L&PS – Logic and Philosophy of Science
Vol. XII, No. 1, 2014

L&PS – Logic and Philosophy of Science
http://www2.units.it/~episteme
ISSN: 1826-1043

L&PS – Logic & Philosophy of Science
Vol. XII, No. 1, 2014, pp. 3-39

Testimonianze esperte e probabilità delle ipotesi

Roberto Festa
Dipartimento di Studi Umanistici
Università di Trieste
e-mail: festa@units.it

SOMMARIO. Nella pratica clinica e in quella giudiziaria un importante ruolo viene svolto dalle cosiddette testimonianze esperte, cioè dalle testimonianze degli esperti interpellati da medici e giudici nelle diverse fasi del processo diagnostico e di quello penale. Nella prima parte di questo articolo, introdurremo alcune nozioni fondamentali dell'approccio bayesiano all'epistemologia della testimonianza. Nella seconda parte, mostreremo che tale approccio può far luce sul modo in cui le testimonianze esperte dovrebbero governare la valutazione probabilistica delle ipotesi diagnostiche considerate dai medici e delle ipotesi ricostruttive prospettate nel processo penale.

PAROLE CHIAVE: epistemologia bayesiana, epistemologia della pratica clinica, epistemologia giudiziaria, testimonianza esperta, epistemologia della testimonianza.

L'epistemologia include diverse aree di ricerca, quali l'epistemologia generale, che si occupa dei problemi che si presentano in relazione a qualsiasi forma di conoscenza, e diverse epistemologie speciali, che si interessano ai problemi relativi a specifici campi del sapere. Un'importante epistemologia speciale è la filosofia della scienza che si occupa, ormai da alcuni secoli, dei problemi re-

Published by L&PS – Logic and Philosophy of Science
http://www2.units.it/~episteme – ISSN: 1826-1043

lativi alla natura, all'acquisizione e alla dinamica delle conoscenze scientifiche. Occorre attendere, invece, gli ultimi decenni del Novecento per assistere allo sviluppo di altre, e non meno importanti, epistemologie speciali, rivolte all'analisi di quelle che chiameremo *pratiche esperte*. Due notevoli esempi di pratiche esperte sono la *pratica clinica* e quella *giudiziaria*. Le pratiche esperte sono accomunate da un tratto distintivo: mentre l'acquisizione di nuove conoscenze non rientra tra i loro obiettivi, esse affrontano problemi per la cui soluzione si richiede un ampio ricorso alle più aggiornate conoscenze scientifiche. Tali pratiche vanno quindi distinte sia dalle attività conoscitive tipiche della vita quotidiana – condotte senza alcun ricorso ai risultati della scienza –, sia dalla ricerca scientifica, che è volta all'acquisizione di nuove conoscenze. Ne segue che gli interrogativi affrontati dall'epistemologia delle pratiche esperte sono diversi da quelli affrontati dalla filosofia della scienza. Si consideri, per esempio, la medicina, intesa come il campo teorico-pratico che include le scienze mediche e la pratica clinica. L'epistemologia della medicina comprende due epistemologie speciali, vale a dire la filosofia delle scienze mediche – che, a sua volta, costituisce una branca della filosofia della scienza –, e l'epistemologia della pratica clinica. Mentre la filosofia delle scienze mediche affronta la questione: "in che modo la ricerca medica può acquisire nuove conoscenze?", l'epistemologia della pratica clinica cerca di rispondere all'interrogativo: "in che modo le conoscenze rese disponibili dalle scienze mediche possono essere impiegate nella diagnosi e nella cura dei pazienti?".

In questa sede ci occuperemo di alcuni problemi epistemologici relativi alle due pratiche esperte sopra menzionate – la pratica clinica e quella giudiziaria. Più precisamente, considereremo il ruolo che, entro tali pratiche, viene svolto dalle testimonianze degli esperti, cioè da quelle che, sempre più frequentemente, vengono chiamate *testimonianze esperte*. Affronteremo questo tema dal punto di vista del cosiddetto *approccio bayesiano* alla razionalità, che ha conosciuto un impetuoso sviluppo nel corso del Novecento. Il presupposto fondamentale dell'approccio bayesiano consiste nella tesi che la conoscenza ha carattere probabilistico, cioè che la fiducia che un soggetto nutre nella verità di determinate ipotesi può venire espressa mediante appropriate probabilità. L'approccio bayesiano offre una teoria unitaria della razionalità cognitiva e pratica, cioè della formazione di opinioni e decisioni razionali. Infatti, i bayesiani non si limitano a descrivere la natura probabilistica della conoscenza, ma mostrano anche in che modo le nostre conoscenze, cioè le probabilità che attribuiamo alle ipotesi, possono venire usate nell'attuazione di decisioni razionali volte alla soluzione di problemi pratici. In particolare, i bayesiani ritengono che le probabilità debbano venire usate proprio a partire da quei luoghi, co-

me le corsie degli ospedali e le aule dei tribunali, in cui si prendono decisioni che spesso riguardano questioni di vita e di morte. Non sorprende, quindi, che l'approccio bayesiano venga sempre più ampiamente usato, soprattutto a partire dagli anni Ottanta dello scorso secolo, nell'epistemologia della pratica clinica e giudiziaria, caratterizzate entrambe dalla stretta combinazione di problemi cognitivi e pratici. Infatti, entro tali pratiche, la formazione di opinioni razionali costituisce un obiettivo preliminare in vista dell'attuazione di decisioni razionali. Così, per esempio, la formazione di opinioni diagnostiche razionali è un passaggio necessario all'attuazione di appropriate scelte terapeutiche. Allo stesso modo, nel processo penale, la formazione di opinioni razionali circa le ipotesi prospettate dall'accusa e dalla difesa è un obiettivo preliminare in vista della decisione finale, che può consistere in un verdetto di condanna oppure di assoluzione. Nelle pagine che seguono, concentreremo la nostra attenzione sul modo in cui la teoria bayesiana della razionalità cognitiva può far luce sulla formazione di opinioni razionali nella pratica clinica e giudiziaria, con particolare attenzione per le opinioni che medici e giudici si formano alla luce delle testimonianze esperte. Nel *primo* paragrafo introdurremo alcuni indispensabili elementi di epistemologia bayesiana. Nel *secondo* illustreremo i tratti fondamentali dell'epistemologia bayesiana della testimonianza, cioè dell'approccio bayesiano all'epistemologia della testimonianza. Infine, tale approccio verrà applicato all'analisi delle testimonianze – e, più specificamente, del loro ruolo nella determinazione della probabilità delle ipotesi – nella pratica clinica (*terzo* paragrafo) e giudiziaria (*quarto* paragrafo).

1. Elementi di epistemologia bayesiana

1.1. L'approccio bayesiano alla razionalità cognitiva

Nella comunicazione scientifica, nelle pratiche esperte e anche nella vita quotidiana, si dice che un'ipotesi è probabile quando si hanno buoni motivi per credere che sia vera, pur senza esserne certi. Talvolta si precisa il grado di probabilità di un'ipotesi H dicendo, per esempio, che H è molto probabile, o estremamente probabile. Può anche accadere che si attribuisca un valore quantitativo alla probabilità di H dicendo, per esempio, che la probabilità di H è pari al 99%. L'idea che si possano attribuire precisi valori quantitativi alle probabilità costituisce il nocciolo della teoria delle probabilità.

Consideriamo un insieme di enunciati $\mathbf{Z} \equiv \{A, B, C, ...\}$ tale che: (i) se $\mathbf{Z}$ contiene A, contiene anche la sua negazione $\neg A$; (ii) se $\mathbf{Z}$ contiene A e B, con-

tiene anche la loro congiunzione $A \,\&\, B$ e la loro disgiunzione $A \vee B$. I tre assiomi fondamentali della teoria della probabilità affermano che una funzione di probabilità $p(\bullet)$ definita su $\mathbf{Z}$ ha le seguenti proprietà:

(I) $0 \leq p(A) \leq 1$;
(II) Se A è una tautologia, allora $p(A) = 1$;
(III) Se è logicamente impossibile che la congiunzione $A \,\&\, B$ sia vera, allora $p(A \vee B) = p(A) + p(B)$.

Dagli assiomi (I)-(III) deriva che la probabilità $p(\neg A)$ della negazione di A è data da:

(1) $p(\neg A) = 1 - p(A)$.

Parlando di una corsa di cavalli, potremo formulare le ipotesi $A \equiv$ "Tartaruga vincerà la corsa" e $\neg A \equiv$ "Tartaruga non vincerà la corsa". A proposito di tali ipotesi, il teorema (1) ci permette di affermare, per esempio, che se la probabilità di A è 0,75, allora la probabilità di $\neg A$ è 0,25. Poiché, in questo caso, la probabilità di A è tre volte la probabilità di $\neg A$, possiamo dire, usando il linguaggio delle scommesse, che A viene *quotato* tre contro uno. Il concetto di *quota* di A – che corrisponde al termine inglese *odds* ed è quindi abitualmente indicato con "$o(A)$" – viene definito come il rapporto tra la probabilità di un enunciato e quella della sua negazione:[1]

$$(2) \qquad o(A) \equiv \frac{p(A)}{p(\neg A)} = \frac{p(A)}{1 - p(A)}.^{2}$$

La probabilità $p(A)$ attribuita a un enunciato $A \in \mathbf{Z}$ viene spesso chiamata *probabilità assoluta* di A. Analogamente, la quota $o(A)$ viene chiamata *quota assoluta* di A. Data una coppia ordinata (A, B) di membri di $\mathbf{Z}$, la cosiddetta probabilità *condizionata* di A dato B – indicata da "$p(A|B)$" – viene così definita:

$$(3) \qquad \text{Se } p(B) \neq 0 \text{ allora } p(A|B) \equiv \frac{p(A\&B)}{p(B)}.$$

[1] Il termine "quota" viene usato da diversi autori, tra il quali Mura (2003, p. XXIII).
[2] Si noti che $0 \leq o(A) \leq \infty$.

Possiamo ora introdurre, con una definizione strettamente simile alla (2), il concetto di quota condizionata di A dato B, indicato con "$o(A|B)$":

$$(4) \qquad o(A|B) \equiv \frac{p(A|B)}{p(\neg A|B)} = \frac{p(A|B)}{1 - p(A|B)}.$$

Dagli assiomi (I)-(III) e dalla definizione (3) derivano due semplici teoremi di cui faremo uso in seguito:

$$(5) \qquad p(A|A) = 1;$$
$$(6) \qquad p(\neg A|B) = 1 - p(A|B).$$

Un fondamentale teorema della teoria delle probabilità, noto come teorema di Bayes, concerne le relazioni tra la probabilità condizionata $p(A|B)$ e la probabilità assoluta $p(A)$. La più semplice versione del teorema è la seguente:

$$(7) \qquad p(A|B) = p(A) \times \frac{p(B|A)}{p(B)}.$$

Un'altra versione del teorema di Bayes – di cui faremo ampio uso nel seguito – viene formulata in termini di quote:

$$(8) \qquad o(A|B) = o(A) \times \frac{p(B|A)}{p(B|\neg A)}.$$

Il rapporto $p(B|A)/p(B|\neg A)$, che appare nel termine destro dell'uguaglianza (8), viene abitualmente chiamato *rapporto di verosimiglianza* (*likelihood ratio*) di A rispetto a B, e indicato con "$L(A,B)$". Questa notazione ci consente di riformulare (8) nel seguente modo:

$$(9) \qquad o(A|B) = o(A) \times L(A,B).$$

La teoria delle probabilità rappresenta, per così dire, il nocciolo matematico dell'approccio bayesiano alla razionalità.[3] Infatti, il principio fondamentale

[3] Per una dettagliata esposizione dell'approccio bayesiano alla razionalità si veda Festa (1996).

dell'approccio bayesiano – che potremmo chiamare principio di *conformità probabilistica*; in breve: (CP) – può venire così formulato:

(CP) In qualunque istante t, i gradi di credenza $p(A)$, $p(B)$, $p(C)$, ..., che esprimono le opinioni di un soggetto razionale X circa un insieme di enunciati $\mathbf{Z} \equiv \{A, B, C, ...\}$, sono probabilità, cioè obbediscono agli assiomi (I)-(III).

Le probabilità $p(A)$, $p(B)$, $p(C)$, ..., che esprimono le opinioni di un soggetto razionale vengono spesso chiamate probabilità *epistemiche*. Va osservato che (CP) è un principio *statico*, poiché non pone alcun limite alla *cinematica delle opinioni*, cioè alla variabilità temporale delle probabilità epistemiche che un soggetto razionale può attribuire, in momenti diversi, agli enunciati di $\mathbf{Z}$. Tuttavia, i bayesiani ritengono che la cinematica delle opinioni debba venire regolata da appropriati principi. Date due funzioni di probabilità $p(\cdot)$ e $p_n(\cdot)$, definite su $\mathbf{Z}$, che esprimono, rispettivamente, le "vecchie" opinioni di un soggetto razionale X nell'istante t e le sue "nuove" opinioni in un successivo istante t', il passaggio da $p(\cdot)$ a $p_n(\cdot)$ deve essere effettuato, secondo i bayesiani, in accordo con un principio cinematico noto come principio di *condizionalizzazione* – in breve: (Co) – che può venire così formulato:

(Co) Se le uniche informazioni acquisite da X fra gli istanti t e t' sono costituite dall'accertamento dell'enunciato $E \in \mathbf{Z}$ allora, per qualunque enunciato $H \in \mathbf{Z}$, la vecchia probabilità assoluta $p(H)$ deve essere sostituita con una nuova probabilità assoluta $p_n(H)$ determinata in accordo con la seguente *regola di condizionalizzazione* (RC):

(RC) $p_n(H) = p(H|E)$.

Risulta spesso naturale considerare l'istante t, in cui viene assegnata la vecchia probabilità $p(H)$, come l'istante iniziale di un'indagine e l'istante t', in cui viene assegnata la nuova probabilità $p(H|E)$, come l'istante finale dell'indagine. Di conseguenza, $p(H)$ e $p(H|E)$ vengono spesso chiamate, rispettivamente, probabilità *iniziale* e probabilità *finale* di H. Occorre notare che l'ambito di applicazione di (Co) è piuttosto ristretto. Infatti (Co) si applica solo ai casi in cui l'aggiornamento della vecchia funzione di probabilità $p(\cdot)$ viene effettuato in risposta all'acquisizione di un'*evidenza certa*, cioè in risposta all'accertamento di un enunciato $E \in \mathbf{Z}$. Tuttavia, (Co) non dice nulla su come un soggetto razionale dovrebbe aggiornare la sua vecchia funzione di probabilità $p(\cdot)$ nel-

le situazioni, tutt'altro che insolite, in cui viene acquisita un'evidenza che non presenta alcun carattere di certezza.

La ricerca epistemologica ha individuato molti tipi di evidenza incerta tra i quali, ai nostri fini, basterà considerarne uno, cioè l'evidenza ambigua. Un soggetto X acquisisce un'*evidenza ambigua* $p_n(E)$ circa la coppia di enunciati E e $\neg E$ quando si verificano le seguenti condizioni: (i) le sue vecchie probabilità $p(E)$ e $p(\neg E)$ (= $1 - p(E)$) vengono sostituite da nuove probabilità $p_n(E)$ e $p_n(\neg E)$ (= $1 - p_n(E)$), entrambe diverse dai valori estremi 1 e 0, (ii) tale sostituzione viene effettuata in *risposta diretta* a determinati input sensoriali o a informazioni di altro genere acquisite da X. Le evidenze ambigue vengono spesso ottenute nelle cosiddette *osservazioni a lume di candela*, che possiamo illustrare con il seguente esempio.

ESEMPIO 1. ROSE INTRAVISTE ALLA LUCE DI UNA CANDELA. Tornando a casa tardi dal lavoro, Xavier trova sul tavolo del salotto, illuminato dalla sola luce di una candela, un mazzo di rose, ma non riesce a distinguere con certezza il loro colore. Potrebbe tuttavia accadere che, in risposta alla sua osservazione a lume di candela, egli sia in grado di attribuire all'enunciato $E \equiv$ "Le rose sul tavolo del salotto sono rosse" una nuova probabilità $p_n(E) = 2/3$.

Fino a pochi decenni or sono, i bayesiani non avrebbero saputo dare alcuna indicazione su come un soggetto razionale dovrebbe aggiornare le proprie probabilità in risposta a evidenze ambigue. Ai nostri giorni, invece, molti bayesiani sarebbero pronti ad affermare – sulla scia di Richard Jeffrey (1965/1983)[4] – che tale aggiornamento dovrebbe essere effettuato sulla base del seguente principio cinematico, noto come principio di *condizionalizzazione generalizzata* – in breve (CoG):

(CoG) Se le uniche informazioni acquisite da X fra gli istanti t e t' sono costituite dall'evidenza ambigua $p_n(E)$ allora, per qualunque enunciato $H \in \mathbf{Z}$, la vecchia probabilità assoluta $p(H)$ deve essere sostituita con una nuova probabilità assoluta $p_n(H)$ determinata in accordo con la seguente *regola di condizionalizzazione generalizzata* (RCG):

(RCG) $p_n(H) = p(H|E) \times p_n(E) + p(H|\neg E) \times p_n(\neg E)$.[5]

[4] Si veda anche Jeffrey (1992; 2004).

[5] La regola (RCG) identifica la nuova probabilità assoluta $p_n(H)$ con la media ponderata delle vecchie probabilità condizionate $p(H|E)$ e $p(H|\neg E)$, ove i pesi sono dati, rispettivamente,

1.2. Conferma delle ipotesi

Supponiamo che, in risposta a una nuova evidenza, un soggetto X aggiorni le sue probabilità sulla base di appropriati principi cinematici. Converrà denominare la nuova evidenza acquisita da X – non importa, per il momento, precisare se si tratta di evidenza certa o incerta – con "*Ev*". Se concentriamo l'attenzione su una determinata ipotesi H e confrontiamo la vecchia probabilità $p(H)$ con la nuova probabilità $p_n(H)$, che X attribuisce ad H in risposta a *Ev*, possiamo valutare l'impatto che *Ev* ha avuto sulla fiducia di X nella verità di H. A tale scopo viene comunemente usata la nozione *qualitativa* di conferma, che viene così definita:

(10) *Ev* conferma $H \equiv p_n(H) > p(H)$.

Secondo la (10), *Ev* conferma H nel caso in cui $p_n(H)$ è maggiore di $p(H)$, cioè nel caso in cui, in risposta a *Ev*, X accresce la sua fiducia nella verità di H. Le intuizioni alla base di (10) hanno suggerito anche l'introduzione di svariate nozioni *quantitative* di conferma. In particolare, le cosiddette misure *incrementali* di conferma identificano il grado di conferma di H da parte di *Ev* con una determinata quantità $c(H,Ev)$ che dipende solo da $p(H)$ e $p_n(H)$ e, per qualsiasi valore di $p(H)$, cresce al crescere di $p_n(H)$. Una semplice misura incrementale è data dal rapporto $p_n(H)/p(H)$ tra nuova e vecchia probabilità di H. Un'altra misura incrementale, nota come *fattore di Bayes*, è data dal rapporto fra la nuova quota $o_n(H)$ – definita, nel modo usuale, come $o_n(H) \equiv p_n(H)/(1 - p_n(H))$ – e la vecchia quota $o(H)$:

$$(11) \quad c_B(H,Ev) = \frac{o_n(H)}{o(H)}.$$ [6]

È interessante notare che $c_B(H,Ev)$ è connessa alla nozione qualitativa di conferma (10) dalla seguente relazione:

---

da $p_n(E)$ e $p_n(\neg E)$. La recente ricerca epistemologica ha messo in luce che le evidenze ambigue sopra descritte non sono l'unico genere di evidenza incerta che un soggetto può acquisire. Di conseguenza, i principi cinematici (Co) e (CoG) – applicabili, rispettivamente, alle evidenze certe e ambigue –, devono essere integrati da ulteriori principi cinematici, in grado di indicare come si dovrebbero aggiornare le proprie probabilità in risposta a svariati tipi di evidenze incerte. A questo riguardo si veda Festa (1996, cap. 4).

[6] Si noti che $0 \leq c_B(H,Ev) \leq \infty$.

(12) *Ev* conferma *H* se e solo se $c_B(H,Ev) > 1$.

Ai nostri fini è di particolare interesse considerare l'applicazione del fattore di Bayes $c_B(H,Ev)$ nel caso particolare in cui la nuova evidenza *Ev* acquisita da *X* è costituita da un'evidenza certa *E*. In tal caso la nuova probabilità $p_n(H)$ di *X* dovrà essere determinata applicando il principio cinematico (Co) che richiede di porre $p_n(H) = p(H|E)$. Ciò equivale a porre $o_n(H) = o(H|E)$, cioè a identificare la nuova quota $o_n(H)$ di *X* con la sua vecchia quota condizionata $o(H|E)$. Dalla definizione (11) e dall'uguaglianza $o_n(H) = o(H|E)$ segue che la conferma $c_B(H,E)$ apportata ad *H* da *E* è data da:

$$(13) \quad c_B(H,E) = \frac{o(H|E)}{o(H)}.$$

Le uguaglianze (8) e (9), che esprimono il teorema di Bayes in termini di quote, ci dicono che $o(H|E) = o(H) \times L(H,E) = o(H) \times p(E|H)/p(E|\neg H)$. Ne segue che $o(H|E)/o(H) = L(H,E) \equiv p(E|H)/p(E|\neg H)$ e quindi, in base alla (13), che:

$$(14) \quad c_B(H,E) = L(H,E) = \frac{p(E|H)}{p(E|\neg H)}.$$

La formula (14) ci dice che il grado di conferma $c_B(H,E)$ è identico al rapporto di verosimiglianza $L(H,E) \equiv p(E|H)/p(E|\neg H)$, cioè al rapporto tra la probabilità di *E* alla luce di *H* e la probabilità di *E* alla luce di $\neg H$. L'uguaglianza $c_B(H,E) = L(H,E)$ ci permette di riformulare nel seguente modo il teorema di Bayes (9) per le quote:

$$(15) \quad o(H|E) = o(H) \times c_B(H,E).$$

Il contenuto intuitivo di (15) può venire espresso dicendo che la quota finale di un'ipotesi *H* alla luce di *E* è pari al prodotto della sua quota iniziale e del grado di conferma che *E* apporta ad *H*.

1.3. Evidenze consonanti, indirette e ambigue

Vedremo ora in che modo le ipotesi possono venire confermate da due tipi di evidenze certe – cioè le evidenze consonanti e quelle indirette – e dalle evidenze ambigue.

Evidenze consonanti. Diremo che *E* ed *E** sono evidenze *consonanti a favore*

dell'ipotesi H se ciascuna di esse conferma H o, equivalentemente, se $c_B(H,E)$, $c_B(H,E^*) > 1$ (vedi (12)). Possiamo ora chiederci a quali condizioni anche la congiunzione E & E^* conferma H, cioè a quali condizioni $c_B(H,E \,\&\, E^*) > 1$. Prima di rispondere a questo interrogativo occorre introdurre il concetto di *separazione*.

