個局部遞迴函數的定義域都是遞迴可枚舉的。又，如果該定理證明器所模擬的是一個可被決定的系統，則若所輸入的語句不是一個定理，它也會輸出某個數字來表示，不過它的定義域仍會是遞迴可枚舉的。

> 就這面向來看，數學的想法，就其如何被產生而言，不會是機器式的 (Feferman, 1995: 4.2 and 4.3)。

如果機器能不能模擬心靈指的是機器是否能夠依心靈的思惟方式來得到證明，大部分的邏輯學家或數學家應會肯定地回答「不能」。或許會有人認爲這樣的回答過於武斷，因爲人工智慧的發展常就是試圖用機器來模擬人類的思維方式。但如此一來，要考慮機器是否能模擬心靈就必須考量心靈的種種數學思惟方式，但這些思惟方式是無法清楚界定的（更遑論要將之形式化了），因此不是一個用邏輯就能處理的問題，更不用說僅依靠不完備性定理了。

林德斯仲曾區分強及弱的兩個「心靈不是機器」的主張。弱的主張：機器不可能完整地模擬人類發現數學證明的方式；強的主張：不可能存在有某個機器其所能證明的數學命題所成的集合包含了(或等於)心靈所能證明的數學命題所成的集合 (Lindström, 2001: 242)。剛才已經提到，弱的主張，在我們的討論脈絡下，不會有什麼爭議，所以若有問題應是在強的主張上。同樣的，這裏的數學命題也可以限定在自然數的算術理論上。換言之，機器在這個脈絡下可以用它所能證明的自然數算術的定理所形成的集合來取代。如果我們這樣子界定問題，則就相當明確，因而能從形式邏輯的觀點來考慮。很明顯的，剛提到的弱的主張其實是反對強的機器論的主張，亦即機器可以模擬人類數學證明的方式，而強的主張是反對弱的機器論的主張，亦即心靈所能證明的數學命題所成的集合包括於（或等於）某個機器所能證明的數學命題所成的集合。所謂的強弱之分是根據邏輯蘊涵的方向來定義：強的主張蘊涵弱的主張。據此，潘若斯等人的反機器論立場應該界定爲反對弱的主張。當然，若潘若斯等人的論證成立的話，強的機器論主張也不能成立。

要注意的是，就算潘若斯等人的論證成立，也不蘊涵「實際上」

機器一定無法模擬人類的心靈，這是因爲機器論者與反機器論者之間的爭辯所涉及的其實是「理想化」的心靈及「理想化」的機器。如同夏皮洛所指出的 (Shapiro, 2003: 20-21)，現實中的機器有記憶體的限制，硬體有時會故障，軟體有時會有錯誤，但理想化的機器沒有記憶體的限制，且我們常用一個演算法或形式化的演繹系統或語句所構成的集合來取代實體的機器，故也不需要考慮硬體可能會發生的問題。而理想化的心靈所指的並非任何特定個體的心靈，它沒有生命的限制，不會疲勞，不論一個計算要花多長的時間，它都能專注在上面，並且不會犯錯。據此及上文所述，文獻中所討論的問題，精確地說，應是：「是否有一理想化的機器能證明理想化的心靈能證明的所有的自然數算術的語句」。

二、 什麼是「證明」？

「證明」這個概念在形式邏輯上有很清楚的定義（見註解 3），但是此定義通常僅適用於公理化系統，可是要給出一個數學證明，不見得要先給出一個公理化的系統。在早期，許多數學家大都訴諸一些數學上的直覺來建構證明。公理化系統的提倡是晚近的事，而最爲人所知的提倡者是希爾伯特 (Hilbert)，他的主張一般稱之爲「希爾伯特計畫」(Hilbert's program)，其內容大致如下：所有的數學理論都應該進行公理化，並且要能證明它本身是一致的，而對其一致性的證明僅能運用「有限方式的方法」(finitary methods)。[29] 希

[29] 希爾伯特對有限主義 (finitism) 的詮釋訴諸某種康德式的直覺，這是一種對有限符號的操作的直覺。根據他自己所舉的例子，"1" 是一個符號，任何一個以 "1" 開頭，以 "1" 結尾，且任兩個 "1" 中間有個 "+" 的序列也是符號，這些符號就是數字，而且構成了所有的數字。它們是某種「邏輯之外的具體事物」(extra-logical