Ciascuna delle coppie di enunciati $\mathbf{H} \equiv (H,\neg H)$, $\mathbf{E} \equiv (E,\neg E)$ ed $\mathbf{E^*} \equiv (E^*,\neg E^*)$ costituisce una variabile con due possibili valori: per esempio, i possibili valori di **H** sono H e $\neg H$. Diremo allora che:

(16) **H** separa **E*** da **E** se e solo se valgono le seguenti condizioni:

(i) $p(E^*|H \,\&\, E) = p(E^*|H \,\&\, \neg E) = p(E^*|H)$;
(ii) $p(E^*|\neg H \,\&\, E) = p(E^*|\neg H \,\&\, \neg E) = p(E^*|\neg H)$.

Ciò significa che **H** separa **E*** da **E** nel caso in cui la conoscenza del valore di **H** – cioè il fatto di sapere se H è vera o falsa – rende irrilevante l'ulteriore conoscenza del valore di **E** per determinare la probabilità di E^*. Occorre notare che la relazione di separazione è simmetrica nel senso che, se **H** separa **E*** da **E**, allora separa anche **E** da **E***.[7] Si può inoltre dimostrare che (16) equivale alla seguente definizione:

(17) **H** separa **E*** da **E** se e solo se valgono le seguenti condizioni:

(i) $p(E \,\&\, E^*|H) = p(E|H) \times p(E^*|H)$;
(ii) $p(E \,\&\, E^*|\neg H) = p(E|\neg H) \times p(E^*|\neg H)$.

Siamo ora in grado di rispondere all'interrogativo circa le condizioni in cui un'ipotesi H, confermata da ciascuna delle evidenze E ed E^*, viene confermata anche dalla loro congiunzione E & E^*. Infatti segue da (17) e dall'uguaglianza $o(H|E) = o(H) \times c_B(H,E)$ (vedi (15)) che:

(18) Se **H** separa **E*** da **E** allora $c_B(H,E \,\&\, E^*) = c_B(H,E) \times c_B(H,E^*)$.[8]

A sua volta, (18) implica che:

[7] La relazione "**H** separa **E*** da **E**" potrebbe quindi venire espressa, con un'espressione leggermente più lunga ma più precisa, con "**H** separa **E** ed **E*** l'una dall'altra".

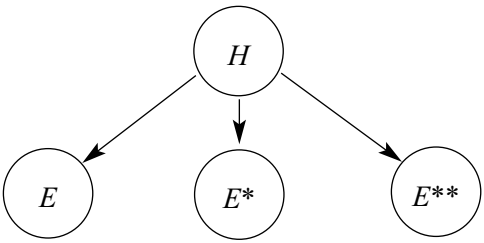
[8] Infatti $c_{or}(H,E \,\&\, E^*) \equiv p(E \,\&\, E^*|H)/p(E \,\&\, E^*|\neg H)$ (per (14)) = $[p(E|H) \times p(E^*|H)]/[p(E|\neg H) \times p(E^*|\neg H)]$ (per (17)) = $c_{or}(H,E) \times c_{or}(H,E^*)$ (per (14)).

(19) Se E ed E^* sono evidenze consonanti a favore di H e, inoltre, **H** separa **E*** da **E**, allora $c_B(H,E \,\&\, E^*) > c_B(H,E)$, $c_B(H,E^*) > 1$.

Ciò significa che, se l'antecedente di (19) è vero, allora non solo la congiunzione $E \,\&\, E^*$ conferma H, ma tale conferma è maggiore di quella apportata da ciascuno dei congiunti E ed E^*.

Le nozioni appena illustrate possono venire estese al caso di tre o più evidenze. Per esempio, diremo che E, E^* ed E^{**} sono evidenze consonanti a favore di H se $c_B(H,E)$, $c_B(H,E^*)$, $c_B(H,E^{**}) > 1$. Con riferimento alle variabili $\mathbf{H} \equiv (H,\neg H)$, $\mathbf{E} \equiv (E,\neg E)$, $\mathbf{E^*} \equiv (E^*,\neg E^*)$ ed $\mathbf{E^{**}} \equiv (E^{**},\neg E^{**})$, possiamo introdurre una naturale generalizzazione del concetto di separazione definito nella (16). Tale generalizzazione può venire illustrata, in termini intuitivi, come segue: **H** separa ciascuna delle variabili **E**, **E***, ed **E**** dalle altre due nel caso in cui, data la conoscenza del valore di **H**, l'ulteriore conoscenza del valore di due delle variabili **E**, **E***, ed **E**** è irrilevante per determinare la probabilità dei possibili valori della terza. Se **H** separa ciascuna delle variabili **E**, **E***, ed **E**** dalle altre due, allora le relazioni tra **H**, **E**, **E***, ed **E**** possono venire rappresentate nella Figura 1, dove il fatto che **E**, **E***, ed **E**** non siano direttamente collegate fra loro da alcuna freccia significa che **H** le separa l'una dalle altre:[9]

FIG. 1: Rete bayesiana con connessioni divergenti



Possiamo ora generalizzare i teoremi (18) e (19) al caso di tre evidenze consonanti:

(20) Se **H** separa ciascuna delle variabili **E**, **E*** ed **E**** dalle altre due, allora $c_B(H,E \,\&\, E^* \,\&\, E^{**}) = c_B(H,E) \times c_B(H,E^*) \times c_B(H,E^{**})$.

(21) Se E, E^* ed E^{**} sono evidenze consonanti a favore di H e, inoltre, **H** separa ciascuna delle variabili **E**, **E*** ed **E**** dalle altre due, allora $c_B(H,E \,\&\, E^* \,\&\, E^{**}) > c_B(H,E)$, $c_B(H,E^*)$, $c_B(H,E^{**}) > 1$.

[9] Il lettore che ha qualche familiarità con le reti bayesiane si sarà accorto che la Figura 1 è un esempio di rete bayesiana con connessioni divergenti: cfr. Taroni *et al*. (2006, p. 40).

Evidenze indirette. La nozione di evidenza indiretta può venire illustrata con il seguente esempio.

ESEMPIO 2. EVIDENZA INDIRETTA SULLA PRESENZA DEL NOTO PIROMANE FUEGO. Xavier vede del fumo uscire dalla boscaglia. Questa osservazione (E^*) conferma l'enunciato E il quale afferma che c'è un incendio nella boscaglia. A sua volta E conferma l'ipotesi H che da quelle parti è passato il noto piromane Fuego. Poiché E^* conferma E che, a sua volta, conferma H, diremo che E^* è un'*evidenza indiretta* a favore di H.

Nell'Esempio 2 le nostre intuizioni suggeriscono che H viene confermata dall'evidenza indiretta E^*. Tuttavia vi sono molti altri casi in cui E^* conferma E ed E conferma H, ma E^* *non* conferma H. Ciò significa che il principio di transitività della conferma non è universalmente applicabile.[10] Dobbiamo quindi chiederci a quali condizioni un'ipotesi viene confermata da un'evidenza indiretta a suo favore. Come vedremo qui sotto, il concetto di separazione, definito nella (16), ci consente di rispondere a questo interrogativo.

Date le variabili $\mathbf{H} \equiv (H,\neg H)$, $\mathbf{E} \equiv (E,\neg E)$ ed $\mathbf{E^*} \equiv (E^*,\neg E^*)$, si può dimostrare che:

(22) Se $\mathbf{E}$ separa $\mathbf{E^*}$ da $\mathbf{H}$, allora $p(H|E^*)$ è data da:

$$p(H|E^*) = p(H|E) \times p(E|E^*) + p(H|\neg E) \times p(\neg E|E^*).$$

Il teorema (22) ci mostra che, nel caso in cui $\mathbf{E}$ separa $\mathbf{E^*}$ da $\mathbf{H}$, $p(H|E^*)$ può venire calcolata sulla base delle probabilità $p(H|E)$, $p(H|\neg E)$ e $p(E|E^*)$.[11] Un'interessante conseguenza di (22) è la seguente:

(23) Se E^* conferma E ed E conferma H e, inoltre, $\mathbf{E}$ separa $\mathbf{E^*}$ da $\mathbf{H}$, allora $c_B(H,E) \geq c_B(H,E^*) > 1$.

Ciò significa che, se l'antecedente di (23) è vero, allora l'evidenza indiretta E^* conferma H, anche se – proprio come ci aspetteremmo – la conferma apporta-

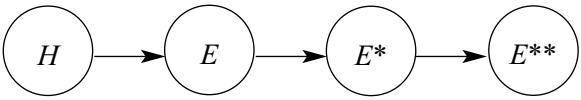
[10] Un esempio nel quale la conferma non sembra affatto transitiva è il seguente. Dati gli enunciati $E^* \equiv$ "Giovanni è stato accoltellato", $E \equiv$ "Giovanni è morto" e $H \equiv$ "Giovanni ha avuto un ictus", si può facilmente ammettere che E^* conferma E ed E conferma H, mentre E^* non conferma H.

[11] Non abbiamo menzionato la probabilità $p(\neg E|E^*)$, che appare nella (22), poiché il suo valore è determinato da quello di $p(E|E^*)$; infatti, segue da (1) che $p(\neg E|E^*) = 1 - p(E|E^*)$.

ta da E^* ad H non supera quella che viene apportata ad H dall'evidenza "diretta" E.

La nozione di evidenza indiretta può venire generalizzata al caso in cui un enunciato E^{**} conferma un'ipotesi H attraverso una catena formata da due o più anelli intermedi. Per esempio, date le variabili $\mathbf{H} \equiv (H,\neg H)$, $\mathbf{E} \equiv (E,\neg E)$, $\mathbf{E^*} \equiv (E^*,\neg E^*)$ ed $\mathbf{E^{**}} \equiv (E^{**},\neg E^{**})$, possiamo considerare il caso – rappresentato nella Figura 2 – in cui $\mathbf{E^*}$ separa $\mathbf{E^{**}}$ da $\mathbf{E}$ ed $\mathbf{E}$ separa $\mathbf{E^*}$ da $\mathbf{H}$.[12]

FIG. 2: Rete bayesiana con connessioni seriali



Si può dimostrare che in questo caso vale anche la seguente relazione di separazione: $\mathbf{E}$ separa $\mathbf{E^{**}}$ da $\mathbf{H}$. Varrà quindi, per il teorema (22), l'uguaglianza:

(24) $p(H|E^{**}) = p(H|E) \times p(E|E^{**}) + p(H|\neg E) \times p(\neg E|E^{**})$.

Il teorema (24) ci mostra che $p(H|E^{**})$ può venire determinata sulla base delle probabilità $p(H|E)$, $p(H|\neg E)$ e $p(E|E^{**})$.[13] A sua volta l'ultima di queste probabilità, cioè $p(E|E^{**})$, può venire facilmente determinata sulla base di $p(E|E^*)$, $p(E|\neg E^*)$ e $p(E^*|E^{**})$; infatti, poiché abbiamo supposto che $\mathbf{E^*}$ separa $\mathbf{E^{**}}$ da $\mathbf{E}$, il teorema (22) ci permette di ottenere l'uguaglianza:

(25) $p(E|E^{**}) = p(E|E^*) \times p(E^*|E^{**}) + p(E|\neg E^*) \times p(\neg E|{}^*E^{**})$.

Le uguaglianze (24) e (25) ci mostrano che $p(H|E^{**})$ può venire determinata sulla base delle probabilità $p(H|E)$, $p(H|\neg E)$, $p(E|E^*)$, $p(E|\neg E^*)$ e $p(E^*|E^{**})$. Tali uguaglianze consentono, inoltre, di dimostrare la seguente, del tutto naturale, generalizzazione di (23):

(26) Se E^{**} conferma E^*, E^* conferma E ed E conferma H e, inoltre, $\mathbf{E^*}$ separa $\mathbf{E^{**}}$ da $\mathbf{E}$ ed $\mathbf{E}$ separa $\mathbf{E^*}$ da $\mathbf{H}$, allora $c_B(H,E) \geq c_B(H,E^*) \geq c_B(H,E^{**}) > 1$.

[12] Il lettore che ha qualche familiarità con le reti bayesiane si sarà accorto che la Figura 1 è un esempio di rete bayesiana con connessioni seriali: cfr. Taroni *et al*. (2006, p. 40).

[13] Non abbiamo menzionato la probabilità $p(\neg E|E^*)$, che appare nella (24), poiché il suo valore è determinato dall'uguaglianza $p(\neg E^*|E^{**}) = 1 - p(E^*|E^{**})$ (vedi (6)).

Evidenze ambigue. Vediamo ora in che modo le ipotesi possono venire confermate dalle evidenze ambigue. Supponiamo che la nuova evidenza Ev acquisita da X sia costituita dall'evidenza ambigua $p_n(E)$. In tal caso la nuova probabilità $p_n(H)$ che X dovrebbe attribuire all'ipotesi H dovrà essere determinata applicando il principio cinematico (CoG) che richiede di porre $p_n(H) = p(H|E) \times p_n(E) + p(H|\neg E) \times p_n(\neg E)$. Per stabilire se $p_n(E)$ conferma una determinata ipotesi H – cioè se $p_n(H) > p(H)$ –, occorre confrontare la nuova probabilità $p_n(H)$ con la vecchia probabilità $p(H)$. In particolare, ai nostri fini è interessante notare che dalla definizione (10) segue che $p_n(E)$ conferma E se e solo se $p_n(E) > p(E)$. Dobbiamo ora chiederci, più in generale, a quali condizioni un'ipotesi H viene confermata da $p_n(E)$. Una risposta a questo interrogativo viene fornita dal seguente teorema:

(27) Se $p_n(E)$ conferma E ed E conferma H, allora $c_B(H,E) > c_B(E,p_n(E)) > 1$.[14]

Ciò significa che, se l'antecedente di (27) è vero, allora $p_n(E)$ conferma H, anche se – proprio come ci aspetteremmo – la conferma apportata ad H dall'evidenza ambigua $p_n(E)$ è inferiore quella apportata dall'evidenza certa E.

2. Epistemologia bayesiana della testimonianza

2.1. Come la probabilità delle ipotesi dipende dall'attendibilità delle testimonianze

Diciamo che una persona testimonia un enunciato H quando esprime la propria credenza nella verità di H. Possiamo intendere "testimonianza" in un senso abbastanza ampio da includere non solo le testimonianze nelle quali ci imbattiamo nella vita quotidiana, ma anche le testimonianze scritte, le notizie apparse sui giornali e altri media, e così via. Crediamo nella verità di un enunciato H sulla base di una testimonianza ogni volta che la nostra fiducia in H si basa, in qualche modo, sul fatto che qualcuno testimonia H. Sarebbe difficile sottovalutare il ruolo della testimonianza per l'acquisizione delle nostre conoscenze. Ci basiamo sulla testimonianza non solo per le conoscenze necessarie alla vita di ogni giorno – per esempio, per sapere dove si trova il calzolaio più vicino al nostro albergo –, ma anche per l'acquisizione delle conoscenze scientifiche. Infatti gli scienziati formano le loro conoscenze di base attraverso le te-

[14] Vedi Crupi, Festa e Mastropasqua (2008).

stimonianze di insegnanti e manuali e, successivamente, vengono informati dei risultati sperimentali ottenuti nei laboratori di tutto il mondo dalle testimonianze degli sperimentatori, riportate in articoli di riviste scientifiche. Non deve quindi stupire che la testimonianza sia stata tradizionalmente considerata come una delle fonti della conoscenza. È invece piuttosto sorprendente che, almeno fino alla metà del secolo scorso, il ruolo epistemico della testimonianza non sia stato fatto oggetto di alcuna analisi approfondita.[15] Tuttavia, negli ultimi trent'anni la situazione è rapidamente cambiata, cosicché possiamo oggi disporre di una vasta letteratura sull'epistemologia della testimonianza.[16] Le recenti ricerche sul tema sono state condotte sia nell'ambito dell'epistemologia generale sia in quello dell'epistemologia delle pratiche esperte e, in particolare, dell'epistemologia della pratica giudiziaria. Molte fra queste ricerche sono state effettuate nella prospettiva dell'approccio bayesiano.

L'analisi bayesiana della testimonianza deve mostrare come essa contribuisca a determinare i nostri gradi di credenza in determinati enunciati.[17] Le testimonianze possono apparirci più o meno attendibili cosicché, di fronte a una testimonianza relativa all'ipotesi H, il nostro grado di credenza in H verrà necessariamente influenzato dalla nostra valutazione dell'attendibilità di tale testimonianza. Indichiamo con "$T(H)$" la testimonianza con la quale un individuo T esprime la propria credenza nella verità di H, con "$A(T(H))$" un'appropriata misura dell'attendibilità che un soggetto razionale X attribuisce a $T(H)$ e, infine, con "$p(H|T(H))$" la probabilità che X deve attribuire ad H alla luce di $T(H)$.

Diversi autori hanno definito $A(T(H))$ come il rapporto tra la probabilità che T testimoni H nel caso in cui H è vera e la probabilità che testimoni H nel caso in cui è falsa:[18]

[15] Dopo essere stata, per circa duemila anni, la Cenerentola dell'epistemologia, la testimonianza entrò nell'esclusivo club dei problemi filosofici "rispettabili" attorno alla metà del Settecento, grazie all'opera di David Hume (1748, cap. X) e Thomas Reid (1764), i quali delinearono due diverse concezioni del ruolo della testimonianza nella formazione e giustificazione delle nostre conoscenze. Tuttavia le loro tesi sull'argomento non suscitarono particolare interesse tra i loro contemporanei e non riuscirono a stimolare, neppure nei due secoli successivi, alcuna indagine sistematica. Sulle vicissitudini, e la scarsa fortuna, dell'epistemologia della testimonianza nella storia del pensiero filosofico, si veda Vassallo (2003, pp. 24-32).

[16] Due corpose monografie sulla natura e il ruolo epistemico della testimonianza si devono a Coady (1992) e Shapin (1994). Si vedano anche la raccolta di saggi curata da Lackey e Sosa (2006) e le eccellenti rassegne di Adler (2008) e Kusch e Lipton (2002).

[17] Su tale argomento si vedano Goldman (1999, cap. 4.2-4.4) ed Earman (2000).

[18] Questa definizione di attendibilità è stata adottata sia nell'ambito dell'epistemologia generale (si vedano, fra gli altri, Goldman 1999 ed Earman 2000) sia in quello dell'epistemologia giudiziaria (si vedano, fra gli altri, Dawid 1987 e Mura 2004).

$$(28) \quad A(T(H)) \equiv \frac{p(T(H)|H)}{p(T(H)|\neg H)} \equiv L(H,T(H)).$$

Dalla definizione (28), assieme ai teoremi (14) e (15), segue che l'attendibilità di $T(H)$ è uguale alla conferma che $T(H)$ apporta ad H:

$$(29) \quad A(T(H)) = c_B(H,T(H)) = \frac{o(H|T(H))}{o(T(H))}.$$

Il teorema (29) suggerisce, fra l'altro, una naturale definizione del concetto qualitativo di *testimonianza attendibile*. Possiamo infatti dire che:

(30) $T(H)$ è attendibile nel caso in cui $A(T(H)) = c_B(H,T(H)) > 1$,

cioè nel caso in cui $T(H)$ conferma H. Dalle uguaglianze (29) e (15) segue che la quota finale di H alla luce di $T(H)$ è pari alla sua quota iniziale moltiplicata per l'attendibilità di $T(H)$:

$$(31) \quad o(H|T(H)) = o(H) \times c_B(H,T(H)) = o(H) \times A(T(H)).$$

Il contenuto intuitivo di (31) può venire espresso dicendo, in pieno accordo con le nostre intuizioni, che la fiducia di X nella verità di H, alla luce di $T(H)$, cresce al crescere della sua fiducia iniziale nella verità di H e dell'attendibilità da lui attribuita a $T(H)$.

2.2. Testimonianze consonanti, indirette e conflittuali

La testimonianza $T(H)$ con la quale un individuo T esprime la propria credenza nella verità di H può essere vista come una testimonianza "diretta" a favore di H, nel senso che essa riguarda direttamente H. Vedremo ora che un'ipotesi può venire confermata anche da appropriate combinazioni di testimonianze dirette, oppure da una testimonianza indiretta, relativa a un enunciato che a sua volta conferma H.

Testimonianze consonanti. Supponiamo che due testimoni T e T^* affermino di credere nella verità di H. Se le loro testimonianze $T(H)$ e $T^*(H)$ sono attendibili, cioè se entrambe confermano H (vedi (30)), diremo che sono *testimo-*

nianze consonanti a favore di H. Tali testimonianze costituiscono evidenze consonanti a favore di H, nel senso illustrato all'inizio del paragrafo 1.3. Inoltre, applicando la definizione (16) alle coppie di enunciati $\mathbf{H} \equiv (H,\neg H)$, $\mathbf{T(H)} \equiv (T(H),\neg T(H))$ e $\mathbf{T^*(H)} \equiv (T^*(H),\neg T^*(H))$, possiamo dire che:

(32) **H** separa **T*(H)** da **T(H)** se e solo se valgono le condizioni:

(i) $p(T^*(H)|H \,\&\, T(H)) = p(T^*(H)|H \,\&\, \neg T(H)) = p(T^*(H)|H)$;
(ii) $p(T^*(H)|\neg H \,\&\, T(H)) = p(T^*(H)|\neg H \,\&\, \neg T(H)) = p(T^*(H)|\neg H)$.

Ciò significa che **H** separa **T*(H)** da **T(H)** nel caso in cui la conoscenza del valore di **H**, cioè il fatto di sapere se H è vera o falsa, rende irrilevante l'ulteriore conoscenza del valore di **T(H)** – cioè del fatto che T abbia testimoniato, oppure no, H –, per determinare la probabilità che T^* testimoni H.

Se T e T^* hanno fornito le testimonianze $T(H)$ e $T^*(H)$ e, inoltre, **H** separa **T*(H)** da **T(H)**, diremo che $T(H)$ e $T^*(H)$ sono *testimonianze indipendenti a favore* di H. Dai teoremi (18), (19) e (29) segue che:

(33) Se $T(H)$ e $T^*(H)$ sono testimonianze indipendenti a favore di H, allora $c_B(H,T(H) \,\&\, T^*(H)) = c_B(H,T(H)) \times c_B(H,T^*(H)) = A(T(H)) \times A(T^*(H))$.

(34) Se $T(H)$ e $T^*(H)$ sono testimonianze indipendenti e consonanti a favore di H, allora $c_B(H,T(H) \,\&\, T^*(H)) > c_B(H,T(H))$, $c_B(H,T^*(H)) > 1$.

Il teorema (33) afferma che la conferma apportata ad H dalla congiunzione delle testimonianze indipendenti $T(H)$ e $T^*(H)$ è pari al prodotto dei gradi di attendibilità di tali testimonianze. Il teorema (34) afferma, invece, che la conferma apportata ad H dalla congiunzione delle testimonianze indipendenti e consonanti $T(H)$ e $T^*(H)$ è maggiore di quella apportata da ciascuna di esse presa isolatamente.

Testimonianze indirette. Nel paragrafo 1.3 abbiamo introdotto la nozione di evidenza indiretta: E^* è un'evidenza indiretta a favore dell'ipotesi H nel caso in cui E^* conferma un enunciato E che a sua volta conferma H. Abbiamo poi dimostrato (vedi (23)) che, se E^* è un'evidenza indiretta a favore di H e, inoltre, $\mathbf{E} \equiv (E,\neg E)$ separa $\mathbf{E^*} \equiv (E^*,\neg E^*)$ da $\mathbf{H} \equiv (H,\neg H)$, allora E^* conferma H, anche se in misura minore dell'evidenza diretta E.

In molti casi, data un'evidenza diretta E che conferma H, un'evidenza indiretta a favore di H è costituita da una testimonianza attendibile $T(E)$, cioè da una

testimonianza che conferma E. In tal caso, è naturale parlare di *testimonianza indiretta a favore* di H. Se vale la condizione – del tutto plausibile nella maggior parte delle testimonianze acquisite nella vita quotidiana[19] – che $\mathbf{E} \equiv (E,\neg E)$ separa $\mathbf{T(E)} \equiv (T(E),\neg T(E))$ da $\mathbf{H} \equiv (H,\neg H)$, allora $T(E)$ confermerà H. Questa possibilità viene bene illustrata dalla seguente versione, leggermente modificata, dell'Esempio 2.

Esempio 3. Testimonianza indiretta sulla presenza del noto piromane Fuego. Parlando con Forestale (in simboli: T), Xavier riceve la testimonianza $T(E)$ secondo la quale si è appena sviluppato un incendio nella boscaglia (E). L'enunciato E conferma l'ipotesi H che da quelle parti è passato il noto piromane Fuego. Se $T(E)$ è attendibile allora conferma E ed è, quindi, una testimonianza indiretta a favore di H. Un attimo di riflessione basterà a convincerci che **E** separa **T(E)** da **H**. Possiamo quindi concludere – grazie al teorema (23) – che $T(E)$ conferma H.

Testimonianze conflittuali. Vedremo ora che certi tipi di testimonianze conflittuali possono venire intese come particolari forme di evidenze ambigue. Questa possibilità viene illustrata nel seguente esempio.

Esempio 4. Testimonianze esperte conflittuali di due spie in Corea del Nord. Xavier ha due spie in Corea del Nord, che chiameremo T e T^*. Agli inizi del gennaio 2016 Xavier chiede a entrambe un parere circa l'eventualità E che quello stato riuscirà a completare il suo impianto missilistico segreto entro l'anno. Dopo un mese arrivano contemporaneamente sul tavolo di Xavier gli opposti pareri di T e T^*: il primo crede che E si realizzerà, il secondo che non si realizzerà. Possiamo considerare i pareri $T(E)$ e $T^*(\neg E)$, forniti da T e T^*, come *testimonianze esperte conflittuali*. In base alla sua valutazione della competenza di T e T^*, quale emerge dalla loro carriera di spie, Xavier ritiene che T sia due volte più competente di T^*, cosicché attribuisce alla testimonianza $T(E)$ un peso doppio di quello attribuito a $T^*(\neg E)$. Di conseguenza, aggiorna le sue vecchie probabilità $p(E)$ e $p(\neg E)$ in modo tale che la nuova probabilità $p_n(E)$ sia il doppio di $p_n(\neg E)$; pone quindi $p_n(E) = 2/3$ e $p_n(\neg E) = 1/3$. Xavier potrà poi usare l'evidenza ambigua $p_n(E) = 2/3$ per determinare, in accordo con il principio cinematico (CoG), la nuova probabilità di qualunque ipotesi di suo interesse come, per esempio, l'ipotesi H che la Corea del Nord sia in grado di lanciare missili intercontinentali entro il 2018.

[19] Cfr. Shogenji (2003, p. 615).

Per quanto riguarda la probabilità $p_n(E)$ che Xavier attribuisce ad E in risposta alle informazioni acquisite, l'evidenza ambigua descritta in questo esempio non si differenzia da quella delle rose intraviste alla luce di una candela, descritta nell'Esempio 1: in entrambi i casi, infatti, $p_n(E) = 2/3$. La differenza tra i due esempi consiste, invece, nella fonte di $p_n(E) = 2/3$: mentre nell'Esempio 1 tale probabilità viene ottenuta in risposta a un'osservazione effettuata in condizioni sfavorevoli, nell'Esempio 4 essa viene ottenuta in risposta alle testimonianze conflittuali di due esperti, tra i quali quello che testimonia E è ritenuto due volte più competente di quello che testimonia $\neg E$.

3. Testimonianze esperte e probabilità delle ipotesi nella pratica clinica

Gli studiosi che, in questi ultimi decenni, hanno sviluppato l'approccio bayesiano all'analisi della pratica clinica hanno dedicato molta attenzione all'epistemologia del processo diagnostico.[20] Nell'ambito di tale processo il medico si avvale di testimonianze esperte di diverso genere. In primo luogo opereremo una distinzione tra due tipi di testimonianze esperte, vale a dire le testimonianze strumentali e quelle interpretative (paragrafo 3.1). Successivamente, illustreremo il ruolo delle testimonianze interpretative nella determinazione della probabilità delle ipotesi diagnostiche (paragrafo 3.2).

3.1. Testimonianze esperte nel processo diagnostico

Una parte dell'evidenza usata nel processo diagnostico viene acquisita direttamente dal medico: si pensi, per esempio, alle informazioni ottenute attraverso l'esame obiettivo del paziente. In molti casi, tuttavia, la maggior parte dell'evidenza acquisita dal medico è costituita da *testimonianze esperte*, vale a dire dai responsi forniti da analisti e specialisti di vario genere – quali biologi, tossicologi, radiologi, genetisti, biologi molecolari e così via –, in seguito all'effettuazione di test diagnostici. Per esempio, il responso dell'analista che comunica al medico che il suo paziente rivela una copremia positiva può venire inteso come una testimonianza esperta. Le testimonianze esperte fornite in seguito all'effettuazione di test diagnostici si dispongono in una sorta di *conti-*

[20] Si vedano, per esempio, Scandellari (2005) e Weinstein e Fineberg (1980). Ci permettiamo di segnalare anche i contributi di Festa (2004; 2005) e Festa, Buttasi e Crupi (2009) che contengono ampi riferimenti alla recente letteratura sull'argomento.

nuum ai cui estremi si trovano quelle che potremo chiamare *testimonianze strumentali* e *testimonianze interpretative*.

Con "testimonianza strumentale" ci riferiamo al responso che il medico curante riceve da un analista che ha effettuato un test diagnostico il cui risultato è completamente (o quasi) determinato dal funzionamento di determinati dispositivi strumentali. In questi casi il compito dell'analista ha carattere essenzialmente pratico: consiste, cioè, nella corretta effettuazione del test, nell'attenta lettura del risultato rivelato dai dispositivi sperimentali e, infine, nella fedele trascrizione e comunicazione del risultato al medico. Le testimonianze strumentali includono, per esempio, i responsi di test diagnostici volti a stabilire determinati valori numerici – come il dosaggio dei trigliceridi o il numero di globuli rossi –, oppure a determinare la presenza o assenza di determinate caratteristiche, come nel test della copremia, che mira all'identificazione di sangue occulto nelle feci. Con riferimento a test di questo genere ci sembra appropriato parlare di testimonianze strumentali, poiché la testimonianza dell'analista è interamente determinata dalla lettura del *dispositivo strumentale* usato per effettuare il test, lettura che non presenta particolari difficoltà interpretative. Tali difficoltà, al contrario, sono l'aspetto distintivo di quelle che abbiamo chiamato "testimonianze interpretative". Con questo termine ci riferiamo ai casi in cui il responso dell'esperto che ha effettuato il test diagnostico si basa sulla sua *interpretazione* del risultato ottenuto dai dispositivi strumentali. In genere, le testimonianze interpretative vengono fornite in seguito all'effettuazione di test consistenti nell'acquisizione e interpretazione di immagini. Si pensi, per esempio, ai test diagnostici basati sull'interpretazione di immagini al microscopio di frammenti di tessuti, o di immagini fotografiche (quali radiografie, ecografie, risonanze magnetiche e TAC) o, ancora, di tracciati (quali elettrocardiogrammi ed elettroencefalogrammi). In questo genere di test il risultato non è costituito dalle immagini visive in quanto tali, bensì dalle immagini interpretate, cioè dall'opinione dell'analista circa il corretto significato delle immagini. Occorre notare che la maggiore complessità delle testimonianze interpretative, rispetto a quelle strumentali, non dipende tanto dalla sofisticazione della strumentazione usata per effettuare il test, quanto dall'alto grado di competenza e abilità richieste per una corretta interpretazione delle immagini ottenute dal test.[21]

[21] A quanto pare, non tutte le abilità e conoscenze degli esperti che forniscono testimonianze interpretative possono venire espresse in termini espliciti. Esse possono forse venire intese – per usare una celebre espressione di Michael Polanyi (1966) – come una forma di conoscenza tacita.

3.2. Testimonianze esperte e probabilità delle ipotesi diagnostiche

Dopo avere visitato un paziente, il medico si trova spesso a chiedersi se una determinata ipotesi diagnostica H sia corretta, oppure no. Per rispondere a tale interrogativo, può sottoporre il paziente a un test diagnostico. Supponiamo che il responso del test sia E. Allora il medico dovrà aggiornare la vecchia probabilità $p(H)$, attribuita ad H prima dell'effettuazione del test, sostituendola con la nuova probabilità $p(H|E)$. Sembra ragionevole richiedere che entrambe le probabilità – $p(H)$ e $p(H|E)$ – non debbano riflettere semplicemente le opinioni soggettive del medico, ma debbano avere un carattere oggettivo, cioè essere in accordo con le più aggiornate conoscenze acquisite dalle scienze mediche. L'esempio che segue mostra in che modo il medico possa operare una *valutazione oggettiva* sia di $p(H)$ sia di $p(H|E)$.

ESEMPIO 5. LA PROBABILITÀ CHE LA PAZIENTE ABBIA UN CARCINOMA MAMMARIO ALLA LUCE DEL RESPONSO POSITIVO DEL RADIOLOGO. Xavier si trova di fronte una donna di quarant'anni senza sintomi nella quale un esame fisico di controllo ha evidenziato un nodulo al seno. Allo scopo di appurare se il nodulo è maligno, cioè se la paziente ha un cancro al seno, decide di sottoporla al test della mammografia. Un *risultato positivo* di questo test è costituito da un responso in cui il radiologo afferma che l'immagine ottenuta dal test indica che la paziente ha il cancro. La natura complessa e sofisticata dell'attività interpretativa del radiologo ci permette di considerare tale responso come un tipico caso di testimonianza interpretativa. Indicheremo con "C" l'ipotesi che la paziente ha un cancro al seno, con "Pos" un risultato positivo del test e con "$\neg Pos$" un risultato negativo. Se il risultato è positivo, allora Xavier dovrà passare dalla sua vecchia probabilità $p(C)$ alla nuova probabilità $p(C|Pos)$ o, equivalentemente, dalla sua vecchia quota $o(C) \equiv p(C)/(1 - p(C))$ alla nuova quota $o(C|Pos) \equiv p(C|Pos)/(1 - p(C|Pos))$. Dai teoremi (8) e (15) segue che $o(C|Pos)$ è data da:

$$(35) \quad o(C|Pos) = o(C) \times \frac{p(Pos|C)}{p(Pos|\neg C)} = o(C) \times c_B(C,Pos).^{22}$$

[22] Possiamo prendere alla lettera l'idea che un risultato positivo del test della mammografia è costituito da una testimonianza interpretativa del radiologo T. Possiamo, cioè, identificare, Pos con la testimonianza esperta $T(C)$ con la quale T afferma di credere che l'ipotesi C sia vera. Questo implica che $c_B(C,Pos) \equiv p(Pos|C)/p(Pos|\neg C) = p(T(C)|C)/p(T(C)|\neg C) \equiv A(T(C))$, cioè che il grado di conferma $c_B(C,Pos)$ che Pos apporta a C è uguale all'attendibilità di $T(C)$. L'ugua-

Ciò significa che $o(C|Pos)$ può venire determinata a partire dalle probabilità $p(C)$, $p(Pos|C)$ e $p(Pos|\neg C)$. Di tutte e tre queste probabilità Xavier può fornire una valutazione oggettiva, in accordo con le conoscenze acquisite dalle scienze mediche e, più precisamente, con le acquisizioni dell'epidemiologia e dell'epidemiologia clinica. Consideriamo, anzitutto, la probabilità iniziale $p(C)$. La ricerca epidemiologica è giunta alla conclusione che, tra le quarantenni senza sintomi che rivelano un nodulo al seno, una su cento ha il cancro, cioè che la frequenza del cancro è pari all'1%.[23] In assenza di ulteriori informazioni sulla storia clinica della paziente, Xavier dovrà allora attribuire a $p(C)$ un valore basato sulle sole conoscenze epidemiologiche; dovrà cioè attribuire a $p(C)$ un valore pari a 0,01. Equivalentemente, dovrà porre $o(C) \equiv p(C)/(1 - p(C)) = 0{,}01/0{,}99 = 0{,}\overline{01}$. Il carattere oggettivo dei valori così attribuiti a $p(C)$ e $o(C)$ è garantito dal loro accordo con le conoscenze epidemiologiche. Consideriamo ora le probabilità $p(Pos|C)$ e $p(Pos|\neg C)$ e il grado di conferma $c_B(C,Pos) \equiv p(Pos|C)/p(Pos|\neg C)$ che *Pos* apporta a *C*. Le ricerche di epidemiologia clinica hanno consentito di determinare che, su 1000 donne che presentano un cancro al seno, 792 ottengono un risultato positivo al test della mammografia e che, su 1000 donne che non lo presentano, 96 ottengono comunque un risultato (falsamente) positivo. Sulla base di tali conoscenze Xavier dovrà porre $p(Pos|C) = 0{,}792$ e $p(Pos|\neg C) = 0{,}096$; di conseguenza dovrà attribuire a $c_B(C,Pos) \equiv p(Pos|C)/p(Pos|\neg C)$ un valore pari a 8,25.[24] Anche in questo caso, il carattere oggettivo dei valori così attribuiti a $p(Pos|C)$, $p(Pos|\neg C)$ e $c_B(C,Pos)$ è garantito dal loro accordo con conoscenze mediche di sfondo e, più precisamente, con i risultati dell'epidemiologia clinica.

Una volta determinati i valori di $o(C)$ e $c_B(C,Pos)$, l'uguaglianza (35) permetterà di calcolare $o(C|Pos)$, che risulterà approssimativamente pari a 0,083. Tale quota equivale a una probabilità $p(C|Pos)$ approssimativamente pari a 0,08. Ciò significa che, data una mammografia positiva, la probabilità che la paziente abbia un cancro al seno è pari all'incirca all'8%. Poiché, come si è visto, tutte le probabilità usate per determinare la nuova probabilità $p(C|Pos)$ e la corrispondente quota $o(C|Pos) \equiv p(C|Pos)/(1 - p(C|Pos))$ sono state determinate sulla base di valutazioni oggettive – cioè di valutazioni scientificamente ben fondate –, possiamo concludere che $o(C|Pos)$ e $p(C|Pos)$ hanno ca-

glianza $c_B(C,Pos) = A(T(C))$ ci consente di riformulare il teorema (35) nel seguente modo: $o(C|Pos) = o(C) \times A(T(C))$.

[23] Tutti i dati epidemiologici utilizzati in questo esempio provengono da Eddy (1982).

[24] Ciò significa che un risultato positivo della mammografia accresce di più di otto volte la quota iniziale $o(C)$.

rattere oggettivo. In maniera perfettamente analoga possiamo calcolare la probabilità finale $p(C|\neg Pos)$ dell'ipotesi che la paziente abbia un cancro dato un risultato negativo del test: tale probabilità sarà approssimativamente pari a 0,002, cioè al 2‰. Naturalmente anche $p(C|\neg Pos)$ avrà carattere oggettivo.

L'Esempio 5 mostra che si può determinare, in accordo con i risultati dell'epidemiologia clinica, il "valore oggettivo" di $c_B(C,Pos)$. Tale valore ci informa, per così dire, sull'attendibilità media della totalità dei responsi positivi formulati dai radiologi in determinate circostanze – cioè in determinati paesi o sistemi sanitari, con determinate tecnologie, e così via. In mancanza di informazioni specifiche circa lo specifico grado di competenza del radiologo da cui ha ricevuto il responso positivo, Xavier potrà calcolare la probabilità che la sua paziente abbia il cancro in base al presupposto che l'attendibilità del responso positivo da lui ottenuto è identica al valore oggettivo di $c_B(C,Pos)$.[25] Anche se di norma non sono disponibili precisi dati quantitativi sul grado di competenza dei radiologi che operano in una determinata zona, spesso i medici dispongono di informazioni largamente condivise – espresse di solito in forma comparativa – sul loro grado di competenza.[26] Nel prossimo esempio – che ha la stessa struttura cognitiva dell'Esempio 4, relativo alle testimonianze conflittuali di due spie – mostreremo come le informazioni sulla competenza di due radiologi che hanno fornito responsi conflittuali possono venire usate per determinare la probabilità che una paziente abbia il cancro.

ESEMPIO 6. LA PROBABILITÀ CHE LA PAZIENTE ABBIA UN CARCINOMA MAMMARIO ALLA LUCE DEI RESPONSI CONFLITTUALI DI DUE RADIOLOGI. La paziente alla quale Xavier ha prescritto una mammografia si rivolge, per eccesso di scrupolo, a due radiologi, che chiameremo T e T^*, i quali forniscono due responsi conflittuali, il primo positivo e il secondo negativo. Xavier potrebbe interpretare tali responsi come testimonianze esperte conflittuali – che indicheremo con "$T(Pos)$" e "$T^*(\neg Pos)$" – circa il "corretto" risultato del test, cioè circa il responso che verrebbe fornito dalla maggior parte dei radiologi che operano nelle circostanze date. Xavier ritiene che la competenza di T sia due volte maggiore di quella di T^* cosicché attribuisce a $T(Pos)$ un peso doppio di quello attribuito a $T^*(\neg Pos)$. Di conseguenza aggiorna le sue vecchie probabilità $p(Pos)$ e $p(\neg Pos)$ passando alle nuove probabilità $p_n(Pos) = 2/3$ e $p_n(\neg Pos) = 1/3$. Successivamente, Xavier usa l'evidenza ambigua $p_n(Pos) = 2/3$ per determinare,

[25] Si veda la nota 22.

[26] Sulla misurazione della qualità del giudizio esperto, si veda Cooke (1991).

in accordo con il principio cinematico (CoG), la nuova probabilità $p_n(C) = p(C|Pos) \times p_n(Pos) + p(C|\neg Pos) \times p_n(\neg Pos)$ che la paziente abbia il cancro. Tenendo conto del fatto che $p(C|Pos)$ e $p(C|\neg Pos)$ sono approssimativamente uguali a 0,08 e 0,002 (vedi Esempio 5), Xavier conclude che $p_n(C)$ è approssimativamente uguale a 0,04, cioè al 4%.[27]

4. Testimonianze esperte e probabilità dell'ipotesi di colpevolezza nella pratica giudiziaria

Nell'ambito dell'analisi bayesiana della pratica giudiziaria notevole attenzione è stata dedicata alla determinazione della probabilità dell'ipotesi di colpevolezza nel processo penale.[28] In questo paragrafo considereremo la natura della probabilità iniziale e finale dell'ipotesi di colpevolezza e illustreremo il modo in cui la probabilità finale può venire determinata in base a certi tipi di testimonianze indipendenti o indirette (paragrafo 4.1). Successivamente, ci occuperemo della determinazione della probabilità dell'ipotesi di colpevolezza alla luce di testimonianze esperte (paragrafo 4.2).

4.1. Probabilità dell'ipotesi di colpevolezza nel processo penale

La struttura cognitiva del processo penale appare, nei suoi tratti generali, identica a quella del processo diagnostico. Infatti, come l'evidenza acquisita nel processo diagnostico viene usata per determinare la probabilità delle ipotesi diagnostiche, allo stesso modo l'evidenza acquisita nel processo penale viene usata per determinare la probabilità dell'ipotesi di colpevolezza e, più in generale, di tutte le cosiddette ipotesi ricostruttive prospettate nel processo. Possiamo notare, tuttavia, anche alcune notevoli differenze cognitive tra processo diagnostico e penale. Tali differenze, che verranno ora brevemente illustrate, sono connesse alla peculiare natura della probabilità iniziale e finale dell'ipotesi di colpevolezza.

Come si è visto nel paragrafo 3.2, la probabilità iniziale delle ipotesi diagnostiche dovrebbe essere determinata sulla base di valutazioni oggettive, effettua-

[27] Sull'uso di evidenze e testimonianze incerte nella pratica clinica, si veda Festa, Buttasi e Crupi (2009).

[28] Si vedano, per esempio, Anderson, Schum e Twining (1991/2005), Dawid (2005), Frosini (2002), Lucy (2005) e Tillers e Green (1988).

te in accordo con i risultati delle scienze mediche. D'altra parte, la natura della probabilità iniziale attribuita all'ipotesi di colpevolezza nel processo penale appare molto diversa. Le conoscenze di un giudice bene informato sui risultati delle scienze forensi, a partire da quelli della sociologia criminale, potrebbero – almeno in linea di principio – consentirgli di operare una valutazione oggettiva della probabilità iniziale $p(C)$ dell'ipotesi di colpevolezza C, prospettata dall'accusa, secondo la quale l'imputato è colpevole: tale valutazione potrebbe venire effettuata, per esempio, sulla base di sesso, età, gruppo etnico e precedenti penali dell'imputato. Tuttavia, questo genere di valutazione non può venire adottato nella pratica giudiziaria dei paesi evoluti, dove entrano in gioco considerazioni extraepistemiche che impongono forti vincoli sulla determinazione di $p(C)$. Ci riferiamo qui al principio della *presunzione di innocenza* – nel seguito: (PI) – il quale afferma che occorre partire dalla supposizione che l'imputato sia innocente. (PI) viene comunemente tradotto in termini bayesiani con la richiesta che il valore di $p(C)$ sia molto basso; a fini illustrativi, possiamo identificare tale richiesta con la condizione che la probabilità iniziale di C sia pari all'uno per mille o, equivalentemente, che la quota iniziale di C sia pari a 1 contro 999:

(PI) $p(C) = 0{,}001$ o, equivalentemente, $o(C) = 1/999 = 0{,}\overline{001}$.[29]

Un principio di decisione giudiziaria ampiamente adottato nei paesi evoluti è quello dell'"*oltre ogni ragionevole dubbio*" – nel seguito: (RD) – il quale afferma che un verdetto di colpevolezza può venire emesso solo se l'evidenza a favore di C è schiacciante.[30] (RD) viene comunemente tradotto in termini bayesiani con la richiesta che un verdetto di colpevolezza possa essere emesso solo a condizione che la probabilità finale di C sia superiore al 99,9% o, equivalentemente, che la quota finale di C sia superiore a 999 contro 1:

(RD) Un verdetto di colpevolezza può venire emesso solo se $p(C|E) \geq 0{,}999$ o, equivalentemente, solo se $o(C,E) \geq 999$.