爾伯特在上個世紀的二零年代提出上述的計畫。該計畫的產生肇因於十九世紀末及二十世紀初所發現的一些悖論，其中最有名的就是羅素悖論 (Russell's paradox) 。[30] 羅素悖論帶來了一個警訊：見似理所當然的數學直覺卻有可能導致矛盾。羅素悖論所牽涉的直覺是：考慮任何的性質，具備這性質的事物可視爲一個整體，而我們也可考察這個整體是否具備某些性質。在羅素悖論出現之前，大概不會有人去質疑這樣的直覺有何不妥，事實上，眾所周知的，弗雷格 (Frege) 還把它用在其邏輯體系的建構。[31] 但羅素悖論的發現，使得一些邏輯學家及數學家懷疑僅訴諸見似無可質疑的數學直覺來進行證明的方式，有可能暗藏矛盾，而這也是希爾伯特鼓吹公理化系統的原因，如此一來，所有證明能用的憑藉（除了純邏輯的真理以外），都必須事先被交代清楚，也就是以公理方式列出，這增

---

concrete objects)，而我們能夠立即、直接地掌握它們的呈現（這蘊涵了數字不會是無限長的序列）。例如，我們可以不假任何論證，馬上看出在“1+1=1+1”等號兩端有同樣多的“1”及“+”出現。對希爾伯特而言，這類型的等式為真的命題 (true propositions)，見Hilbert (1922)。希爾伯特的有限主義並不是很清楚，也必須面對哲學上的一些難題，見 Zach (1998) 的討論。

[30] 羅素悖論是利用「不屬於自己」這個一元述詞所構造出來的，現簡述如下：讓我們用 x∈y 來表示 x 是 y 的一個元素（或 x 屬於 y）。現在考慮 x∉x 這個性質（這是一個一元述詞，故可視為一種性質）。將所有具備這個性質的集合放在一起形成一個集合 R，亦即 R={x: x∉x}。顯然地，R∈R 或 R∉R。若R∈R，則因為 R 的元素必須滿足 x∉x 這個性質，所以我們得到 R∉R。若 R∉R，則因為 R 滿足 x∉x 這個性質，所以 R∈R。換句話說，無論是 R∈R 或 R∉R，我們都推導出矛盾。

[31] 見弗雷格的 “Conceptual Notation” 一文，並見羅素與弗雷格的通信。兩者可見 Beaney, 1997。

加了數學證明的嚴謹性 (rigor)。但更重要的是，一個公理化的系統必須能證明其自身的一致性，這樣才能一勞永逸，不會再受到悖論的困擾。然而，如同哥德爾的第二個不完備性定理所示，有一些重要的公理化的系統不能證明自身的一致性，希爾伯特的計畫也因而破滅。[32]

公理化系統雖然增加了證明的嚴謹與清晰，但也可能帶來意想不到的限制。例如集合論最爲人所知的獨立性結果是康托爾的連續統假設 (continuum hypothesis, CH)，也就是 $\omega_1=2^{\omega}$ 的猜測，無法被集合論的公理所證明或者否證（若集合論是一致的）。[33] 但數學家不見得要接受這個結果，因爲他可能相信集合論的公理並未窮盡所有的數學直覺，所以仍可能堅持尋找 CH 的證明或否證。[34] 總之，我們應可接受：數學證明並不一定要在一個公理化形式系統下提出，只要它是由有限的數學命題所構成，而且一般人都能檢查，理解這證明的推衍過程，一旦無誤，檢查者會以「數學的確定度」(mathematical certitude) 來接受被證明的命題。

通常，數學家所持的形上學立場常會影響他如何理解「數學證

---

[32] 在此我們不考慮費弗曼曾提出的「相對化的希爾伯特計畫」(relativized Hilbert's program)。這計畫大致上是將一個理論 T 的一致性「化約」(reduced) 到另一個理論 T′，亦即若 T′ 是一致的，則 T 也必須是一致的。這種化約的用意在於：如果我們有更多的理由相信 T′ 是一致的，則我們也會接受 T 的一致性。見 Feferman, 1988: 364-84。

[33] 哥德爾在 1939 年證明了：如果 ZFC 是一致的，則 ZFC 不能證明 ¬CH；而柯亨 (Cohen) 在 1963 年證明：如果 ZFC 是一致的，則 ZFC 也不能證明 CH。見 Kunen, 1980: Ch.6 and Ch.7。