[29] Si noti che il valore 0,001 qui attribuito a $p(C)$ ha significato puramente indicativo e dovrebbe, in molti casi, essere sostituito da un valore ancora più basso. Secondo alcuni autori $p(C)$ dovrebbe essere identificato con la probabilità che il crimine sia stato commesso da un membro qualunque della "popolazione rilevante" cui appartiene l'imputato, cioè del gruppo di persone che, in linea di principio, potrebbero avere commesso il crimine. Se, per esempio, vi sono dieci milioni di persone che, in linea di principio, avrebbero potuto commettere il crimine, allora il valore attribuito a $p(C)$ dovrebbe essere pari a 0,0000001, cioè a uno su dieci milioni.

[30] Sulla natura e la funzione di (RD), si vedano Canzio (2004) e Stella (2003, pp. 161 ss. e 178 ss.).

Naturalmente, l'elevato valore di $p(C|E)$ richiesto in (RD) potrebbe dipendere, in ampia misura, da un'elevata probabilità iniziale $p(C)$. Tuttavia, l'adozione di (PI) scongiura questa possibilità. Possiamo infatti dimostrare che, se si attribuisce – in accordo con (PI) – un valore molto piccolo a $p(C)$, allora un elevato valore di $p(C|E)$ può essere ottenuto solo se l'evidenza E a favore di C è schiacciante, cioè solo se E conferma molto fortemente C. Più precisamente, segue dall'uguaglianza $o(C|E) = o(C) \times c_B(C,E)$ (vedi (15)) che:

(36) Se vale (PI) allora (RD) equivale alla seguente condizione:

> (RD*) Un verdetto di colpevolezza può venire emesso solo se $c_B(C,E) \geq 998\,001$ o, equivalentemente, solo se $p(E|C) \geq 998\,001 \times p(E|\neg C)$.

Il contenuto intuitivo di (RD*) può venire espresso dicendo che un verdetto di colpevolezza può venire emesso solo se il grado di conferma $c_B(C,E)$ ha un valore vicino al milione, cioè solo se la probabilità $p(E|C)$ dell'evidenza alla luce dell'ipotesi di colpevolezza è quasi un milione di volte più grande della probabilità $p(E|\neg C)$ dell'evidenza alla luce della supposizione che l'ipotesi di colpevolezza sia falsa.

Vedremo ora, con l'aiuto di alcuni esempi, che l'evidenza a favore dell'ipotesi di colpevolezza può includere sia testimonianze indipendenti sia testimonianze indirette.

Secondo un celebre motto, attribuito al cardinale John Henry Newman (1870), tre indizi fanno una prova.[31] Certamente questo motto vale per le testimonianze a favore dell'ipotesi di colpevolezza come si vede dal seguente esempio, il quale mostra che tre testimonianze indipendenti e molto attendibili a favore dell'ipotesi di colpevolezza rappresentano un'evidenza schiacciante.[32]

ESEMPIO 7. TRE TESTIMONIANZE INDIPENDENTI E MOLTO ATTENDIBILI A FAVORE DELL'IPOTESI DI COLPEVOLEZZA. Secondo l'ipotesi C, il noto piromane Fuego, accusato di aver appiccato il recente incendio nella boscaglia, è colpevole. I testimoni T_1, T_2 e T_3 affermano di aver visto Fuego mentre appiccava l'incendio. Se sappiamo che i tre testimoni non si conoscono fra loro e hanno potuto osservare la scena del delitto da tre diverse posizioni, allora possiamo ragio-

[31] Newman ha reso famoso questo motto che compare, in diverse versioni, anche in altri autori. Il motto è poi entrato nella letteratura probabilistica grazie a de Finetti (1970, cap. 4.15.4).

[32] Sulla congiunzione di testimonianze indipendenti si veda Dawid (1987).

nevolmente concludere che le loro testimonianze – vale a dire, $T_1(C)$, $T_2(C)$ e $T_3(C)$ – sono indipendenti. Di conseguenza, grazie ai teoremi (20) e (29), valgono le uguaglianze $c_B(C,T_1(C)\ \&\ T_2(C)\ \&\ T_3(C)) = c_B(C,T_1(C)) \times c_B(C,T_2(C)) \times c_B(C,T_3(C)) = A(T_1(C)) \times A(T_2(C)) \times A(T_3(C))$. Supponiamo, inoltre, che $T_1(C)$, $T_2(C)$ e $T_3(C)$ siano *molto attendibili*, nel senso che $A(T_1(C))$, $A(T_2(C))$, $A(T_3(C)) > 100$. In tal caso, dalle uguaglianze appena viste segue che $c_B(C,T_1(C)\ \&\ T_2(C)\ \&\ T_3(C)) > 1\,000\,000$ e quindi, a maggior ragione, che $c_B(C,T_1(C)\ \&\ T_2(C)\ \&\ T_3(C)) > 998\,001$. Questo implica – per il teorema (36) – che, data la presunzione di innocenza (PI), secondo la quale $p(C) = 0{,}001$, il grado di conferma $c_B(C,T_1(C)\ \&\ T_2(C)\ \&\ T_3(C))$ supera la soglia necessaria per l'emissione di un verdetto di colpevolezza.

Il seguente esempio mostra che l'evidenza a favore dell'ipotesi di colpevolezza può essere costituita da una testimonianza indiretta.

ESEMPIO 8. TESTIMONIANZA INDIRETTA A FAVORE DELL'IPOTESI DI COLPEVOLEZZA. Forestale (in simboli: T) fornisce la testimonianza secondo la quale, circa due ore prima della segnalazione dell'incendio nella boscaglia, il noto piromane Fuego – accusato di aver appiccato l'incendio – era da quelle parti, alla guida della sua auto (E). L'enunciato E conferma l'ipotesi C che Fuego sia colpevole. Se la testimonianza $T(E)$ di Forestale è attendibile, allora $T(E)$ conferma E ed è, quindi, una testimonianza indiretta a favore di C. Un attimo di riflessione basterà a convincerci che $\mathbf{E} \equiv (E,\neg E)$ separa $\mathbf{T(E)} \equiv (T(E),\neg T(E))$ da $\mathbf{C} \equiv (C,\neg C)$. Possiamo quindi concludere – grazie al teorema (23) – che $T(E)$ conferma C.

4.2. Testimonianze esperte e probabilità dell'ipotesi di colpevolezza

L'evidenza utilizzata nel processo penale include svariati tipi di prove scientifiche, che di solito vengono comunicate al giudice attraverso le testimonianze esperte fornite dai periti: si pensi, per esempio, all'esame grafologico, all'analisi di fibre, capelli e impronte digitali, o al test del DNA.[33] Nel seguito di questo paragrafo considereremo brevemente, con riferimento al test del DNA, alcuni problemi relativi all'uso delle testimonianze esperte nella determinazione della probabilità dell'ipotesi di colpevolezza.

[33] Sull'uso delle prove scientifiche e delle testimonianze esperte nel processo penale si vedano Brewer (1998), De Cataldo Neuburger (2008), Golanski (2001) e Walton (1997, cap. 6).

Come è noto l'acido desossiribonucleico, o DNA, è il fondamento chimico dell'ereditarietà. L'applicazione delle conoscenze sul DNA nelle scienze forensi risale alla metà degli anni Ottanta, quando fu sviluppato un metodo, piuttosto elaborato, per isolare e visualizzare appropriati frammenti di DNA estratti dal sangue o da altro materiale organico. Tale metodo consente di ottenere una particolare immagine, nota come *profilo del DNA*. Il test del DNA consiste nel confronto, effettuato da un analista genetico, tra i profili del DNA tratti da due campioni di materiale organico: si dice che il test ha dato un risultato positivo quando l'analista genetico dichiara che i profili *concordano*, cioè che sono sostanzialmente identici. La natura complessa e sofisticata dell'attività interpretativa dell'analista ci permette di considerare il test del DNA come un esempio di testimonianza interpretativa.[34]

Supponiamo che il test del DNA venga applicato a due campioni di materiale organico – uno prelevato da una traccia trovata sul luogo del delitto e l'altro dall'imputato –, e che l'analista dichiari che i due profili concordano: in tal caso parleremo di *concordanza dichiarata*. La concordanza dichiarata non implica che l'imputato è colpevole. Infatti, la catena inferenziale dalla concordanza dichiarata alla colpevolezza è costituita dai seguenti tre passi, ciascuno dei quali caratterizzato da una buona dose d'incertezza:

Passo 1. A partire dalla concordanza dichiarata tra i due profili (*CD*), si determina la probabilità della *concordanza effettiva* (*CE*), cioè la probabilità che i profili siano davvero sostanzialmente identici.
Passo 2. A partire dalla supposizione che vi sia una concordanza effettiva, si determina la probabilità che l'imputato sia la *fonte* della traccia (*F*).
Passo 3. Infine, a partire dalla supposizione che l'imputato sia la fonte, si determina la probabilità della sua *colpevolezza* (*C*).[35]

Anche se le possibilità di errore insite in ciascuno di questi passi sono note, non sembra che il significato di tale circostanza sia stato ampiamente compreso. Di conseguenza, ancora oggi si tende ad attribuire alla concordanza dichiarata lo status di prova inattaccabile, che può in taluni casi condurre alla virtuale certezza che l'imputato è colpevole. Allo scopo di rimuovere alcuni

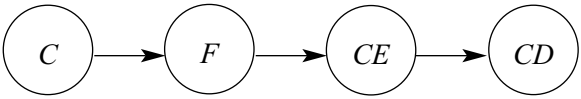
[34] Sul test del DNA e le sue applicazioni forensi, si vedano Goodwin, Linacre e Hadi (2007) e Lucy (2005, capp. 14 e 15). Per una concisa esposizione di carattere divulgativo, è consigliabile la lettura di Gigerenzer (2002, cap 10).

[35] Le seguenti pagine possono essere lette come una ricostruzione in termini bayesiani della catena inferenziale descritta da Gigerenzer (2002, pp. 192-204).

comuni fraintendimenti, occorre descrivere con qualche dettaglio la catena inferenziale che conduce dalla concordanza dichiarata alla determinazione della probabilità dell'ipotesi di colpevolezza.

Possiamo considerare la concordanza dichiarata come un'evidenza indiretta che conferma l'ipotesi di colpevolezza attraverso due anelli intermedi, costituiti dalla concordanza effettiva e dalla fonte, cioè dalla supposizione che l'imputato sia la fonte della traccia. Le relazioni tra concordanza dichiarata e colpevolezza possono venire rappresentate dalla seguente Figura 3 – del tutto simile alla Figura 2 –, nella quale le variabili **C** ≡ (C,$\neg C$), **F** ≡ (F,$\neg F$), **CE** ≡ (CE,$\neg$CE) e **CD** ≡ (CD,$\neg CD$) indicano, rispettivamente, la colpevolezza, la fonte, la concordanza effettiva e la concordanza dichiarata:

FIG. 3: Rete bayesiana con connessioni seriali



La Figura 3 mostra che **CE** separa **CD** da **F** ed **F** separa **CE** da **C**. Come già sappiamo, questo implica che varrà anche la seguente relazione di separazione: **F** separa **CD** da **C**. Possiamo quindi applicare i teoremi (24) e (25) dai quali, fatte le debite sostituzioni, segue che:

(37) $p(C|CD) = p(C|F) \times p(F|CD) + p(C|\neg F) \times p(\neg F|CD)$.

(38) $p(F|CD) = p(F|CE) \times p(CE|CD) + p(F|\neg CE) \times p(\neg CE|CD)$.

Le uguaglianze (37) e (38) mostrano come possiamo determinare $p(C|CD)$, cioè la probabilità della colpevolezza alla luce della concordanza dichiarata. A tale scopo occorrerà anzitutto determinare, mediante la (38), i valori di $p(F|CD)$ e $p(\neg F|CD) \equiv 1 - p(F|CD)$; come si vede, tali valori possono venire calcolati sulla base delle probabilità $p(F|CE)$, $p(F|\neg CE)$ e $p(CE|CD)$. Successivamente si determinerà $p(C|CD)$ mediante la (37), sulla base di $p(C|F)$ e $p(C|\neg F)$ e dei valori di $p(F|CD)$ e $p(\neg F|CD)$, precedentemente calcolati mediante la (38). La procedura appena descritta mostra che $p(C|CD)$ può venire determinata sulla base delle seguenti cinque probabilità: $p(C|F)$, $p(C|\neg F)$, $p(F|CE)$, $p(F|\neg CE)$ e $p(CE|CD)$. A loro volta, tali probabilità possono venire determinate effettuando i tre passi sopra menzionati della catena inferenziale che va dalla concordanza dichiarata alla colpevolezza. Tali passi verranno ora illustrati con l'aiuto di alcuni esempi.

Passo 1. Come determinare la probabilità $p(CE|CD)$ della concordanza effettiva alla luce della concordanza dichiarata. La concordanza dichiarata dall'analista genetico tra i profili dei due DNA – provenienti l'uno dalla traccia trovata sul luogo del delitto e l'altro dall'imputato –, non conduce alla certezza che vi sia concordanza effettiva. Infatti, la concordanza dichiarata potrebbe essere il frutto di un errore di laboratorio oppure di un'erronea interpretazione, da parte dell'analista, della somiglianza tra i profili. Occorre quindi determinare la probabilità $p(CE|CD)$ – o, equivalentemente, la quota $o(CE|D)$ –, da attribuire alla concordanza effettiva alla luce della concordanza dichiarata. Segue dal teorema (15) che $o(CE|CD) = o(CE) \times c_B(CE,CD)$ – ove $c_B(CE,CD) \equiv p(CD|CE)/p(CD|\neg CE)$ è il grado di conferma apportato dall'evidenza CD all'ipotesi CE. L'uguaglianza $o(CE|CD) = o(CE) \times c_B(CE,CD)$ mette in luce una circostanza molto importante: poiché $o(CE|CD)$ è pari al prodotto di $o(CE)$ e $c_B(CE,CD)$, se la quota iniziale $o(CE)$ attribuita alla concordanza effettiva è sufficientemente piccola, la quota finale $o(CE|CD)$ può essere piuttosto bassa anche in presenza di un valore molto elevato di $c_B(CE,CD)$, cioè di un risultato di concordanza dichiarata ottenuto sulla base di un test del DNA molto attendibile. Questa possibilità viene illustrata dal seguente esempio.

ESEMPIO 9 – PRIMA PARTE. DOVE SI MOSTRA CHE LA PROBABILITÀ DELLA CONCORDANZA EFFETTIVA ALLA LUCE DELLA CONCORDANZA DICHIARATA IN UN TEST DEL DNA ESTREMAMENTE ATTENDIBILE PUÒ ESSERE MOLTO BASSA. Supponiamo che, nei test del DNA eseguiti da un certo laboratorio, la concordanza dichiarata venga ottenuta 99 999 volte ogni centomila casi di concordanza effettiva, e una sola volta ogni centomila casi senza concordanza effettiva. Questo significa che $p(CD|CE) = 0{,}99999$ e $p(CD|\neg CE) = 0{,}00001$, cioè che il test è estremamente attendibile, dato che $c_B(CE,CD) \equiv p(CD|CE)/p(CD|\neg CE) = 99\,999$. Immaginiamo ora che solo un individuo su un milione abbia una concordanza effettiva con il DNA prelevato dalla traccia, cioè che $p(CE) = 0{,}000001$. Ciò significa che $o(CE) \equiv p(CE)/(1 - p(CE)) = 1/999\,999$. Ne segue che $o(CE|CD) = o(CE) \times c_B(CE,CD)$ è approssimativamente uguale a 0,1 o, equivalentemente, che $p(CE|CD)$ è approssimativamente uguale a 0,091, cioè al 9,1%. Come si vede la probabilità $p(CE|CD)$ da attribuire alla concordanza effettiva alla luce di un risultato di concordanza dichiarata, ottenuto con un test del DNA estremamente attendibile, è molto più bassa di quanto ci si potrebbe aspettare e, in ogni caso, è ben lontana dalla certezza.

Passo 2. Come determinare le probabilità $p(F|CE)$ e $p(F|\neg CE)$ dell'ipotesi della fonte alla luce della concordanza effettiva o della sua assenza. La presenza

di una concordanza effettiva tra i profili dei due DNA non conduce alla certezza che l'imputato sia la fonte della traccia. Può darsi, infatti, che due persone prese a caso abbiano profili uguali e, quindi, che il DNA dell'imputato sia sostanzialmente identico a quello prelevato dalla traccia, anche se egli non ne è la fonte.[36] Occorre quindi determinare le probabilità $p(F|CE)$ e $p(F|\neg CE)$ – o, equivalentemente, le quote $o(F|CE)$ e $o(F|\neg CE)$ –, dell'ipotesi della fonte alla luce della concordanza effettiva o della sua assenza. Occupiamoci, anzitutto, della probabilità $p(F|\neg CE)$ che l'imputato sia la fonte in assenza di una concordanza effettiva con la traccia. Appare ragionevole porre $p(F|\neg CE) = 0$ poiché, data l'assenza di concordanza effettiva tra il DNA dell'imputato e quello della fonte, l'imputato non può essere la fonte. Restano da determinare i valori di $p(F|CE)$ e $o(F|CE)$. Segue dal teorema di Bayes (9) che $o(F|CE) = o(F) \times p(CE|F)/p(CE|\neg F)$: potremmo quindi determinare $o(F|CE)$ – e quindi anche $p(F|CE)$ – sulla base di $o(F)$, $p(CE|F)$, e $p(CE|\neg F)$. In molti casi, tuttavia, non vi è alcun bisogno di ricorrere al teorema di Bayes, poiché la probabilità $p(F|CE)$ può essere determinata mediante una semplice procedura che verrà illustrata nel seguente esempio.

ESEMPIO 9 – SECONDA PARTE. DOVE SI MOSTRA CHE LA PROBABILITÀ DELL'IPOTESI DELLA FONTE ALLA LUCE CONCORDANZA EFFETTIVA PUÒ ESSERE MOLTO BASSA. Supponiamo che gli investigatori abbiano stabilito che la popolazione delle potenziali fonti – cioè delle persone che avrebbero potuto commettere il delitto e lasciare una traccia –, include dieci milioni di membri. Immaginiamo, inoltre, che gli analisti genetici abbiano stabilito – come si è ipotizzato nella prima parte di questo esempio –, che solo un individuo su un milione ha una concordanza con il DNA prelevato dalla traccia. Ciò implica che nella popolazione delle potenziali fonti ci si devono attendere dieci persone che presentano una concordanza effettiva con la traccia, inclusa la fonte effettiva della traccia. Di conseguenza ogni membro della popolazione che presenta una concordanza con la traccia potrebbe esserne la fonte, ma potrebbe anche essere uno dei nove individui che concordano casualmente con la traccia. Ciò significa che, data la concordanza effettiva dell'imputato con la fonte, la probabilità $p(F|CE)$ che l'imputato sia la fonte è pari al 10%, cioè a 0,1. Questo esempio mette in luce una circostanza di grande interesse: se la popolazione delle potenziali fonti è molto vasta, allora il valore di $p(F|CE)$ può essere sorprendentemente basso.

[36] Qui trascuriamo, per semplicità, la possibilità che l'identità del DNA dipenda dalla consanguineità, per esempio dal fatto che la traccia è stata lasciata da un gemello identico dell'imputato.

Passo 3. Come determinare le probabilità $p(C|F)$ e $p(C|\neg F)$ dell'ipotesi di colpevolezza alla luce della condizione della fonte o della sua assenza. La supposizione che F sia vera, cioè che l'imputato sia la fonte della traccia, non significa affatto che sia colpevole: può darsi, infatti, che non abbia commesso il delitto pur essendo la fonte della traccia. L'imputato potrebbe avere lasciato le proprie tracce sulla scena del delitto, prima o dopo il delitto, senza esserne l'autore; oppure il vero colpevole potrebbe avere trasportato intenzionalmente sulla scena del delitto qualche materiale biologico dell'imputato; infine, qualcun altro potrebbe avercelo trasportato, intenzionalmente oppure no.[37] Occorre quindi determinare le probabilità $p(C|F)$ e $p(C|\neg F)$ dell'ipotesi di colpevolezza alla luce della condizione della fonte o della sua assenza.

ESEMPIO 9 – TERZA PARTE. DOVE SI DETERMINA LA PROBABILITÀ DELLA COLPEVOLEZZA ALLA LUCE DELLA CONDIZIONE DELLA FONTE O DELLA SUA ASSENZA. La valutazione delle probabilità $p(C|F)$ e $p(C|\neg F)$ dipenderà, naturalmente, dalle specifiche caratteristiche del delitto. Tuttavia, in linea generale, si può affermare che, per le considerazioni fatte sopra, occorre attribuire a $p(C|F)$ un valore piuttosto inferiore da 1; a scopo illustrativo, porremo quindi $p(C|F) = 0{,}8$. Per quanto riguarda, invece, $p(C|\neg F)$, sembra del tutto naturale attribuire a $p(C|\neg F)$ un valore estremamente basso, approssimativamente uguale a zero; a scopo illustrativo porremo $p(C|\neg F) = 0$.

Disponiamo ora di tutti gli elementi per applicare le uguaglianze (37) e (38) e determinare così, con riferimento all'Esempio 9, la probabilità $p(C|CD)$ della colpevolezza alla luce della concordanza dichiarata.

ESEMPIO 9 – QUARTA PARTE. DOVE SI MOSTRA CHE LA PROBABILITÀ DELL'IPOTESI DI COLPEVOLEZZA ALLA LUCE DELLA CONCORDANZA DICHIARATA IN UN TEST DEL DNA ESTREMAMENTE ATTENDIBILE PUÒ ESSERE MOLTO BASSA. Nelle prime tre parti di questo esempio abbiamo determinato le seguenti probabilità: $p(CE|CD) \simeq 0{,}091$. $p(F|CE) = 0{,}1$, $p(F|\neg CE) = 0$, $p(C|F) = 0{,}8$ e $p(C|\neg F) = 0$. A partire dai valori di $p(CE|CD)$, $p(F|CE)$ e $p(F|\neg CE)$ possiamo ora calcolare, applicando l'uguaglianza (38), i valori di $p(F|CD)$ e $p(\neg F|CD)$ i quali sono dati da $p(F|CD) \simeq 0{,}1 \times 0{,}091 + 0 \times 0{,}909 = 0{,}0091$ e $p(\neg F|CD) \equiv 1 - p(F|CD) \simeq 0{,}9909$. Successivamente, a partire dai valori di $p(C|F)$ e $p(C|\neg F)$ e da quelli, appena cal-

[37] Per esempio, nel caso di O. J. Simpson la difesa riuscì a convincere la giuria che una traccia di sangue – di cui l'imputato era molto probabilmente la fonte – , era stata messa nel luogo del delitto dalla polizia: cfr. Gigerenzer (2002, p. 192).

colati, di $p(F|CD)$ e $p(\neg F|CD)$, possiamo calcolare, applicando l'uguaglianza (37), il valore di $p(C|CD)$ che è dato da $p(C|CD) \simeq 0{,}8 \times 0{,}0091 + 0 \times 0{,}9909 = 0{,}00728 = 0{,}728\%$. Ciò significa che la probabilità $p(C|CD)$ della colpevolezza alla luce della concordanza dichiarata è inferiore all'1%.

Anche se i dati utilizzati nell'Esempio 9 hanno carattere illustrativo, essi non sono troppo lontani da quelli che caratterizzano l'applicazione del test del DNA in molti processi penali. Ciò significa che – con buona pace dei più entusiasti sostenitori dell'uso delle prove scientifiche nei processi penali –, in molte applicazioni del test del DNA, la probabilità dell'ipotesi di colpevolezza, alla luce della sola evidenza costituita dalla concordanza dichiarata, è piuttosto bassa. Vale comunque la pena notare che nell'Esempio 9 (seconda parte) si era ipotizzato che i potenziali autori del delitto erano 10 milioni: ciò significa che il valore attribuito alla probabilità iniziale $p(C)$ dell'ipotesi di colpevolezza era pari a una su 10 milioni. Di conseguenza la probabilità finale $p(C|CD) \simeq 1\%$ ottenuta nell'Esempio 9 risulta essere all'incirca centomila volte più grande della probabilità iniziale $p(C)$. L'Esempio 9 ci permette così di trarre un'importante lezione di carattere generale sull'uso del test del DNA nel processo penale: l'evidenza costituita dalla concordanza dichiarata ottenuta nel test del DNA può accrescere in modo spettacolare la probabilità dell'ipotesi di colpevolezza senza però essere in grado – se presa da sola – di portare tale probabilità oltre la soglia necessaria all'emissione di un verdetto di colpevolezza.

In questo contributo abbiamo considerato alcuni interessanti problemi posti dall'applicazione dell'epistemologia bayesiana della testimonianza all'analisi della pratica clinica e giudiziaria ma, per motivi di spazio, abbiamo dovuto trascurarne altri non meno interessanti. Ci limitiamo a segnalarne due, che ci proponiamo di affrontare in altra occasione: il primo riguarda la ricerca dei modi più appropriati per valutare l'attendibilità delle testimonianze (comuni ed esperte) acquisite nel processo diagnostico e in quello penale;[38] il secondo, invece, concerne la ricerca dei modi più appropriati per determinare la probabilità delle ipotesi diagnostiche o ricostruttive alla luce dell'enorme ed eterogenea massa di evidenze, di carattere testimoniale e non, che vengono acquisite in molti processi diagnostici e penali.[39]

[38] Per quanto riguarda la valutazione dell'attendibilità delle testimonianze comuni acquisite nel processo penale si vedano Friedman (1987) e Wells e Olsen (2003).

[39] Sui modi più appropriati per determinare la probabilità delle ipotesi diagnostiche alla luce di numerose ed eterogenee informazioni, si veda Sackett *et al*. (1985); per quanto riguarda,

RIFERIMENTI BIBLIOGRAFICI

ADLER, J. (2008). Epistemological Problems of Testimony. In *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/archives/fall2008/entries/testimony-episprob/.

ANDERSON T., SCHUM D. e TWINING W. (1991/2005). *Analysis of Evidence*. Cambridge: Cambridge University Press.

BAYES, Th. (1763). "An Essay towards Solving a Problem in the Doctrine of Chances". In: *Philosophical Transactions of the Royal Society*, **53**, pp. 370-418. Trad. it. Saggio sulla soluzione di un problema della dottrina delle *chances*. In P. Garbolino (a cura di), *Sulla probabilità*, Ferrara: Librit, 1994, pp. 74-110.

BREWER, S. (1998). Scientific Expert Testimony and Intellectual Due Process. *Yale Law Journal*, **107**, pp. 1535-1681.

BUTLER, J. (1736). *The Analogy of Religion, Natural and Unrevealed, to the Constitution and Course of Nature*. London.

CANZIO G. (2004). L'"oltre il ragionevole dubbio" come regola probatoria e di giudizio nel processo penale. *Rivista italiana di diritto processuale e penale*, **47**, pp. 303-308.

COADY, C. A. J. (1992). *Testimony: A Philosophical Study*. Oxford: Oxford University Press.

COOKE, R. M. (1991). *Experts in Uncertainty. Opinion and Subjective Probability in Science*. Oxford: Oxford University Press.

CRUPI, V., FESTA, R. e MASTROPASQUA, T. (2008). Bayesian Confirmation by Uncertain Evidence: A Reply to Huber (2005). *The British Journal for the Philosophy of Science*, **59**, pp. 201-211.

DAWID, A. P. (1987). The Difficulty About Conjunction. *The Statistician*, **36**, pp. 91-97.

——— (2005). Probability and Proof. Appendice online alla seconda edizione di Anderson T., Schum, D. e Twining, W. (1991/2005), http://tinyurl.com/7g3bd.

DE CATALDO NEUBURGER, L. (2008) (a cura di). *La prova scientifica nel processo penale*. Padova: CEDAM.

DE FINETTI, B. (1970). *Teoria delle probabilità,* vol. I. Torino: Einaudi.

EARMAN, J. (2000). *Hume's Abject Failure*. Oxford: Oxford University Press.

EDDY, D. M. (1982). Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities". In D. Kanhneman, P. Slovic e A. Tversky (a cura di), *Judgment*

invece, la determinazione della probabilità delle ipotesi ricostruttive nel processo penale, si vedano Kadane e Schum (1996) e Taroni *et al*. (2006).

under Uncertainty: Heuristic and Biases, Cambridge: Cambridge University Press, pp. 249-267. Trad. it. Il ragionamento probabilistico nella medicina clinica: problemi e opportunità. In V. Crupi, G. F. Gensini e M. Motterlini (2006) (a cura di): *La dimensione cognitiva dell'errore in medicina*. Milano: Franco Angeli, pp. 45-67.

Festa, R. (1996). *Cambiare opinione. Temi e problemi di epistemologia bayesiana*. Bologna: CLUEB.

——— (1999). Bayesian Confirmation. In M. C. Galavotti e A. Pagnini (a cura di), *Experience, Reality, and Scientific Explanation*. Dordrecht: Kluwer, pp. 55-87.

——— (2004). Principio di evidenza totale, decisioni cliniche ed *Evidence Based Medicine*. In: G. Federspil e P. Giaretta (a cura di), *Forme della razionalità medica*. Soveria Mannelli: Rubbettino, pp. 47-82.

——— (2005). Il reverendo Thomas Bayes entra in corsia. L'uso dell'evidenza nella pratica clinica. *Nuova civiltà delle macchine*, **23**, pp. 39-54.

Festa, R., Buttasi, C. e Crupi, V. (2009). Evidenza incerta e probabilità delle diagnosi: estensioni dell'approccio bayesiano alla pratica clinica. In: P. Giaretta, P. Moretto, G. F. Gensini e M. Trabucchi (a cura di), *Filosofia della medicina*. Bologna: Il Mulino, 565-609.

Friedman, R. (1987). Route Analysis of Credibility and Hearsay. *Yale Law Journal*, **96**, pp. 667-742.

Frosini, B. V. (2002). *Le prove statistiche nel processo civile e nel processo penale*. Milano: Giuffrè.

Golanski, A. (2001). Why Legal Scholars Get Daubert Wrong: A Contextualist Explanation of Law's Epistemology. *Whittier Law Review*, **22**, pp. 653-721.

Gigerenzer, G. (2002). *Calculated Risks. How To Know When Numbers Deceive*. New York: Simon & Schuster. Trad. it. *Quando i numeri ingannano. Imparare a vivere con l'incertezza*. Milano: Raffaello Cortina, 2003.

Goldman, A. I. (1999). *Knowledge in a Social World*. Oxford: Oxford University Press.

Goodwin, W., Linacre, A. e Hadi, S. (2007). *An Introduction to Forensic Genetics*. Chichester: John Wiley & Sons.

Hume, D. (1748). *Enquiries Concerning the Human Understanding and Concerning the Principles of Morals*. Edizione a cura L. A. Selby-Bigge e P. H. Nidditch, Oxford: Oxford University Press, 1975. Trad. it. a cura di E. Lecaldano, *Ricerca sull'intelletto umano*. Bari: Laterza, 2009.

Jeffrey, R. (1965/1983). *The Logic of Decision*. New York: McGraw-Hill.

——— (1992): *Probability and the Art of Judgment*. Cambridge: Cambridge University Press.

——— (2004). *Subjective Probability. The Real Thing*. Cambridge: Cambridge University Press.

KADANE, J. B. e SCHUM, D. A. (1996). *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*. Chichester: John Wiley & Sons.

KUSCH, M. e LIPTON, P. (2002). Testimony: A Primer. *Studies in History and Philosophy of Science*, **33**, pp. 209-217.

LACKEY, J. e SOSA, E. (2006) (a cura di). *The Epistemology of Testimony*. Oxford: Oxford University Press.

LUCY, D. (2005). *Introduction to Statistics for Forensic Scientists*. Chichester: John Wiley & Sons.

MURA, A. (2003), *Per un bayesianesimo critico*. Introduzione all'edizione italiana di Tilliers e Green (1988), pp. IX-XLVI.

——— (2004). Teorema di Bayes e valutazione della prova. *Cassazione penale*, **44**, pp. 1808-1818.

NEWMAN, J. H. (1870). *An Essay in Aid of a Grammar of Assent*. London: Burns, Oates & Co. Trad. it. *Grammatica dell'assenso*. Milano: Jaca Book, 2005.

POLANYI, M. (1966). *The Tacit Dimension*. London: Routledge. Trad. it. *La conoscenza inespressa*. Roma: Armando, 1979.

REID, T. (1764). *An Inquiry into the Human Mind*. Ristampato a cura di R. E. Beanblossom e K. Lehrer, *Inquiry and Essays*. Indianapolis: Hackett, 1983

SACKETT, D.L., HAYNES, R.B. e TUGWELL, P. (1985). *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Boston: Little-Brown. Trad. it. *Epidemiologia clinica. Scienza di base per la medicina*. Torino: Centro Scientifico, 1988.

SCANDELLARI, C. (2005). *La diagnosi clinica. Principi metodologici del processo decisionale*. Milano: Masson.

SHOGENJI, T. (2003). A Condition for Transitivity in Probabilistic Support. *The British Journal for the Philosophy of Science*, **54**, pp. 613-616.

STELLA, F. (2003). *Giustizia e modernità. La protezione dell'innocente e la tutela delle vittime*. III ed., Milano: Giuffrè.

SHAPIN, S. (1994). *A Social History of Truth*. Chicago: University of Chicago Press.

TARONI, F., AITKEN, C., GARBOLINO, P. e BIEDERMANN, A. (2006). *Bayesian Networks and Probabilistic Inference in Forensic Science*. Chichester: John Wiley & Sons.

TILLERS, P. e GREEN, E. (1988) (a cura di). *Probability and Inference in the Law of Evidence: The Limits and Uses of Bayesianism*. Dordrecht: Kluwer. Trad. it. *L'inferenza probabilistica nel diritto delle prove. Usi e limiti del bayesianesimo*. Milano: Giuffrè, 2003.

VASSALLO, N. (2003). *Teoria della conoscenza*. Bari: Laterza.

WALTON, D. (1997). *Appeal to Expert Opinion*. Pennsylvania: Pennsylvania State Press.

WEINSTEIN, M. C. e FINEBERG, H. V. (1980). *Clinical Decision Analysis*. Philadelphia: Saunders. Trad. it. *L'analisi della decisione in medicina clinica*. Milano: Franco Angeli, 2008.

WELLS, G. L. e OLSEN, E. A. (2003). Eyewitness Testimony. *Annual Review of Psychology*, **54**, pp. 277-295.——— (2004). Teorema di Bayes e valutazione della prova. *Cassazione penale*, **44**, pp. 1808-1818.*Networks and Probabilistic Inference in Forensic Science*. Chichester: John Wiley & Sons.

TILLERS, P. e GREEN, E. (1988) (a cura di). *Probability and Inference in the Law of Evidence: The Limits and Uses of Bayesianism*. Dordrecht: Kluwer. Trad. it. *L'inferenza probabilistica nel diritto delle prove. Usi e limiti del bayesianesimo*. Milano: Giuffrè, 2003.

VASSALLO, N. (2003). *Teoria della conoscenza*. Bari: Laterza.

WALTON, D. (1997). *Appeal to Expert Opinion*. Pennsylvania: Pennsylvania State Press.

WEINSTEIN, M. C. e FINEBERG, H. V. (1980). *Clinical Decision Analysis*. Philadelphia: Saunders. Trad. it. *L'analisi della decisione in medicina clinica*. Milano: Franco Angeli, 2008.

WELLS, G. L. e OLSEN, E. A. (2003). Eyewitness Testimony. *Annual Review of Psychology*, **54**, pp. 277-295.

L&PS – Logic & Philosophy of Science
Vol. XII, No. 1, 2014, pp. 41-69

On the Relation Between Models and Hypotheses and the Role of Heuristic Hypotheses in the Construction of Scientific Models

Tjerk Gauderis
Center for Logic and Philosophy of Science
Ghent University
e-mail: tjerk.gauderis@ugent.be

ABSTRACT. In our understanding of model-based scientific practice, it has become unclear what the role of hypotheses is. Many take models and hypotheses to be more or less on the same footing; others take hypotheses to be claims about the intended representational features of models; some have even argued against the use of hypotheses in model-based science. In this paper, I argue that the first and third of these positions are untenable, while the second position applies only to a subclass of the many hypotheses actually employed in model-based scientific practice, which I call fully interpretable hypotheses. Next, I show, based on some case studies from astronomy, that many scientific hypotheses are in fact of a different type, which I call heuristic hypotheses. Therefore, I argue for a fourth position which complements the second position to provide an account of the role of these two kinds of hypotheses in model-based scientific practice.

KEYWORDS: Models, Hypotheses, Scientific Practice, Model-Based Reasoning, Heuristic Reasoning.

Published by L&PS – Logic and Philosophy of Science
http://www2.units.it/~episteme – ISSN: 1826-1043

Imagination creates events.
(Giovanni Francesco Sagredo,
letter to Galileo, 1612)

1. Introduction

As a result of a shift in focus in the philosophy of science from dealing largely with issues of scientific confirmation towards studying actual scientific practices and the questions they invoke, philosophical interest in the use of *models in science* has steadily increased in recent decades. Although early interest was mostly fueled by adherents of the so-called *semantic and structuralist views of theories*,[1] who tried to tailor their formal analyses towards the type of models actually used by scientists, it is now recognized that the structural set-theoretical meaning of models is best not equivocated with the actual practices of *model-based science*,[2] which elicit many ontological and epistemological questions in their own right. The study of these questions in relation to many actual scientific cases has led many, nowadays, to appreciate that much of science can be adequately described as model-based science, which should not be seen so much as a division between the various disciplines, but rather as a strategy that any discipline can employ to address theoretical scientific research (Godfrey-Smith 2006). The construction, manipulation and refinement of models are also now generally considered to be key scientific practices (Frigg and Hartmann 2012).

The recognition of the use of models in science has elicited a substantial amount of research to clarify the relation between this rather new addition to the jargon of scientific methodology and older inhabitants of this conceptual

[1] Until the 1970s, the received view of theories (also called the *syntactic view*) maintained the Euclidean or Aristotelian ideal of a theory as a set of axioms and a suitable logic to infer all true sentences in an ideal scientific language, supplemented with a set of correspondence rules to link theoretical terms to empirical observations. The heavy language-dependency of this view has led various scholars to develop the so-called *semantic view* of theories, in which theories are equated with a class of models, abstract mathematical structures for which the theory was true (Suppes 1960; Suppe 1977, 1989; Van Fraassen 1980). A related, *structuralist view* of scientific theories was developed by, among others, Balzer, Moulines and Sneed (1987). For a recent paper incorporating these structuralist ideas, see Leuridan (2013).

[2] The two main arguments for this distinction are that many models from actual scientific practice cannot be accommodated within the set-theoretical view of models (Downes 1992) and that, while the semantic view aims to analyze all of science in terms of models, not all actual scientific practice relies on the manipulation of models (Godfrey-Smith 2006, Weisberg 2007).

jungle, such as the relation between models and theories (e.g. the semantic view, Giere 1988), models and discovery (Redhead 1980, Morrison and Morgan 1999), models and laws of nature (Cartwright 1983, Giere 1999a) and models and data (Suppes 1962, Harris 2003). Yet, no substantial attention has been paid so far to the relation between models and hypotheses in science. The main reason why this relation has been left unattended may have to do with the fact that models and hypotheses are generally considered to belong to the jargon of two mutually exclusive conceptions of the scientific method, i.e. the inductive (model-based) view and the hypothetico-deductive view.[3]

In this paper, I investigate how hypotheses and models relate in actual model-based scientific practice, show that both are necessary concepts in understanding this practice, and show that they are mutually supportive. Apart from touching upon recent debates in the literature on models, such as those concerning the nature of their representational function and their construction, this research will reinstate a modernized concept of a scientific hypothesis, in line with model-based scientific practice, by shrugging off some of the unrealistic intuitions with which it has been burdened by the old Popperian hypothetico-deductive view.

After delineating my precise usage of the main concepts of this paper (Section 2), I will identify four stances on the relation between hypotheses and models by examining the scattered remarks that have been made in the literature and consider what objections might threaten these stances (Section 3). Then, I will look into actual scientific practice and present three case studies to expose the nature of the interplay between models and hypotheses (Section 4). This will allow me to develop my own account of how hypotheses and their role should be understood in the context of model-based science (Sections 5 and 6).

2. Some Conceptual Issues

Before I present the main arguments of this paper, some preliminaries about its scope and topic are in order. More specifically, as 'model' and certainly 'hypothesis' are often used as umbrella terms and as their meaning is often thought to be more or less self-evident, I need to specify more precisely the kind of hypotheses and models this paper will deal with. Unavoidably, this is a trade-off between catching as much as possible of the actual usage of these concepts

[3] This position is advanced, for instance, in the article of Glass and Hall (2008) that I discuss in Section 3.3.

in scientific practice and defining sufficiently coherent concepts to allow for analysis.

> **Scientific hypotheses**. I take *scientific hypotheses* to be (1) statements (2) about the empirical world (3) that have an unknown or underdetermined truth status and (4) are advanced as a tentative answer to a particular research question.

Let me expand on each part of this characterization. First, scientific hypotheses are linguistic statements or propositions, by virtue of which it always makes sense to talk about their truth status.

Second, this paper focuses only on hypotheses that make a reference to the empirical world. This excludes, along with mathematical conjectures and metaphysical claims, also hypotheses that refer exclusively to parts of and relations within a particular model. Studying the internal properties of scientific models is an important aspect of theoretical science, but conjectures of this kind are generally not what scientists refer to with the notion 'scientific hypothesis'.[4]

Third, although it makes sense to speak (typically in retrospect) of confirmed hypotheses, it is assumed that hypotheses are not known to be true. Yet this does not exclude that scientists can have a firm and even justified belief in them, certainly in later stages of research. Also, I do not assume that hypotheses are fully determined or have an unambiguous reference. As the case studies in this paper show, many actual hypotheses in early stages of research unavoidably have ambiguous or vague references. It is only afterwards, when the conceptual apparatus, requisite models and governing conditions have been developed in subsequent stages of research, that the intended hypothesis can be formulated unambiguously.

Finally, scientific hypotheses are not mere conjectural statements; they are advanced in an attempt to answer particular research questions. In other words, they are *truth-purposive*. Scientists advance them with the purpose of finding the answer to a research question by trying to determine the suggested hypotheses' truth value, even if they know that any particular hypothesis can be rejected or refined later on. Importantly, it is not required that hypotheses be compatible with the agent's background knowledge: many valuable truth-purposive

[4] This relates to Contessa's (2007) distinction between external sentences (e.g. "The emission spectrum of hydrogen can be calculated with the Bohr model") and internal sentences (e.g. "In the Bohr model of the atom, electrons orbit around the nucleus in well-defined orbits"). I consider scientific hypotheses to be external, while internal sentences belong to the model itself or a description of it.

hypotheses presented in history firmly contradicted large portions of the adopted set of beliefs or (assumed) knowledge of those who suggested them. In such cases, the agent thought that pursuing the truth value of the hypothesis he had in mind might anyway lead to certain answers to his research question, even when he was well aware that parts of his background knowledge would need revision if this particular hypothesis turned out to be true.

With this final condition, I have excluded a large class of hypotheses from my characterization of scientific hypotheses: explicit counterfactuals and belief-negating hypotheses. Although these *truth-denying hypotheses* have their role in science by virtue of, for instance, thought experiments (De Mey 2006), I consider them fundamentally different from the *truth-purposive hypotheses* this paper deals with, as (a) their truth value is explicitly known or believed to be false and (b) they neither provide a direct answer to any particular question, nor is it their suggestor's aim to determine the hypothesis' truth value (as he already assumes it to be false). Their purpose is generally to set up a line of reasoning that can lead to certain sought-for answers via a detour, such as a thought experiment or a *reductio ad absurdum* argument.[5]

> **Scientific models**. I take *scientific models* to be (1) abstract or concrete artifacts (2) purposefully created in order to be manipulated to perform particular scientific tasks (such as prediction or explanation) by exploiting certain representational relations.

Although this characterization is in line with much of the actual usage of the notion '(scientific) model' by scientists and in the contemporary literature on models,[6] I have made some restricting choices.

[5] My distinction between truth-purposive and truth-denying hypotheses relates to Rescher's (1964) classic distinction between hypotheses with an unknown truth status, on the one hand, and belief-negating hypotheses and counterfactuals, on the other. However, there is one caveat: Rescher operates in a logical framework (which assumes logical omniscience). Therefore, for Rescher, it makes no difference whether the agent explicitly believes (or knows) that the hypothesis is false, or that this is only a consequence of his set of beliefs (or knowledge). For my purposes, this distinction does matter. A hypothesis is only truth-denying if the agent explicitly believes (or knows) that it is false. When the agent thinks that it might be true, it is truth-purposive, even if it is in contradiction with his set of beliefs (or knowledge). This situation actually occurs frequently in science: as many problems are overdetermined, scientists are often willing to accept that part of their set of beliefs (or assumed knowledge) is wrong in advancing a new hypothesis.

[6] This characterization is inspired by, amongst others, the views of Giere (2004, 2010), Hughes (1997), Teller (2001), Bailer-Jones (2003), Nersessian (2008) and Knuuttila (2011) and fits accounts of actual scientists reporting on their use of models (Bailer-Jones 2002).