[34] 這事實上也促使一些邏輯學家提出尋找集合論新公理的計畫，見 Koellner, 2003。

明」這個概念。著名的數學家哈地 (Hardy) 曾如是說：

> ……我一直認為一位數學家主要是一個「觀察者」：從遠處凝望著山脈並記錄下他的觀察。他的目的就是要就其能力所及清楚地區分不同的山峰，並將這區分告知他人……。當他看到一座山峰，他相信它在那裏純因為他看到那座山峰。如果他希望別人能看到那座山峰，他就直接指向它或是經由一系列的峰頂引導他去認出該山峰。當他的學生也看到那座山峰時，相關的探究 (research)、論證 (argument)、證明 (proof) 就已完成。這是一個粗糙的類比，但我相信它並不全然讓人誤解。如果我們要把這類比推到極限，會得到一個相當弔詭的結論：嚴格來說，沒有數學證明這種東西，我們終究不能做任何事而僅能「指出」(point)……(Hardy, 1929: 18)

哈地很明顯地是一個徹底的數學上的柏拉圖主義者 (Platonist)。潘若斯的立場就算沒有那麼強，也是一個柏拉圖主義者，而且他認爲大部分的數學家（包含哥德爾）都是柏拉圖主義者 (Penrose, 1996: Section 9)。對柏拉圖主義者而言，數學真理是關於某種抽象實在的客觀事實，如果有某種方式讓人能清楚地看到某個數學上的事實，則柏拉圖主義者也可能將之視爲一種「證明」，因此對柏拉圖主義者來說，要得到具有「數學確定度」的信念，不見得要透過形式化的證明才可以。問題是，如果對人類而言確實存在著某些不能被形式化的數學證明，則心靈不是機器這個命題已經直接被蘊涵，因爲所有的機器能給予的證明都是可被形式化的（如前所述，機器本身可被等同於一個形式化的理論），但如此一來也不用訴諸不完備性定理了。據此，我們可以看出機器論者的負擔是相當大的，他必須先說明心靈能製造出的所有數學證明都可以被形式化，然後又必須論證機器能證明心靈所能證明的所有數學命題（可

以只限定在自然數的算術上）。但前者是未決的議題，況且要面對諸多柏拉圖主義者，因此這顯然不是簡單的任務。

三、 什麼是「一致」？

一般而言，在邏輯上「一致」(consistency) 有兩種定義方式，語法的 (syntactical) 或語意的 (semantic)。前者定義：一個理論 T 爲一致的若且唯若它不能證明矛盾句（這裏的「證明」是嚴格形式定義下的概念）；後者定義：一個理論 T 爲一致的若且唯若它在某個詮釋下爲真，或更形式化地說，有某個 T 的語言的結構 (a structure of the language of T) 滿足T。根據完備性定理 (completeness) 及健全性 (soundness) 定理，上述兩個定義會是等價的。[35] 故看似選擇何者都沒有差別，但其實不然，因爲這其中會牽涉到形式主義者 (formalist) 與柏拉圖主義者之間的糾葛。如果我們接受語意的定義，則似乎難以避免承認有所謂的結構（大多是抽象的）存在，而所考慮的理論的每一個語句即是關於該結構某個事實的陳述。有些理論甚至會預設有所謂的意欲的模型 (intended model) 存在。例如 PA 之所以是一致的是因爲它所能證明的語句都在自然數算術理論的意欲的模型上爲真（見第一節所提到的自然數算術結構）。柏拉圖主義者當然樂於接受這樣的理解，但形式主義者不見得會願意。或許有人會認爲：我們可以接受語意的定義但不一定要抱持柏拉圖主義，因爲做爲某理論 T 之一致性的見證者的結構，其存在

[35] 完備性及健全性二者合起來等價於：T 證明某個語句 α 若且唯若任何滿足 T 的結構也會滿足 α。否定左右兩個子句即得到：T 不能證明某個語句 α 若且唯若最少有一個結構滿足 T 但是不滿足 α。因為沒有任何結構能滿足矛盾句，我們得到 T 不能證明矛盾句若且唯若最少有一個結構滿足 T。