First, ontologically, I consider models to be either concrete or abstract models, yet my focus will be on the abstract type. It is commonly accepted that the human imagination can create such things as abstract objects and that many scientific models, such as the ideal pendulum or the Bohr model of the atom, should be understood as such.[7] As it is my purpose to determine the relation between models and hypotheses, my analysis will unavoidably focus on abstract models. In principle, this would exclude from the analysis any tangible model, such as plastic models, diagrams, descriptive texts or annotated drawings. But this should not unduly concern us, as we can straightforwardly interpret most such tangible concrete models as representations of a particular abstract model,[8] while tangible models used for the direct representation of real target phenomena, such as a wooden bridge model, are not a part of our concern here.

Second, functionally, I take models to be used to represent some target system in the real (or empirical) world.[9] This is what Giere (1999b) has called the *representational conception* of models, as opposed to the *instantial conception* of models used in the semantic and structuralist analysis of theories. According to the representational conception, the intended representational relations can be exploited for predictive or explanatory purposes by manipulating the model.

Finally, the target system the representation of which the model is used for can also be a set of data points or measurements. Such models of data (Suppes 1962) or phenomenological models, which are generally constructed via statistical methods of data analysis, are sometimes seen as temporary models requiring further explanation by deeper explanatory or constitutive models (the 1885 Balmer formula for the hydrogen emission spectrum lines, for example,

[7] For the current debates on how this should be understood, see, amongst others, Godfrey-Smith (2009), Giere (2009) and Contessa (2010) but also Teller (2001) or French (2010) for an alternative position.

[8] Interpreting concrete or tangible models as representations of abstract models only makes sense if one adopts a three-place analysis of the representation relation: representation is not purely a relation between a model M and a target T, but a relation of an agent S who uses a model M to represent a target T for some purpose (Giere 2004). As such, concrete tangible models, such as a double helix made from cardboard, can be used in two ways: either to represent directly a target phenomenon (actual DNA), or to represent an abstract model (the Crick and Watson double-helix model), which is itself used to represent that same initial target (actual DNA). Although the particular form in which an abstract model is represented does influence the scientist's actual manipulations (Knuuttila 2011, Vorms 2011), I will pay no further attention to individual (tangible) models in the present paper.

[9] This is a choice. Although this characterization fits large classes of models in science, it does not fit all models (Downes 2011).

was explained by the 1913 Bohr model of the hydrogen atom). Yet, this type of model is often employed in actual scientific practice, especially for predictive purposes (consider, for instance, the importance of the discipline of data analysis), and is highly esteemed by scientists with a strongly inductivist mindset (see e.g. Glass and Hall in Section 3.3). Therefore, it is important that our analysis of the relation between models and hypotheses should apply to this type of model as well.

3. Four Stances on the Relation Between Models and Hypotheses

In this section I review four stances that can be found in the literature. However, it should be kept in mind that none of the authors I will associate with these stances was explicitly concerned to specify the relation between hypotheses and models. In each case, the characterization was embedded in a broader research goal.

3.1. Models Are (a Particular Form of) Hypotheses and the Concepts Can Be Used Interchangeably

Although the stance that models are just a form of hypotheses is never explicitly articulated in the current literature, it is often implied by the fact that the terms 'model' and 'hypothesis' are sometimes used more or less interchangeably. The idea is that models are just a particular form of hypotheses: they are a bit more elaborate, and often have some figurative elements, but in essence they are just hypothetical suggestions which can be tested to confirm whether they conform to reality. This view is particularly appealing to people focused on explanatory and mechanistic models, as for this kind of models it is intended that the parts of the model should have an accurate one-to-one correspondence relation with the parts of the target system.

This stance, however, neglects to take into account the important and currently hot issue of the representational relation between models and the world.[10] The representational relation between hypotheses and the world is rather straightforward to specify: hypotheses are linguistic entities. Therefore,

[10] See also footnote 8. For a discussion about the representational relation between models and the world see Van Fraassen (1980, 2008), Giere (1989, 2010), Knuuttila (2010) and Downes (2011).

whether they represent the world can be indicated by stating whether they are true or false. But models are not linguistic entities.[11] Therefore, one cannot determine whether a model is literally true or false. When a model is called true (or false), this attribution normally has to be understood in a *metaphorical* or *pragmatic sense*: it indicates that the model meets the purpose for which it was designed, such as accurate prediction or explanatory power, not that it consists of literally true sentences.[12] Even if one replaces truth with a gradual notion such as accuracy, it makes for some models no sense to assess whether or not they are accurate, because they were never intended to be so because of their use of idealizations, simplifications and fictional entities.

Finally, one might suggest that, although models are maybe not linguistic in nature and hypotheses are, they might still be interchangeable if every model would have a full characterization that is purely defined in linguistic terms (and which could, hence, act as the hypothesis of this model). After all, many models in science are known purely from a textual description, and Craver (2006) has introduced in the mechanism literature the notion of the *ideally complete description of a mechanism* as the ideal for a mechanistic model. Let us grant this for a moment, and assume that there exists for each model in science an ideal fully characterizing and fully linguistic description. Such a description of a model would indeed have a truth value. But it would be true only by reference to the model itself. If we were to determine its truth value by reference to the world, it would always be false. Models include fictitious entities (e.g. point masses or frictionless planes) or describe unreal and simplified conditions (e.g. no air resistance or uniform mass density) and even if a model is very descriptive, as are particular mechanism models in biology, its (ideally) full description would be false by reference to the world because of the simplifications and abstractions it incorporates.[13] For instance, the description

[11] There is a minority position that does take models literally as linguistic entities. This view, which is embedded in a syntactic view of theories, takes models (just like theories) to be a set of statements about a target system, simplified or idealized for certain purposes (Achinstein 1968, Redhead 1980). This position, however, has to cope with similar concerns as the syntactic view of theories. Moreover, it faces the obvious objection that there can be many different linguistic descriptions of the same model. How should the canonical description be determined? As a matter of fact, I have found no recent adherents of this position.

[12] Mäki (2011) has, however, tried to define a literal truth relation for models (see also Perini 2005 on the possibility of such a truth relation for pictorial representations), but, in essence, Mäki's proposal boils down to defining the truth of a model as the truth of the assertion that the driving mechanism of the model is the same as its target mechanism (which makes him rather fit the stance discussed in Section 3.2).

[13] See also Niiniluoto (2012, 2013) about the verisimilitude of models.

might state that one part is directly adjacent to another part, while in reality there are blood vessels, tissues and fat cells in between. Or, turning the argument around, if the ideally full description of a model were to be completely true with respect to the world, there would be no model defined, as the description would be just a direct description of this part of the world. We can conclude that if such a thing as the (ideally) full description of a model existed, it would be literally false with respect to the world and, hence, counterfactual.[14] Therefore, if we were to use this construction to call models hypotheses, they would be truth-denying hypothesis and not truth-purposive hypotheses, as their creators had likely intended.[15]

3.2. Hypotheses Are Statements about the Relation Between Fully Interpreted Models and their Target Systems

This is the idea Giere has been arguing for since his book *Explaining Science: A Cognitive Approach* (1988). According to him, (theoretical) hypotheses (which, he claims, overlap considerably with the use of the notion by scientists themselves) are assertions of some sort of relationship between a model and the system it is intended to represent. In his more recent work (2004, 2008, 2010), Giere specifies this notion of hypotheses further, holding that hypotheses are claims that a fully specified and interpreted model (a model of which each element is provided with a physical interpretation) fits a particular real system more or less well, or any generalization of such claims.

If one has come to appreciate that the relation between models and the world is not simply a matter of truth (or falsehood), but may include a plenitude of possible representational relations depending on the purposes of the agent, it is quite natural to understand scientific hypotheses as specifications of the nature and fit of these representational relations. For instance, many hypotheses state that the values calculated using a particular model fit particular measurements of the target system of the model (within certain error margins), or that the mechanism represented by a particular simplified and idealized model is the same mechanism driving a real target system. Perhaps because it is

[14] This analysis relates to the analysis of the falsehood of models by Cartwright (1983) and Wimsatt (2007[1987]).

[15] An exception to the general idea that modelers aim to be truth-purposive might be toy models, which are purposefully built not to represent much but rather to experiment with the theoretical tools themselves. Toy models could also be characterized as counterfactuals, and allow, therefore, for analysis both as models and as thought experiments.

natural to understand hypotheses in this bridging role, I have found no dissenting voices on this issue amongst scholars working on scientific models.

However, although this analysis is compelling and very suitable to account for a number of hypotheses used in actual scientific practice, it does not fit the majority of hypotheses advanced and defended in this practice. The reason for this is actually straightforward. Giere's characterization of a hypothesis depends on the existence of a model that can be fully interpreted. This means that this kind of hypotheses can be stated only once a fully interpretable model has been developed, which is typically only in the closing stages of the discovery process. Giere is not to blame for this. His project is to analyze how accomplished science is structured—the starting point of his 1988 investigation was a mechanics text book. But if we want to understand the role of hypotheses in scientific practice, we should take into account that hypotheses are much more closely linked to the discovery process than to the presentation of well-established science. In the process of scientific discovery, advanced hypotheses are seldom well-specified and fully interpretable (as the case studies in the next section show).

Therefore, although we can use Giere's account for a subclass of scientific hypotheses, i.e. the fully interpretable hypotheses (see section 5), we must complement it with an account of hypotheses used in the actual process of scientific discovery.

3.3. Radical Inductionism: Hypotheses Should be Avoided in Model Construction

Recently, Glass and Hall (2008) launched a well-argued attack in the top-ranked journal *Cell* on the use of hypotheses in scientific practice. The use of hypotheses, they argue, is a relic from the old hypothetico-deductive perspective on science, which denied induction as a valid form of reasoning. According to them, the latest articulation of this obsolete view, Popper's Critical Rationalism, has been successfully challenged in the second half of the 20th century by, amongst others, Kuhn, while probability and Bayesianism gave the inductivist better tools to defend his position.

Apart from summarizing the main historical and philosophical positions in this well-known debate, Glass and Hall also argue on a pragmatic level that scientists would do better to replace top-down hypothesis testing with bottom-up inductive model-building. Framing research by hypotheses adds severe biases. Not only are negatives less valued than positives (confirmation bias), but

also researchers are rendered blind to alternative routes, as negatives are not differentiated (categorization bias). Furthermore, not all interesting research (or research proposals) can be framed by a hypothesis. A telling example was the Human Genome Project, of which, when pressed to state a research hypothesis, J. Craig Venter, a major player in the project, stated that "It is our hypothesis that this approach will be successful" (Glass 2006, p. 18).

Therefore, Glass and Hall suggest that research (and research proposals) would better start by asking an open research question, after which data collection could begin. From this data, which is more and more abundant and elaborated in this Era of Big Data, the methods of statistical data analysis might extract a first model, which would lead to new questions, further data gathering and model refinement. Nowhere should one, according to this view, have to introduce unproven premises or hypotheses.

Glass and Hall's argument has the merit that it points out to scientists and funding organizations the danger of bias if research hypotheses are given too much weight. In fact, their suggestion to frame research proposals by open research questions instead of hypotheses (as is sometimes required by funding agencies) is an interesting one, but, philosophically, their suggestion to literally eradicate all hypotheses from scientific practice in favor of model-building cannot be taken seriously.

First, hypotheses are (implicitly) present at all stages of inductive model-building. Even when the research project is framed by a research question, choices will have to be made as to which variables should be tested for in obtaining the first data set. And such choices rely on (hidden) assumptions about which variables are plausible and which are not. For instance, if one is looking for the causal factors and catalysts of a particular disease, the data set will probably contain variables such as air quality or the diet or medical history of the test subjects, but not whether they are left- or right-handed or what their favorite ice cream topping is. These decisions as to which variables to include rely on initial hypotheses concerning what might plausibly be factors in the investigated disease.

Further, inductive model-building or statistical data analysis is a discipline crucially dependent on the introduction of assumptions to mold vast data sets into models that can be manipulated for scientific purposes. The discipline has been described as being "more an art, or even a bag of tricks, than science" (Good 1983). An often cited and telling example is the curve-fitting problem: given the simplest data set of only two variables, there are already an infinity of fitting mathematical functions. Data analysts constantly have to make decisions (based on assumptions) on how to handle outliers, on the tradeoff bet-

ween simplicity and data fitting, on how the data is best represented (as this influences model construction), on how the variable is spread in the population (is it normally distributed or not?), and so on.

Finally, Glass and Hall's analysis is very focused on scientific experimentation, and their generalization is based on the old inductive idea that the whole process of scientific discovery can be reduced to inferences from data. It was precisely against this view that Nickles (1980) and other philosophers of scientific discovery have argued: discovery, they hold, is not separate from theoretical considerations and choices. As the examples in the next section will show, many models originate from theoretical considerations. Only later on, when sufficient detail is attained, can they be compared with experimental data or models of data.

However, Glass and Hall's analysis is not completely without value as their analysis does (largely) apply to the models of data and phenomenological models mentioned above. It does, however, not apply to explanatory or constitutive models about which the literature on models in science generally talks. To them is the burden to argue how this latter kind of models could be constructed without hypotheses.

3.4. Heuristic View: Hypotheses are Necessary Guidelines in Model Construction

A view opposite to the previous stance is that hypotheses somehow have a heuristic and methodological role in the process of model construction. Although this idea is sometimes mentioned (e.g. Nola and Sankey 2007, p. 25), it is more often implicitly assumed. In the remainder of this paper, I will give an explicit account of this stance, which could then complement the second position to give a full analysis of the relation between models and hypotheses.

In my view, heuristic hypotheses are direct attempts to initially answer the research question, but, precisely because the research still needs to be done, they unavoidably contain vague filler terms or black boxes and can do little more than hint at a particular direction of research. Yet, by this hinting they sketch an outline or rough blueprint, or even maybe just identify the type of the model(s) needed to substantiate the initial hypothesis. As such, they reduce the initial research problem to the more specific problem of filling in the black boxes of the model outline, resulting finally in an adequate model, of which a fully specified and interpreted hypothesis (in Giere's sense, see section 3.2), if confirmed, can provide an answer to the initial research question.

Before I give a detailed account of this position in Sections 5 and 6, I will first present in Section 4 three case studies that will allow me to benchmark this analysis.

4. Three Cases from Astronomy

In this section, I introduce three historical cases that illustrate my analysis of the role of hypotheses in model-based science. Due to space restrictions, only the first case will be fully elaborated; for the other two cases, only the key steps in my analysis will be indicated, together with further references to the literature.

4.1. The Energy Source of the Stars (1920-1930s)[16]

Around 1920, the source of stellar energy was still a mystery. By that time, Eddington had crafted the basic structural model of a (stable) star, largely confirmed by the observations at the time. His model represented stars as spheres of gas in which, at each internal point, there was an equilibrium between the inward gravitational pressure and the outward gas and radiation pressure, resulting in concentric layers of increasingly lower pressures and temperatures towards the surface.

But the source of the stellar radiant energy was still a mystery. Clearly, it could not be the result of a chemical reaction, such as exothermic oxidation (fire). Even if the Sun would be totally composed of carbon, its mass would be barely enough to radiate the Sun's current luminosity for a few thousand years. To solve this problem, von Helmholtz and Kelvin had defended in the 19th century what was later referred to as the *contraction hypothesis*, which was in turn inspired by the *nebular hypothesis* for the origin of our solar system by Kant and Laplace.[17] Von Helmholtz and Kelvin took as the source of stellar energy the inward gravitational energy provided, at first, during the accretion of the star and, after it has started to radiate, by the contraction of the star as it cools down. Using this model, Kelvin estimated the age of the solar system to be on

[16] For a thorough and detailed version of this history, see Shiaviv (2010). For a good introduction see Bahcall (2000) or Mazumdar (2005).

[17] This hypothesis situates the origin of our solar system in the gravitational collapse of a gaseous nebula (Kant 2012[1755]).

the order of ten million years—in contradiction to estimates based on the biological and geological record. For instance, Darwin suggested in the *Origin of Species*, based on some geological calculations, that the Earth was at least three hundred million years old, the time he thought to be necessary for the evolution of our current biodiversity.[18] As a matter of fact, this whole situation led to a public controversy between these two leading scientists.

At the dawn of the 20th century, better geological observations and the discovery of radioactivity quickly discredited the contraction model. The Earth (and, hence, the Sun) must be older than Kelvin's estimate. Therefore, the contraction model could not supply the requisite energy. Many looked at the new physics that was emerging, hoping it could provide an answer. Rutherford and the young Eddington suggested that radioactive elements might be the source of stellar energy, and Jeans, upon learning of Einstein's $E = mc^2$, suggested that in the extremely hot interior of stars, protons and electrons might annihilate each other, turning their mass into energy.

The experimental breakthrough that led to Eddington's initial suggestion of nuclear fusion was Ashton's measurements of the mass of *He* and *H* nuclei, finding that the mass of a *He* nucleus was only 99,3% of the combined mass of the four hydrogen nuclei it contained. This led Eddington to the hypothesis of *nuclear fusion*:

> Now mass cannot be annihilated, and the deficit can only represent the mass of the electrical energy set free in the transmutation. [...] If 5 per cent of a star's mass consists initially of hydrogen atoms, which are gradually being combined to form more complex elements, the total heat liberated will more than suffice for our demands, and we need look no further for the source of a star's energy. (Eddington 1920, p. 353)

This suggestion, although defended fiercely, is clearly just a hypothesis. Apart from Ashton's measurements, he had little or no evidence to back it up, nor did he understand how and when such a fusion process might occur. After all, one should not forget that at the time, neither the neutron nor any nucleus of atomic mass 2 or 3 had yet been discovered. Quantum mechanics had not yet been developed and the amount of hydrogen in the Sun was not yet determined. So, Eddington's hypothesis suggested that somehow four protons and two electrons (which it was thought, at the time, the *He* nucleus consisted of) come to-

[18] At the moment, it is widely accepted that the age of our solar system is approximately 4.6 billion years old, while the earliest evidence of life on Earth is about 3.5 billion years old.

gether at one position at a given time, something which Eddington knew was probabilistically nearly impossible, as is illustrated by the following quote:

> Indeed the formation of helium is necessarily so mysterious that we distrust all predictions as to the conditions required. [...] How the necessary materials of 4 mutually repelling protons and 2 electrons can be gathered together in one spot, baffles imagination. (Eddington 1926, p. 301)

Therefore, it is understandable that throughout the 1920s his hypothesis still met with competitors: Jeans kept defending a proton-electron annihilation, while Bohr even thought that in stars the conservation of energy was violated.[19] It was only after numerous contributions of the likes of Gamov, Houterman, Atkinson and Weizsäcker that Bethe (1939) finally put forward a model of stellar energy production in satisfactory agreement with the observational record, which consisted of two well-described processes that converted hydrogen into helium: the p-p chain and the CNO cycle (the latter occurring only in stars more massive than the Sun).

Let us review the various characteristics of Eddington's hypothesis of nuclear fusion. Clearly, it fits our characterization: it is a claim about the world with an unknown truth value in answer to a particular research question. In fact, it would be better to state that its truth value is underdetermined. Eddington had no idea how energy could be liberated by combining atoms. There are many possible models—some even totally different from Bethe's model with completely different concepts, elements and forces—that could still be seen as a specification of Eddington's hypothesis.[20]

Still, the credit that Eddington received for this suggestion is justified, as his suggestion was immensely important in redirecting research. In a sense, it simplified the problem of what the source of stellar energy was to the question of how hydrogen nuclei can combine so as to form helium nuclei, a process involving entities that could also be studied in laboratories on Earth. This sim-

[19] Bohr's suggestion (1986[1929]) to renounce energy conservation must be linked primordially with the problem of the continuous β spectrum (Gauderis 2014), but the way in which Bohr combined it with this problem of astrophysics, a field to which he has not contributed at all, shows how pressing the problem of stellar energy still was around 1930.

[20] Consider, for instance, also the briefly mentioned nebular hypothesis. Our current model of the origin of our solar system differs completely from what Kant had in mind (Kant 2012[1755], Palmquist 1987). Still, our current model for the origin of our solar system can be seen as a specification of Kant's severely underdetermined original hypothesis.

plification is achieved by providing an initial answer to the question of stellar energy, using a sketchy outline of a stellar model containing a black box process that somehow turns present hydrogen into helium. This is why his idea was so hugely important and why he kept on defending it and urging research in that direction for twenty years, until, finally, Bethe was able to crack open the black box.

So what is the nature of the relation here between model and hypothesis? Eddington's model was largely a black box or at most a rough outline, so Giere's characterization of hypotheses does not apply to his hypothesis, because Eddington's model could not be fully specified or provided with a physical interpretation. His hypothesis was heuristic in nature. Only once Bethe's model was available could one say that Eddington's hypothesis, refined by stating that the "combination of hydrogen atoms" has to occur according to Bethe's model, is a hypothesis in Giere's sense: a claim that a fully interpreted model fits a target system.

4.2. The Nice Model (2000s)

In 2001, simulations of the model specified by the nebular hypothesis (describing the origin of our solar system), with reasonable assumptions for the initial conditions, confirmed the idea raised a few years earlier that Neptune could not have become such a large planet at such a great distance from the Sun (Stewart & Levison 1998, Levison & Stewart 2001)—a research problem that triggered, amongst other possible solutions, the hypothesis that Neptune initially formed nearer to the Sun and then migrated out (Thommes *et al.* 1999). Yet this hypothesis was nearly meaningless, as no available model showed how such a migration could have occurred. In 2005, in a series of three papers in *Nature*, the Nice model[21] was presented (Tsiganis *et al.* 2005, Morbidelli *et al.* 2005, Gomes *et al.* 2005). This model postulates that four billion years ago there was a period in which Jupiter and Saturn were in 2:1 orbital resonance.[22] This led to a global gravitational instability in our solar system that caused the outer pla-

[21] This model, named after the French Mediterranean city where the research was conducted, is generally represented and explored via computer simulations. It is yet an open debate how models and simulations relate. See among others Humphreys (2004), Frigg and Reiss (2009), Winsberg (2010).

[22] This means that one orbit of Saturn takes exactly as long as two orbits of Jupiter. Hence, the direction where they line up (with respect to the Sun) and coerce their combined gravitational pull on the rest of the solar system, remained the same for several thousands of years.

nets to move from orbits much nearer to the Sun outwards to their current trajectories. Furthermore, simulations of this model showed that it also explained many other curious features of our solar system, such as the Late Heavy Bombardment (that caused the many lunar craters), the heavy eccentricities of the outer planets' orbits, and the Trojan satellites locked in Jupiter's orbit. In subsequent years, improved simulations and new explanations of further features of the solar system, such as the characteristics of the Kuiper belt, have made the Nice Model generally accepted (Crida 2009).

The Nice model is clearly a very different type of model than the stellar model from Section 4.1. Where the stellar model was mainly a very general theoretical model applicable to any star, the Nice model is an applied model tailored to our solar system, established by numerous computer simulations, in which mainly the initial conditions were sought that, given the principles of a well-known theory (Newtonian dynamics), could result in the observed specificities of our solar system.

Still, we find here the same type of relation between the model and the heuristic hypothesis that led to its development. The initial suggestion, i.e. that Neptune formed closer to the Sun and then migrated out due to gravitational forces in our solar system, provided a first tentative but direct answer to the research question of why Neptune was so massive. Yet, this suggestion was largely vacuous without an exact model or initial conditions to specify how such a migration might have occurred. On the other hand, it was precisely the persuasive plausibility of this initial heuristic hypothesis that motivated and coordinated a large research effort to conduct the numerous computer simulations that led to the substantiation of this claim by explicating the unknown mechanism of Neptune's migration. Only now that this model has been built can we reformulate the hypothesis as a fully interpretable hypothesis in Giere's sense: Neptune formed closer to the Sun and then migrated out according to the conditions and the mechanism described by the Nice model.

4.3. Dark Matter (1930s-Present)[23]

Notwithstanding some earlier references to dark stars or matter, the start of the modern search for dark matter is to be found in Zwicky (2009[1933]). Having found that the galaxies in the Coma Cluster rotate way too high around their center to be explained by the gravitational forces of the visible stars, he sug-

[23] Classic histories of dark matter are Trimble (1987), Van den Bergh (1999), Rubin (2003).

gested that dynamical models of galaxies should incorporate the presence of non-visible dark matter to explain the observed rotational speeds. In the following decades, the problem was largely cast aside, although a growing number of studies for different galaxies confirmed the high rotational speeds. Gradually, more galactic models incorporating dark matter were advanced, attributing more and more features to it. For instance, Ostriker and Peebles (1973) calculated that, in contrast with visible matter which is mostly found in the galactic disk, dark matter is mostly present in the galactic halo. The enumeration of the various indications of its existence in a highly-influential review paper of Faber and Galagher (1979) convinced most astrophysicists of its existence by 1980. In subsequent decades, we saw an enormous increase in the number of suggestions to characterize dark matter, while some possibilities, such as neutrinos or brown dwarfs and other massive dark astronomical bodies (so-called MACHOs), could already be ruled out. At the same time, other hypotheses have been raised to address the initial problem of the galactic rotation curves (e.g. the MOND hypothesis proposed a modification of Newtonian dynamics), but we also saw an increase in the use of the concept 'dark matter' in models that explain other features of our galaxy, such as gravitational lensing or fluctuations in the cosmic background radiation. Nowadays, the fact that the concept is incorporated in virtually any successful galactic or cosmological model is considered by almost everyone to be sufficient proof of its existence. On the other hand, although some possibilities have already been ruled out and some characteristics have already been determined, there is still no satisfactory account of the nature of dark matter. The best guess at present is that it consists of unknown weakly interacting massive particles (so-called WIMPs).

This final case, about a not yet specified hypothetical entity, might seem different from the other two cases. Yet, also here we can find the same interplay between hypotheses and models, the only difference being that, in this case, most of our present models cannot be fully interpreted and specified (in Giere's sense), as dark matter is not yet fully understood. Zwicky's initial heuristic hypothesis, i.e. that there exists a large amount of dark matter in galaxies, has, despite its neglect at the time it was proposed, redirected much research toward specifying the nature of this unknown type of matter and supplementing this claim with suitable models. But, although galactic and cosmological models including dark matter have been substantially refined over the years and have become the only widely accepted models, and even if these models can be operationalized for some explanatory or predictive purposes, the notion 'dark matter' still remains something of a black box in these models.

5. Heuristic and Fully Interpretable Hypotheses

Before turning to the relation between models and hypotheses in model-based scientific practice, let me first draw more precisely the distinction I have been hinting at between two types of hypotheses: *heuristic hypotheses* and *fully interpretable hypotheses*, a distinction that draws on Craver's (2006) distinction between *mechanism sketches* and *ideally complete descriptions of mechanisms*.[24]

A *fully interpretable hypothesis* is a hypothesis the meaning of which (or any part of which) leaves no room for vagueness or ambiguity. In other words, expressions of such hypotheses do not contain any unexplained *filler terms*, terms such as 'process,' 'to interact,' or 'entity' that have a broad and generic meaning covering up some uncertainty, imprecision or unknown details.[25] Hence, these hypotheses are fully expressed in terms with a precise meaning, which is provided either by the conceptual framework of the field the researcher is working in, or by the researcher himself by means of suitable models. *Heuristic hypotheses*, on the other hand, do contain such unspecified and generic filler terms.[26]

The main idea is that heuristic hypotheses are both unavoidable and useful in the early stages of scientific discovery, as they sketch an early blueprint or incomplete model without committing one to too much (yet unknown) detail. A heuristic hypothesis suggests that research should proceed in a particular direction, i.e. that it aims to fill gaps in the incomplete model instead of trying to address the general research question directly. Fully interpretable hypotheses, on the other hand, can be put forward only after the construction of a full model that specifies how the hypothesis (which is a claim about a part of reality) should be interpreted precisely and under what conditions it should hold. Therefore, in principle, it is possible to design a conclusive experiment to verify whether a fully interpretable hypothesis holds, while heuristic hypotheses can seldom be tested conclusively due to their vagueness and ambiguity. Experiments in this case mostly aim to refine the model and reduce the vagueness and ambiguity.

[24] The notion of a mechanism sketch had already been introduced in the seminal paper on mechanisms by Machamer, Darden and Craver (2000).

[25] This does not mean that the hypothesis cannot contain any approximations or abstractions.

[26] Heuristic hypotheses are, however, still real truth-conductive hypotheses, which aim to provide directly answers to particular research questions.

Before I add some further remarks and consider some examples, it is useful first to explain how these two types of hypotheses relate. As the main criterion that distinguishes these two types is the amount of precision in the expression of the hypotheses, the two distinguished types are actually the extremes of a continuum. Moreover, as it is an unwieldy (if even possible) task to specify all relevant conditions for a particular hypothesis, it is clear that the idea of a fully interpretable hypothesis is actually an idealization (as Craver could only speak of ideally complete descriptions of mechanisms). Therefore, at first sight, it seems as if there exist only heuristic hypotheses, interpretable to a greater or lesser extent. In scientific practice, however, some hypotheses are clearly considered to be sufficiently unambiguous and interpretable, allowing them to be tested conclusively. Therefore, for our purposes, we can evade this conclusion by allowing for a pragmatic or social epistemological threshold of precision sufficient for full interpretability. A hypothesis can be considered *sufficiently fully interpretable* if it invokes no disagreement in the research community as to which is its meaning. Yet the flip side of adopting this social epistemological criterion is that a single researcher cannot himself decide whether a hypothesis is fully interpretable. Also, that a particular hypothesis is considered to be fully interpretable at a certain point in time does not warrant that it will remain so in the indefinite future.

A few further remarks are in order concerning the concept of filler terms, including some examples. First, what counts as a filler term is topic dependent. For instance, the phrase ‘exerting a force’ has a precise meaning in physics, while in economics this would be a filler term for an unspecified process of influence. Having said this, the fact that so many words in various fields can be considered to have a precise meaning is precisely because of the cumulative processes of abstraction and concept formation in these sciences. Therefore, whether a phrase counts as a filler term or whether it has a precise meaning (in a particular reference framework) is dependent on the stage of development in the field. Let me return to the examples presented in Section 4. When Eddington in 1920 spoke of “the combination of hydrogen atoms” and somewhat later even used the term ‘nuclear energy,’ these concepts were certainly filler terms. Despite having good arguments why focusing on a possible transition from nuclear mass to energy could possibly solve the problem of stellar energy, he did not have any account of how this energy could be released from the nucleus and why this process occurred in stars. It was only after the acceptance of Bethe’s 1939 nuclear fusion models for the pp-cycle and CNO-cycle that the term ‘nuclear energy’ received a precise meaning in astrophysics.

Also, filler terms generally gain precision only gradually. For instance, while the concept 'dark matter' was at first a pure filler term to indicate the possibility of unobserved but present matter, the term has gained some precision and delineation over the past decades. It is now accepted that dark matter mostly resides in galactic halos, that there is at least five times more dark matter than regular matter, that it consists of weakly interacting massive unknown particles (WIMPS), which move at relatively slow speeds (with respect to the speed of light) and which are electrically neutral, etc. Yet no astronomer at present would claim that the concept of dark matter is fully understood and precisely defined.

Finally, the given examples might suggest that in the discovery process filler terms themselves always gain a more precise meaning. This happens, such as in the case of 'nuclear energy' or 'dark matter', but more often vague filler terms are replaced with more meaningful descriptions, names or acronyms, such as 'nuclear fusion' or 'WIMP'.

So how do these hypotheses relate to models? For fully interpretable hypotheses, as indicated in Section 3.2, I follow Giere in the sense that such hypotheses are claims that a fully specified model provided with a physical interpretation fits a target system more or less well. This idea can now be extended to heuristic hypotheses. Heuristic hypotheses are also claims that a particular model or model type fits a target system more or less well, but in this case, as the models are just bare model sketches containing black boxes labeled by filler terms, this claim should be understood as the weaker claim that a full specification and interpretation of the model sketch that would fit the target system is possible. But in providing such a model sketch, the initial research question is already partially answered, while at the same time the direction is shown for future research, i.e. to fill in the black boxes.

6. The Role of Hypotheses in Model-Based Scientific Practice

Let me now spell out the role of these two types of hypotheses in the process of scientific discovery in model-based science. This view will incorporate the two theses I defended above, i.e. that hypotheses are necessary in the process of model construction and that hypotheses that are not fully interpretable are valuable and even needed in this process.

In general, research aimed at constructing models is triggered by a research question or trigger. In her monograph on abductive reasoning (the inference from observations to explanatory hypotheses), Aliseda (2006) distinguishes

between anomalies and novelties as the two types of observational triggers for abductive reasoning. This classification can be adopted for our current purposes if we keep in mind the main criticism developed by Nickles (1980) and other scholars of scientific discovery against the idea that abductive reasoning could be the logic of scientific discovery (as suggested by Hanson 1958), i.e. that abductive reasoning neglects the triggering role of theory in scientific discovery. Much research is fueled by theoretical considerations, but also here we can distinguish between questions triggered by contradictions (related to experimental anomalies) and questions triggered by lacunas (related to experimental novelties). Therefore, I conceive of four triggers for research aiming at the construction of models: *experimental* (or *observational*) *novelties*, *experimental* (or *observational*) *anomalies*, *theoretical gaps* or *lacunas* and *theoretical contradictions*.

In model-based discovery, these triggers or research questions are answered at the end of the research process by proposing a model and claiming that its similarities with the target system can be exploited to sufficiently address the research question, or, in other words, by stating a (sufficiently) interpretable hypothesis whose claim is sufficiently verified.

As the model is only linked to the trigger or research question through a hypothesis claiming its fit, such a linking hypothesis, constituting the (partial) answer to the research question, must be present through all stages of model construction; though in the early stages it will heuristic in nature, not fully interpretable.

Now we have to investigate what the role of these heuristic hypotheses is in the research process itself. If we take a constraints-based view of scientific discovery, the view Nickles (1978) developed in the tradition of scientific research as problem solving, we can conceive of a scientific problem (or research question) as a set of constraints. Progressing on a problem consists in manipulating these constraints such that the problem turns into a simpler problem or a problem that is easier to solve.

In the case of suggesting a heuristic hypothesis as an initial partial answer to a research problem, one deliberately adds a constraint: however vague a heuristic hypothesis might be, it excludes particular solutions and direct research in a particular direction. As such, one progresses on the problem by reducing it to a simpler problem, though always at the risk that one will not find a solution along these lines (if the heuristic hypothesis turns out to have been a wrong path from the start). After reducing the initial research problem to the simpler problem of finding a suitable model to fill in the filler terms, the heuristic hypothesis remains important as the link between the reduced problem

and the initial research question, as it shows how the latter can be answered by means of the answer to the reduced problem.

Let me illustrate this role of heuristic hypotheses with some of the cases of Section 4. Eddington reduced the open problem of stellar energy (a theoretical gap) to the more restricted problem of how hydrogen could combine so as to form helium. After the problem was reduced to finding a suitable model for this combination, Eddington's hypothesis remained the link that allowed the answer to this reduced problem, namely Bethe's model of hydrogen fusion, to be used to answer the initial research question of where stellar energy originated. By the time Bethe's model was developed, Eddington's hypothesis could be considered a fully interpretable hypothesis.

Similarly, the research question of the improbable accretion of Neptune (an observational anomaly) was reduced by the initial heuristic hypothesis to the more straightforward problem of constructing a model and determining the initial conditions for an outward-directed gravitational slingshot of a planet within our solar system. Only when such a model—the Nice model—was constructed through numerous computer simulations could the original hypothesis that Neptune initially formed much closer to the Sun and migrated outwards be considered as the fully interpretable answer to the initial research question or trigger.

A final thing to address is the fact that many research triggers have the form of an anomaly or a contradiction. Heuristic hypotheses addressing such research questions unavoidably sometimes contradict major parts of the agent's (assumed) background knowledge. Yet as history shows, this clearly does not prevent scientists from coming up with heuristic hypotheses for such overdetermined problems. In such cases, scientists reason according to what Rescher (1960) has called belief-negating (or even knowledge-negating) hypothetical reasoning: they assume the hypothesis while retaining all beliefs from their belief set that are compatible with it, and suspending judgment on beliefs that are contradictory to it. For instance, in the case of Neptune, researchers had at first to suspend judgment on the idea that the planets in our solar system were formed where we observe them today, while retaining acceptance of full Newtonian dynamics. The beliefs compatible with the heuristic hypothesis then become the basis for solving the reduced problem of the construction of a suitable model to interpret this heuristic hypothesis. Only once the model is verified and the research question answered, the initially incompatible beliefs on which judgment was suspended can then be revised.

7. Conclusion

In this article I have addressed the relation between models and hypotheses in model-based science. After reviewing and pointing out the shortcomings of various stances in the literature, I presented my own view on the matter.

First, a distinction has to be made between heuristic hypotheses and fully interpretable hypotheses. Heuristic hypotheses are initial and partial answers to research questions that contain necessarily vague filler terms, yet sketch the outline for the type of model that might be needed to answer the research question. Fully interpretable hypotheses, on the other hand, are claims concerning how a fully constructed model can be used to provide an answer to the research question.

Next, I have shown in this article, by examining three cases from astronomy, how initial heuristic hypotheses fuel the process of model construction and how, once the requisite models are built, they gradually evolve into fully interpretable hypotheses that can, if verified, serve as answers to the initial research questions.

REFERENCES

ACHINSTEIN, P. (1968). *Concepts of Science. A Philosophical Analysis*. Baltimore: Johns Hopkins Press.

ALISEDA, A. (2006). *Abductive Reasoning. Logical Investigation into Discovery and Explanation*. Dordrecht: Springer.

BAHCALL, J. (2000). How the Sun Shines. Retreived October 15th, 2013 from www.nobelprize.org/nobel_prizes/themes/physics/fusion/.

BAILER-JONES, D. (2002). Scientists' Thoughts on Scientific Models. *Perspectives on Science*, **10**, 275-301.

——— (2003). When Models Represent. *International Studies in the Philosophy of Science*, **17**, 59-74.

BALZER, W., ULISES MOULINES, C., & SNEED, J. (1987). *An Architectonic for Science. The Structuralist Program*. Dordrect: Reidel.

BETHE, H. (1939). Energy Production in Stars. *Physical Review*, **55**, 434-456.

BOHR, N. (1986[1929]). β Ray Spectra and Energy Conservation. In R. Peierls (Ed.), *Niels Bohr Collected Works. Vol. 9: Nuclear Physics (1929-1952)* (pp. 85-89). Amsterdam: North Holland Physics. (Original unpublished manuscript written 1929.)

CARTWRIGHT, N. (1983). *How the Laws of Physics Lie*. New York: Oxford University Press.

CONTESSA, G. (2010). Scientific Models and Fictional Objects. *Synthese*, **172**, 215-229.

CRAVER, C. (2006). When Mechanistic Models Explain. *Synthese*, **153**, 355-376.

CRIDA, A. (2009). Solar System Formation. Retrieved from arXiv:0903.3008v1 [astro-ph.EP]

DE MEY, T. (2006). Imagination's Grip on Science. *Metaphilosophy*, **37**, 222-239.

DOWNES, S. (1992). The Importance of Models in Theorizing: A Deflationary Semantic View. In D. Hull, M. Forbes, & K. Okruhlik (Eds.), *PSA 1992, vol. 1* (pp. 142-153). East Lansing, MI: Philosophy of Science Association.

——— (2011). Scientific Models. *Philosophy Compass*, **6**, 757-764.

EDDINGTON, A. (1920). Presidential Address to section A of the British Association at Cardiff (24th Augustus 1920). *Observatory*, **43**, 353.

——— (1926). *The Internal Constitution of Stars*. Cambridge: At the University Press.

FABER, S., & GALLAGHER, J. (1979). Masses and Mass-to-Light Ratios of Galaxies. *Annual Review of Astronomy and Astrophysics*, **17**, 135-187.

FRENCH, S. (2010). Keeping Quiet on the Ontology of Models. *Synthese*, **172**, 231-249.

FRIGG, R., & REISS, J. (2009). The Philosophy of Simulations: Hot New Issues or Same Old Stew? *Synthese*, **169**, 593-613.

FRIGG, R., & HARTMANN, S. (2012). Models in Science. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2012 edition)*. Retrieved from http://plato.stanford.edu/archives/fall2012/entries/models-science/.

GAUDERIS, T. (2014). To Envision a New Particle or Change an Existing Law? Hypothesis Formation and Anomaly Resolution for the Curious Case of the β Decay Spectrum. *Studies in History and Philosophy of Modern Physics*, **45**, 27-45.

GLASS, D. (2006). *Experimental Design for Biologists*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

GLASS, D., & HALL, N. (2008). A Brief History of the Hypothesis. *Cell*, **134**, 378-381.

GIERE, R. (1988). *Explaining Science. A Cognitive Approach*. Chicago: University of Chicago Press.

——— (1999a). *Science withouth Laws*. Chicago: University of Chicago Press.

——— (1999b). Using Models to Represent Reality. In: L. Magnani, N. Nersessian & P. Thagard (Eds.), *Model-Based Reasoning in Scientific Discovery* (pp. 41-57). New York: Kluwer/Plenum.

——— (2004). How Models are Used to Represent Reality. *Philosophy of Science*, **71**, 742-752.

——— (2008). Models, Metaphysics and Methodology. In: S. Hartmann, C. Hoefer & L. Bovens (Eds.), *Nancy Cartwright's Philosophy of Science* (pp. 123-133). New York: Routledge.

——— (2009). Why Scientific Models Should not be Regarded as Works of Fiction. In M. Suárez (Ed.), *Fictions in Science. Philosophical Essays on Modeling and Idealisation* (pp. 248-258). London: Routledge.

——— (2010). An Agent-Based Conception of Models and Scientific Representation. *Synthese*, **172**, 269-281.

Godfrey-Smith, P. (2006). The Strategy of Model-based Science. *Biology and Philosophy*, **21**, 725-740.

——— (2009). Models and Fictions in Science. *Philosophical Studies*, **143**, 101-116.

Gomes, R., Levison, H., Tsiganis, K., & Morbidelli, A. (2005). Origin of the Cataclysmic Late Heavy Bombardment Period of the Terrestrial Planets. *Nature*, **435**, 466-469.

Good, I. (1983). The Philosophy of Exploratory Data Analysis. *Philosophy of Science*, **50**, 283-295.

Hanson, N. R. (1958). *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge: Cambridge University Press.

Harris, T. (2003). Data Models and the Acquisition and Manipulation of Data. *Philosophy of Science*, **70**, 1508-1517.

Hughes, R. (1997). Models and Representation. *Philosophy of Science*, **64** (supplement), S325-S336.

Humphreys, P. (2004). *Extending Ourselves. Computational Science, Empiricism, and Scientific Method*. Oxford: Oxford University Press.

Kant, I. (2012[1755]). Universal Natural History and Theory of the Heavens. In: E. Watkins (ed.), *Natural Science. The Cambridge Edition of the Works of Immanuel Kant* (pp. 182-308). Cambridge: Cambridge University Press. (Original published in German in 1755.)

Knuuttila, T. (2011). Modelling and Representing: An Artefactual Approach to Model-Based Representation. *Studies in the History and Philosophy of Science*, **42**, 262-271.

Kroon, F. & Voltolini, A. (2011). Fiction. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2011 Edition)*. Retrieved from http://plato.stanford.edu/archives/fall2011/entries/fiction/.

Leuridan, B. (2013). The Structure of Scientific Theories, Explanation, and Unification. A Causal-Structural Account. *The British Journal for the Philosophy of Science*, in print. doi:10.1093/bjps/axt015.

LEVISON, H. & STEWART, G. (2001). Remarks on Modeling the Formation of Uranus and Neptune. *Icarus*, **153**, 224-228.

MACHAMER, P., DARDEN, L. & CRAVER, C. (2000). Thinking about Mechanisms. *Philosophy of Science*, **67**, 1-25.

MÄKI, U. (2011). Models and the Locus of their Truth. *Synthese*, **180**, 47-63.

MAZUMDAR, I. (2005). Nucleosynthesis and Energy Production in Stars: Bethe's Crowning Achievement. *Resonance*, **10**, 67-77.

MORBIDELLI, A., LEVISON, H., TSIGANIS, K., & GOMES, R. (2005). Chaotic Capture of Jupiter's Trojan Astreroids in the Early Solar System. *Nature*, **435**, 462-465.

MORRISON, M. & MORGAN, M. (1999). Models as Mediating Instruments. In: M. Morgan & M. Morrison (Eds.), *Models as Mediators. Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.

NERSESSIAN, N. (2008). *Creating Scientific Concepts*. MIT Press, Cambridge, Massachussets.

NICKLES, T. (1978). Scientific Problems and Constraints. In: P. Asquith & I. Hacking (Eds.), *PSA 1978. Proceedings of the Biennial Meeting of the Philosophy of Science Association* (pp. 134-148). East Lansing, Michigan: Philosophy of Science Association.

——— (1980). Introductory Essay: Scientific Discovery and the Future of Philosophy of Science. In: T. Nickles (Ed.), *Scientific Discovery, Logic and Rationality* (pp. 173-183). Dordrecht: Reidel.

NIINILUOTO, I. (2012). The Verisimilitude of Economic Models. In: A. Lehtinen, J. Kuorikoski & P. Ylikoski (Eds.), *Economics for Real: Uskali Mäki and the Place of Truth in Economics* (pp. 65-80). London: Routledge.

——— (2013). Models, Simulations and Analogical Inference. In: V. Karakostas and D. Dieks (Eds.), *EPSA11: Perspectives and Foundational Problems in Philosophy of Science* (pp. 19-27). Dordrecht: Springer.

NOLA, R., & SANKEY, H. (2007). *Theories of Scientific Method. An Introduction*. Durham: Acumen.