是被集合論所保證的，亦即我們實際上可以從集合論證明該結構存在，因而也證明 T 是一致的。我們確實可以看到：基本上，模型論 (model theory) 預設了 ZFC 這樣一個公理化的集合論，而一個結構的存在也確實是預設的集合論所能夠證明的。但作爲其他理論之基礎的集合論，其本身的一致性仍會受到質疑，而且根據哥德爾的第二個不完備性定理，集合論不可能證明本身是一致的，除非它是不一致的。如果集合論是不一致的，就算它證明了某個結構的存在也無濟於事，因爲它同樣也可以證明該結構不存在。因此，若接受了語意的定義而不接受柏拉圖主義似乎會難以自圓其說。

如果僅接受語法的定義，則第二個不完備性定理會帶來很大的限制，也就是一個理論，只要強到某個地步，就不能證明自身的一致性。雖然如此，仍有學者願意採取形式主義者的立場，把這個限制視爲必須接受的事實，而尋求其他退而求其次的處理方式。[36] 但若採取形式主義的立場，一些學者視爲理所當然的心靈相對於某些機器的優勢將消失殆盡，例如我們如何知道 PA 是一致的？如果要用更強的理論來證明它的一致性，則同樣的問題會發生在那更強的理論上。但若不引進更強的理論，例如用簡森 (Gentzen) 的證明方式，則那是機器也能做到的事（見註解 36）。所以關於 PA 的一致性，心靈談不上比機器「知道」得多。

---

[36] 見註解 32 費弗曼的「相對化的希爾伯特計畫」。又例如簡森 (Gentzen) 提出了一個 PA 的一致性的「證明」，但這利用了不一樣的證明系統，其中牽涉到所謂的「序數分析」(ordinal analysis)，這是一種在有限符號操作的證明系統中模擬「超限歸納法」(transfinite induction) 的方法。考慮一個理論 T，與一個序數 γ，關於 γ 的序數分析要滿足以下的幾個條件：(1) T 可以形式化一個序數符號系統（當然只有

如果採取極端的柏拉圖主義，例如承認 ZFC 的意欲的模型存在，則心靈不是機器這件事立即確立。這是因爲 ZFC 在目前是公認的最強的數學理論，故任何機器系統所能證明的定理都可被 ZFC 證明，可是根據哥德爾第二個不完備性定理，它不能證明自身的一致性，但我們卻知道它是一致的。不過如同費弗曼所指出的，很少學者抱持這種極端的立場（潘若斯也不持這種立場，(Penrose, 1996: Section 9)）。可是這也不表示我們必須採取形式主義的立場，因爲我們似乎可以合理地採取某種程度的柏拉圖主義：我們或許會接受自然數算術理論的意欲的模型的存在，但不見得會接受集合論有所謂的意欲的模型存在。[37] 這是因爲我們直覺上可以清楚地「看出」或「想像」自然數算術結構是怎樣的一個抽象存在，但集合論的意

有限多的符號）來模擬小於 γ 的序數；(2) T⊢I(γ)→CON(T)，其中 I(γ) 表示 T 可以做到 γ 的歸納法（一般所謂的數學歸納法僅做到 ω，也就是僅證明所有<ω的情形都成立；超限歸納法是其延伸，也就是在更大的序數作歸納）；(3) T⊢I(β)，對任何β<γ。要注意的是，上述的序數分析並不蘊涵 T⊢CON(T)，而且機器可以模擬這分析（因為只牽涉到有限多的符號的運作）。見 Gentzen, 1969: 132-201。又見竹內外史，1987: 101-47。

[37] 根據筆者與幾位數學家就此交換意見所得到的印象是：在實際上，數學家的立場常常會是某種程度的柏拉圖主義。他們會採取這種立場的理由可能不是透過哲學反省而來，而是基於一種實際的考量，也就是採取這立場有助於清楚說明一些事情。另外，前文中曾提到蓋夫曼認為數學家會把他從事數學推論的架構的一致性視為理所當然，這似乎也暗示了某種柏拉圖主義，事實上，根據筆者對蓋夫曼的認識，他確實是採取某種柏拉圖主義的立場。

欲的模型到底是什麼樣子，我們實在很難有清楚的意象。[38] 在這情況下，我們並不需要訴諸更強的理論就知道 PA 是一致的，而且這種知道的方式是 PA 所無法模擬的。

以上的討論揭露了一個問題：就是我們對「一致」的定義會導致某種兩難的處境：採取語意的定義會促使我們走向極端的柏拉圖主義（要不然我們很難解釋什麼是一致的），但這是我們所不樂見的；採取語法的定義會讓我們無法保持對 PA 的證明機器的優勢，不過這也是我們所不願意的。筆者將試圖在其他的文章中詳細說明這種兩難的處境及可能的解決之道，但在此要點出的問題就是：我們如何定義「一致」這個概念其實也會對心靈是否爲機器這個議題有所影響。

伍、結論

首先，我們無法利用哥德爾的不完備性定理來證明心靈不是機器，因爲有兩個可能性是哥德爾的不完備性定理所無法排除的：(1) 面對一部複雜的機器，我們有可能無法證明它是一致的，也因而無法得知它的哥德爾語句是否爲真，因此不完備性定理並未保證盧卡斯等人所認爲的心靈相對於機器的優勢；(2) 有可能存在某機器能

[38] 我們知道 ZF 蘊涵了集合論的定義域會是由良基集所構成的類 (the class of well-founded sets)。但很少人宣稱他確信這樣的一個結構存在。事實上，若加入新的集合論公理，其定義域有可能被重構，故也很難說什麼是集合論的意欲的結構。見 Koellner (2003)。

證明心靈能證明的所有的自然數算術命題，但我們不知道它是如此（如果我們知道，則我們也會知道它是一致的，但它不能證明它自身是一致的），因此不完備性定理並未保證麥扣所認爲的心靈相對於機器的優勢。

但是不完備性定理確實駁斥了以下的兩個強的機器論的主張：(1)「我們知道」存在某種機器使得心靈所能證明的自然數算術命題正好是該機器能證明的自然數算術命題（這蘊涵了該機器是一致的）及 (2)「我們知道」某種機器是一致的而且心靈所能證明的自然數算術命題都是該機器能證明的。但若我們把上述兩個主張中的「我們知道」四個字拿掉，則所得到的兩個較弱的主張都與不完備性定理相容。

總之，我們無法利用不完備性定理來證明心靈不是機器，但我們也不能證明心靈是機器，因爲不完備性定理蘊涵了這樣的證明不會存在。因此在不完備性定理的脈絡下，對心靈是否爲機器這個問題，我們不可能有具備「數學確信度」的答案（筆者將此結論限定在「不完備性定理的脈絡下」是因爲有可能將來會發現新的後設邏輯定理，使得我們能證明心靈不是機器），不過，基於下文將提出的兩個理由，筆者認爲從哲學的觀點而言，機器論者必須負擔很大的舉證責任而這也使得機器論者的主張不容易成立。

首先，如同上一節所提到的，機器論者的負擔很大：他要論證心靈能製造出的所有數學證明都可以被形式化，然後又必須論證有機器能證明心靈所能證明的所有數學命題（這裏的論證指的當然是哲學上的論證而不是數學的證明）。而且我們似乎難以避免採取某種柏拉圖主義的立場，這件事使得機器論要成立的難度更大，因爲在此情況下，心靈有可能透過某種「直覺」來掌握到一些無法給予形式證明的數學真理。

再者，面對一部複雜的機器，心靈有可能透過經驗歸納的方式

來得知該機器是一致的。而這種類型的數學知識似乎是機器所難以獲得的，因爲機器的「知」即是「證明」。蓋夫曼認爲心靈的經驗歸納的能力不見得是機器所無法模擬的，因爲我們可以把很多機器聯結在一起，分享彼此所接受到的新資訊並且可能因此而改變預設的公理 (Gaifman, 2000: 468)。但問題是，哪一台機器會先透過歸納而改變呢? 如果沒有任何一台機器具有這種能力，它們聯結之後又如何能透過經驗歸納而改變任何公理呢？

對於心靈是否爲機器這個問題，或許有人僅期待明確的答案而對哲學的解決方式不感興趣，但在不完備性定理之外的相關的後設邏輯定理尚未被發現之前，我們充其量也只能做到這種地步。

參考文獻

Beaney, M. (ed.). 1997. *The Frege Reader*. Oxford: Blackwell Publishers.

Carnap, R. 1937. *The Logical Syntax of Language*. London: Routledge & Kegan Paul.

Chalmers, D.J. (1995). "Minds, Machines, and Mathematics: A Review of *Shadows of Mind* by Roger Penrose." *Psyche*, 2, 9. http://psyche.cs.monash.edu.au/v2/psyche-2-09-chalmers.html

Davis, M. 1993. "How Subtle Is Gödel's Theorem? More on Roger Penrose." *Behavioral and Brain Sciences* 16: 611-2.

Davis, M. (ed.). 2004. *The Undecidable*. New York: Dover Publications.

Enderton, H.B. 2001. *A Mathematical Introduction to Logic*. San Diego: Harcourt Press.

Feferman, S. 1988. "Hilbert's Program Relativized: Proof-theoretical and Foundational Reductions." *Journal of Symbolic Logic* 53: 364-84.

Feferman, S. 1995. "Penrose's Gödelian Argument: A Review of Shadows of the Mind by Roger Penrose." *Psyche*, 2, 7. http://psyche.cs.monash.edu.au/v2/psyche-2-07-feferman.html

Franzen, T. 2005. *Gödel's Theorem: An Incomplete Guide to Its Use and Abuse*. Wellesley: A K Peters.

Gaifman, H. 2000. "What Gödel's Incompleteness Result Does and Does Not Show." *The Journal of Philosophy* 2000: 462-70.

Gentzen, G. 1969. *The Collected Papers of Gerhard Gentzen*. M. E. Szabo (ed.). Amsterdam: North-Holland.

Gödel, K. 1990. *Collected Works Vol. II*. New York: Oxford University Press.

———. 1995. *Collected Works Vol. III*. New York: Oxford University Press.

Hardy, G. H. 1929. "Mathematical Proof." *Mind* 38: 1-25.

Hilbert, D. 1922. "The New Grounding of Mathematics." In Ewald, W.B. (ed.), 1996. *From Kant to Hilbert (vol.2)*. Oxford: Oxford University Press.

Koellner, P. 2003. *The Search for New Axioms*. PhD Dissertation, MIT.

Kunen, K. 1980. *Set Theory: An Introduction to Independence Proofs*. Amsterdam: North-Holland.

Lindstrom, P. 2001. "Penrose's New Argument." *Journal of Philosophical Logic* 30: 241-50.

Lucas J.R. 1961. "Minds, Machines and Gödel." *Philosophy 36*: 112-27.

———. 1996. "Minds, Machines and Gödel: A Retrospect." In P. J. R. Millican and A. Clark (eds.). *Machines and Thought: The Legacy of Alan Turing, Vol. 1*. Oxford: Oxford University Press.

McCall, S. 1999. "Can a Turing Machine Know that the Gödel Sen-

tence Is True?" *The Journal of Philosophy* 96: 525.

Monk, J.D. 1976. *Mathematical Logic*. New York: Springer-Verlag.

Penrose, R. 1989. *The Emperor's New Mind*. Oxford: Oxford University Press.

———. 1994. *Shadows of the Mind*. Oxford: Oxford University Press.

———. 1996. "Beyond the Doubting of a Shadow." *Psyche*, 2, 23. http://psyche.cs.monash.edu.au/v2/psyche-2-23-penrose.html

Shapiro, S. 2003. "Mechanism, Truth, and Penrose's New Argument." *Journal of Philosophical Logic* 32: 19-42.

Shoenfield, J.R. 1967. *Mathematical Logic*. Natick: Association for Symbolic Logic.

Soare, R.I. 1987. *Recursively Enumerable Sets and Degrees*. Berlin: Springer-Verlag.

Takeuti, Gaisi（竹內外史）. 1987. *Proof Theory*. Amsterdam: Elsevier.

Wang, Hao（王浩）. 1974. *From Mathematics to Philosophy*. London: Routledge and Kegan Paul.

Zach, R. 1998. "Numbers and Functions in Hilbert's Finitism." *Taiwanese Journal of Philosophy and History of Science* 10: 33-60.

Gödel's Incompleteness Theorems and the Mind-Machine Debate

Hsing-Chien Tsai

Department of Philosophy, National Chung-Cheng University

Abstract

Gödel's incompleteness theorems are the most important independent results ever discovered so far in the continuing development of modern logic. Recently in the literature several attempts have been made to use the incompleteness theorems to argue that no machine can fully capture what the human mind can do: for the incompleteness theorems seem to imply that, for any machine, there are always some mathematical propositions which the human mind can know to be true but which machine cannot prove. I shall look into the major arguments, both pro and con, that can be found in the literature and clarify some relevant concepts such as "machine", "proof" and "consistency". Then I shall argue that in the vein of the incompleteness theorems there is no answer with mathematical certitude to the mind-machine debate and that much of the philosophical burden of proof will lie with mechanists in this debate.

Keywords: Gödel, Incompleteness, Mind, Machine, Proof, Consistency