OSTRIKER, J. & PEEBLES, P. (1973). A Numerical Study of the Stability of Flattened Galaxies: Or, can Cold Galaxies Survive? *The Astrophysical Journal*, **186**, 467-480.

PALMQUIST, S. (1987). Kant's Cosmogony Re-Evaluated. *Studies in History and Philosophy of Science*, **18**, 255-269.

PERINI, L. (2005). The Truth in Pictures. *Philosophy of Science*, **72**, 262-285.

REDHEAD, M. (1980). Models in Physics. *The British Journal for the Philosophy of Science*, **31**, 145-163.

RESCHER, N. (1964). *Hypothetical Reasoning*. Amsterdam: North-Holland.

RUBIN, V. (2003). A Brief History of Dark Matter. In M. Livio (Ed.), *The Dark*

Universe. Matter, Energy and Gravity (pp. 1-13). Cambridge: Cambridge University Press.

RUTHERFORD, E. (1911). The Scattering of α and β Particles by Matter and the Structure of the Atom. *Philosophical Magazine*, Series 6, **21**, 669-688.

SHIAVIV, G. (2010). *The Life of Stars: The Controversial Inception and Emergence of the Theory of Stellar Structure*. Heidelberg: Springer Verlag.

STEWART, G. & LEVISON, H. (1998). On the Formation of Uranus and Neptune. *Proceedings of the 29th Annual Lunar and Planetary Science Conference*, abstract no. 1960.

SUPPES, P. (1960). A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences. *Synthese*, **12**, 287-301.

——— (1962). Models of Data. In: E. Nagel, P. Suppes & A. Tarski (Eds.), *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress* (pp. 252-261). Stanford: Stanford University Press.

SUPPE, F. (1989). *The Semantic Conception of Theories and Scientific Realism*. Urbana, Illinois: University of Illinois Press.

——— (Ed.) (1977). *The Structure of Scientific Theories* (2nd ed.). Urbana, Illinois: University of Illinois Press.

TELLER, P. (2001). Twilight of the Perfect Model Model. *Erkenntnis*, **55**, 393-415.

THOMMES, E., DUNCAN, M., & LEVISON, H. (1999). The Formation of Uranus and Neptune in the Jupiter-Saturn Region of the Solar System. *Nature*, **402**, 635-638.

TRIMBLE, V. (1987). Existence and Nature of Dark Matter in the Universe. *Annual Review of Astronomy and Astrophysics*, **25**, 425-472.

TSIGANIS, K., GOMES, R., MORBIDELLI, A., & LEVISON, H. (2005). Origin of the Orbital Architecture of the Giant Planets of the Solar System. *Nature*, **435**, 459-461.

VAN DEN BERGH, S. (1999). The Early History of Dark Matter. *Publications of the Astronomical Society of the Pacific*, **111**, 657-660.

VAN FRAASSEN, B. (1980). *The Scientific Image*. Oxford: Oxford University Press.

——— (2008). *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press.

VORMS, M. (2011). Representing with Imaginary Models: Format Matters. *Studies in the History and Philosophy of Science*, **42**, 287-295.

WEISBERG, M. (2007). Who is a Modeler? *The British Journal for the Philosophy of Science*, **58**, 207-233.

WIMSATT, W. (2007[1987]). False Models as a Means to Truer Theories. In W. Wimsatt (Ed.), *Re-engineering Philosophy for Limited Beings* (pp. 94-132). Cambridge, Mass.: Harvard University Press. (Original article published in 1987.)

WINSBERG, E. (2010). *Science in the Age of Computer Simulation*. Chicago: University of Chicago Press.

ZWICKY, F. (2009[1933]). The Redshift of Extragalactic Nebulae. *General Relativity and Gravitiation*. **41**, 207-224. (Original article published in German in 1933.)

L&PS – Logic & Philosophy of Science
Vol. XII, No. 1, 2014, pp. 71-90

Revisiting and Type-Freeing Church's Approach to Semantic Paradoxes

Pierdaniele Giaretta
Department of Philosophy, Sociology, Pedagogy and Applied Psychology
University of Padua
e-mail: pierdaniele.giaretta@unipd.it

ABSTRACT. We simplify and slightly modify the theory of types that Church provided with semantic primitive predicates. Two goals are pursued. The first goal is to present a simple application of Church's approach to paradoxes and to point out some aspects of this approach. The second, perhaps more interesting, goal is to show that when type distinctions are removed some basic Churchian principles need to be restricted and different restrictions correspond to Tarski's and Kripke's different approaches to truth. Finally, we briefly hint at how to move in the direction of Field's recent approach to truth by giving up some specific essential points of the Churchian framework.

KEYWORDS: Church, Paradox, Type, Semantic Value, Truth.

1. Introduction

From Bertrand Russell's perspective, paradoxes depend on a sort of linguistic inadequacy, which essentially consists in the failure to recognize some funda-

Published by L&PS – Logic and Philosophy of Science
http://www2.units.it/~episteme – ISSN: 1826-1043

mental ontological distinctions. If the relevant ontological distinctions are properly established and respected by the language, paradoxes cannot arise any more. Alonzo Church showed how the Russellian fundamental ontological distinctions should be applied to the relations of meaning between the linguistic expressions and the meant entities. Church's integration does not add anything basically new to Russell's classic approach. However, his integration is provided in a very clean and elegant formulation of Russell's type theory. Here we present a revision of Church's formulation and then free it from the Russellian ontological distinctions. Once we have omitted the type distinctions, some basic principles concerning the semantic relations are to be restricted, and different ways of conceiving and restricting them correspond to the different approaches to truth by Alfred Tarski and Saul Kripke.

Our starting point is Church's article "Comparison of Russell's Resolution of the Semantical Antinomies with that of Tarski", published on *The Journal of Symbolic Logic* in 1976. It was qualified a magisterial work, but because it falls outside of the main lines of research on truth and paradoxes it appears to be now considered just a beautiful piece of antique. Since we are convinced of the relevance of Church's approach beyond the Russellian framework which it belongs to, we'll change it partially so that it can more easily understood and connected with other major approaches to the theme of truth and paradoxes, mainly those of Tarski and Kripke.

We will introduce the following simplifying modifications. First, as concerns the language of Church's theory first expounded in Church (1976) and corrected in Church (1984), it is assumed that some individual constants are available which are to be taken as names for formulas. Second, a ternary val semantic predicate is replaced by a binary val semantic predicate. This modification is motivated by a correction introduced by Church (1984), but the binary val is not meant to serve all the purposes of the ternary val. In what may amount to a Russellian perspective (though not Russell's original perspective), this apparently strange feature can be used to express truth for sentences without resorting to propositions.

Third, a postulate-schema of Church (1976, 1984) is replaced with two different postulate schemata. The final aim is to separate two roles pertaining to the single postulate schema of Church (1976, 1984). If only simple type distinctions are adopted, a paradox only superficially similar to Russell's paradox is derivable. The derivation is simple since it only involves the quantification of properties and, in particular, it does not involve diagonalization. Even if it does not prove anything more than the Grelling paradox, it can demonstrate in a more direct way that the intuitive relation of being-true-of is paradoxical.

As is well known, the introduction of type ramification provides a special way of solving semantic paradoxes, and reducibility does not reintroduce them if distinctions of languages are not ignored. This is mentioned only briefly because both of these facts are based on results already well known from the literature.

Instead, we emphasize the modified Churchian framework independently of type distinctions. Our aim is to show that it is useful to analyze the source of the truth paradoxes. In particular, the different roles of the two kinds of Churchian postulates for the derivation of the Tarskian conditionals are pointed out. With reference to these two postulates, the Tarskian and the Kripkean perspectives can be accounted for as the consequences of two different theoretical choices that correspond to two general ways of conceiving the semantic values.

Existence and role of semantic values interact with logic. A specific interaction can be focused on with reference to our Churchian framework and provide a possible way to think of Field's radically different perspective.

2. From Church (1976) to Church (1984) and beyond

The language of Church (1976) and Church (1984) is a language built from typed variables and primitive constants. Church presents his language *L* with 'an unspecified list of primitive constants, each of definite r-type'; furthermore, he adds in note 5 that, 'it is intended that additions to the list of primitive constants may be made from time to time, so that Russell's formalized language is an open language rather than a language of fixed vocabulary' (Church 1976, p. 749). Such an openness of the language does appear to make the expressibility of its syntax relative to the various stages of the process of enlarging the list of the primitive constants. Indeed, Church does not care that his language expresses its own syntax: he only takes variables and formulas as values of variables for individuals and, since higher order quantification is available, that is enough to express a kind of self-reference that generates paradoxes. However, if 'additions to the list of primitive constants may be made from time to time' and formulas are values of variables for individuals, also being a value of a variable for individuals appear to be relative to a time. This feature is perhaps not so relevant until reducibility is taken into account.[1] It surely makes additional difficulties, if we as-

[1] See below. Some relativization would be needed in order to account for the denial of the expressibility of some concepts, which Church clearly admits.

sume the language is provided with names, precisely individual constants, for the expressions of the language, as we will do. Without denying that such difficulties can be successfully dealt with, let us give up the openness of the language for the sake of simplicity.

Introducing individual constants for the formulas of the language itself, which will allow us to state separately and easily the questions of the existence of a semantic value and of its constraints, is, however, a delicate matter. Some countable individual constants are bound to be interpreted as names of formulas by means of a suitable map such that every formula has a name and only one name. It follows that for every interpretation the domain of individuals is to be infinite. This limitation, which is already a consequence of Church's assumption that variables and formulas are values of variables for individuals, might be granted. However, how to fix a 1-1 function from some individual constants onto formulas? It would be easy, and dispensable, if the setting were so developed to include arithmetic. Without arithmetization available, we can presuppose some sort of systematic 1-1 coupling based on a list of individual constants and an enumeration of the formulas of the language. Identifying such a coupling with naming could be put into question, but we will not discuss the philosophical objections which might be raised when thinking that proper naming has to satisfy some special constraints.

The only constants introduced by Church are ternary val's. A val predicate applies to a variable, to a formula and to a property. Let us omit types and say that 'val(a, v, F)' means that a is a variable and v is a formula having no free variable other than a, and for every value x of the variable a the value of v is F(x). For simplicity, Church considers only the case that the variable a is for individuals and the variable F is for properties of individuals.

We will motivate the adoption of binary val's and we will remark that it suffices to define a notion of truth applying also to sentences in a way that is still Russellian. This modification can be presented and motivated independently of types. Even if we keep Church's notation referring to types in this section, the reader does not need to concern herself with them in this introductory section.

In Church's (1976) analysis of the Grelling paradox, the following definition of 'heterological' is provided:

$$het^{n+1}(v) =_{df} \exists a \exists F^{1/n} \, (val^{n+1}(a, v, F) \wedge \sim F(v))$$

and it is proved that

(6) $\forall x \, ((val^{m+2}(a, v, G^{1/m+1}) \wedge G(x)) \leftrightarrow het^{m+1}(x)) \rightarrow \sim het^{n+1}(v)$, if $m \geq n$.

In the proof, there is the following passage (the second 'hence') that is not correct:

> Hence, by univ. inst. and P [propositional logic], $val^{m+2}(a, v, G)$, $val^{n+1}(a, v, F), \sim F(v) \vdash \sim G(v)$
> Hence, by ex. inst., $val^{m+2}(a, v, G), het^{n+1}(v) \vdash \sim G(v)$

The variable a is bounded by an existential quantifier in $het^{n+1}(v)$, on the left side of $\vdash$, but is free in the hypothesis $val^{m+2}(a, v, G)$. Given the extensional equivalence of F and G, $val^2(a, v, G)$ and $val^1(a, v, F)$, where m = 0, the flaw in the passage is shown by an interpretation of val (val^1 or val^2) such that for a true closed formula v and a specific individual variable x

$$\text{val(a, v, F) is true if } \begin{cases} a = x \text{ and F is always true} \\ a \neq x \text{ and F is always false.} \end{cases}$$

An interpretation of this kind is no longer available after Church's (1984) correction of the proof of (6). Indeed, the correction amounts to the replacement of the postulate

(1) $(val^{m+1}(a, v, F^{1/m}) \wedge val^{n+1}(a, v, G^{1/n})) \rightarrow F = G$

from which it follows that

(2) $(val^{m+1}(a, v, F^{1/m}) \wedge val^{n+1}(a, v, G^{1/n})) \rightarrow \forall x\, (F(x) \leftrightarrow G(x))$

with the stronger postulate

$$(val^{m+1}(a, v, F^{1/m}) \wedge val^{n+1}(b, v, G^{1/n})) \rightarrow F = G$$

from which it follows that

$$(val^{m+1}(a, v, F^{1/m}) \wedge val^{n+1}(b, v, G^{1/n})) \rightarrow \forall x\, (F(x) \leftrightarrow G(x))$$

By adopting this postulate, it seems that the role of the first argument of val, which is meant to be an individual variable, becomes vacuous so that, instead

of Church's ternary val, a binary val can be used, which applies to a formula v with at most one free variable and a property F: val(v, F).

It is worth remarking that, as happens in the case of Church's ternary val, nothing prevents val(v, F) from being true when v is a closed formula. By virtue of Church's postulate schema (3), adapted to the binary val,

(3) $\exists v \exists F\ (val(v, F) \wedge \forall x\ (F(x) \leftrightarrow A))$

where A has at most the x variable as free variable and the predicate variable F is unary, we have that a sentence v expresses, relative to all individuals, a universal property if it is true and an empty property if it is false. A similar schema for a predicate val with a variable for binary relations as the second argument allows us to state that a true sentence expresses a universal binary relation and a false one an empty binary relation. Generalizing, by means of the appropriate val's and the corresponding schemata, one gets that a sentence expresses a property, a binary relation, a ternary relation, and so on, even if it is more natural to think that a sentence expresses a proposition—what can be said by a suitable postulate schema (3) for a predicate val that has a propositional variable as the second argument. No ambiguity ensues for the meaning of a sentence, since all these different semantic values are established by different meaning relations, expressed by different val predicates.

Of course, binary val's are not adequate to all purposes Church had in mind. We might need to refer to a the specific value of a specific variable, as, for example, to define Tarskian satisfaction relations. Informally, and omitting type distinctions, to define that formula v is satisfied by the values $x_1, x_2, ..., x_m$ of the variables $a_1, a_2, ..., a_m$, it should be said that there is a m-ary relation F such that $val(a_1, a_2, ..., a_m, v, F)$ and $F(x_1, x_2, ..., x_m)$ (see Church 1984, p. 301). Even for a broader characterization of the semantic values of formulas, independently of the comparison with other approaches to semantics, reference to variables is surely relevant, as it appears from hints given by Church in fn 20 (see Church 1984, p. 299).

In the next two sections we are going to introduce a very simple paradox, which only superficially resembles Russell's paradox (therefore called the pseudo-Russell) to show how, analogously to the case of the Grelling paradox, a solution can be provided by introducing ramified types. Our aim it to point out some aspects Church's approach to paradoxes, particularly the need to keep separate type distinctions from expressibility matters. A theory of types where the type of propositions is omitted and only binary val predicates for properties are introduced is sufficient for our analysis.

3. Formal Statement and Derivation of the Pseudo-Russell in a Simple Type Theory

In the work of Church (1976) and (1984), ramified types are assigned to entities and to corresponding variables and constants. In order to emphasize the role of ramification as a way of solving antinomies, let us begin by neglecting it and keeping only the simple type distinctions. i is the type of individuals; $(\beta_1, ..., \beta_n)$, where $n \geq 1$ is the type of n-ary relations having, as arguments, entities of types $\beta_1, ..., \beta_n$ respectively. As already mentioned, and motivated as unnecessary to express a notion of truth for sentences, we omit the type of propositions (and, by extension, the variables and constants for propositions in the language).[2]

A corresponding language should be such that every variable and constant is assigned a simple type. Let us take into account a language *L* that, besides satisfying this minimal requirement, is provided with an infinite choice of variables for every type. An infinite amount of individual constants of type i is available. As anticipated, we suppose that they are such that it is possible to systematically assign some countable of them the role of names of formulas of *L*. Such an assignment is supposed to be known and will be used in the statement of the axioms and postulates.

Terms are variable or constants. An atomic formula is a sequence of symbols $F(t_1, ..., t_n)$, where $n \geq 1$, F is a variable or a primitive constant of a type $(\beta_1, ..., \beta_n)$ and $t_1, ..., t_n$ are variables or constants of the types $\beta_1, ..., \beta_n$ respectively.

Non-atomic formulas are built in the standard way, by means of all the usual connectives and quantifiers, i.e., without restricting to negation, disjunction and universal quantifier as in the work of Church and Russell.

Symbols, terms and formulas of *L* are taken to have the type i of individuals, so they can be values of (individual) variables and referred to by (individual) constants.

Suitable binary predicate constants val, here called val predicates, of various types are included in *L* in order to make it possible to 'speak' of the semantic values of formulas with at most one free variable.

[2] As opposed to his earlier assumptions in Russell (1908), Russell (1910) took propositions as non-entities. Church was aware of this and defended the legitimacy, the coherence and, in a way, the need of propositions when propositional functions are endorsed. With reference to Church's considerations, Cocchiarella (1980) disagreed and remarked that it is technically possible to introduce the type hierarchy without the type of propositions.

In addition to strictly logical axioms and rules of inference appropriate for a typed language, the theory is provided with two kinds of axioms.

Comprehension axioms

1.1 $\exists F\, \forall x_1, ..., \forall x_n\, (F(x_1, ..., x_n) \leftrightarrow A)$

where $n \geq 1$, F is a relational variable of type $(\beta_1, ..., \beta_n)$, so that $x_1, ..., x_n$ are distinct variables of types $\beta_1, ..., \beta_n$ respectively, and F does not occur free in A.

Postulates for specific val *predicates*

val predicates are meant to express relations between a formula and a property.[3] Similar to the ternary val of Church (1984, p. 295), and omitting for simplicity type indications, a formula 'val(v, F)' is intended to mean that v is a formula having at most a free variable, and for every value x of the variable the value of v is F(x).

The postulates for the val predicates are, first of all, those expressing extensional univocacy (so-called by Church). Omitting for simplicity type indications, they have the following form:

2.1 $\forall v \forall F \forall G\, ((val(v, F) \wedge val(v, G)) \rightarrow \forall x\, (F(x) \leftrightarrow G(x)))$ *Univocacy*

Church's postulate schema (3) is replaced by two postulate schemata. While still omitting type indications, the instances of one postulate schema are as follows:

2.2 $\exists F\, val([A], F)$ *Existence of semantic value*

where A is a formula in which at most one free variable of some type β occurs, [A] is the individual constant assigned to A and F is a 1-ary relational variable, not occurring free in A, of the type (β). The instances of the other postulate schema are as follows:

2.3 $\forall F\, (val([A], F) \rightarrow \forall x\, (A \leftrightarrow F(x))$ *Semantic adequacy*

[3] F is taken by Church to be a propositional function, but we may take it as a property, leaving open the precise way in which it should be conceived.

where A is a formula in which at most one free variable of some type β occurs, [A] is the constant assigned to A, F is a 1-ary relational variable of type (β) not occurring free in A and 'x' is the variable occurring free in A, if any.

Then a relation for formulas with at most one free variable can be defined in order to represent the intuitive relation being-true-of of type (i, i):

$$T(v, x) =_{df} \exists P\,(val(v, P) \wedge P(x)).$$

T(v, x) is intended to be true when the formula v has a value—referred to by means of the 1-ary relational variable P—that is true of x.

By virtue of comprehension, T can be taken as a relation and not just as a tool for abbreviating the formula ∃P (val(v, P) ∧ P(x)). Then there is also a property R expressed by the formula ~T(x, x).[4] It is intuitively clear that R is paradoxical. Semi-formally, a proof can be given as follows:

Assume ~∃P (val(r, P) ∧ P(r)), where 'r' stands for [~∃P (val(x, P) ∧ P(x)]. Then ∀P (val(r, P) → ~P(r)). By 2.2 there is a R such that val(r, R). val(r, R) → ~R(r) follows from ∀P (val(r, P) → ~P(r)) by universal instantiation. Hence, by modus ponens, ~R(r). Keeping in mind that r is [~∃P (val(x, P) ∧ P(x)], by 2.3 we get val(r, R) → ∀x (~∃P (val(x, P) ∧ P(x)) ↔ R(x)). Thus, from val(r, R), by modus ponens, ∀x (~∃P (val(x, P) ∧ P(x)) ↔ R(x)). Hence ~∃P (val(r, P) ∧ P(r)) ↔ R(r). Since ~R(r), ∃P (val(r, P) ∧ P(r)).

Assume ∃P (val(r, P) ∧ P(r)), where r stands for [~∃P (val(x, P) ∧ P(x)]. Then there is a P′ such that val(r, P′) and P′(r). By 2.3 val(r, P′) → ∀x (~∃P (val(x, P) ∧ P(x)) ↔ P′(x)). Thus, by modus ponens, ∀x (~∃P (val(x, P) ∧ P(x)) ↔ P′(x)). Hence ~∃P (val(r, P) ∧ P(r)) ↔ P′(r). Since P′(r), ~∃P (val(r, P) ∧ P(r).

The paradox is apparently generated by the expression ~T(x, x), which looks like the formula that generates Russell's paradox. However, ~T(x, x) is a defined expression that stands for the formula ~∃P (val(x, P) ∧ P(x)), where differences of simple types are respected. Loosely speaking, the paradox shows the impossibility of taking val and ~∃P (val(x, P) ∧ P(x) on a par with P. It

[4] A comparison with the notion of heterologicality might be useful. Heterological is defined by Church as ∃P (val(x, P) ∧ ~P(x)) and turns out to be equivalent to ~T(x, x), i.e. ~∃P (val(x, P) ∧ P(x)), as a consequence of *Univocacy*, *Existence of semantic value*, and the intended meaning of val. However, ~T(x, x) looks (partially) analogous with Russell's paradox in a more direct way.

does not strictly depend on the interpretations of predicates and predicate variables as properties, even if the solution through ramification which we are going to consider introduces seemingly intensional distinctions.

Finally, it should be remarked that the derivation of $\sim\exists P\ (val(r, P) \wedge P(r))$ from $\exists P\ (val(r, P) \wedge P(r))$ does not contain any appeal to the *Existence of semantic value*. That might suggest that it is possible to avoid the paradox giving up only 2.2, without invalidating the second derivation and without diagnosing that the paradox depends on the simple type theory. However, ramification allows us to retain the *Existence of semantic value* (and *Semantic adequacy*) in forms such that the paradox is no longer derivable.

4. A solution through ramification

Once ramification is introduced, the above proof cannot be carried on any more. Let us define ramified types recursively, as Church does. From now on 'type' will stand for 'ramified type'.

i is a type and, if $\beta_1, ..., \beta_m$, where $m \geq 1$, are types, then $(\beta_1, ..., \beta_m)/n$, where $n \geq 1$, is a type. i is the type of individuals. $(\beta_1, ..., \beta_m)/n$ is the type of m-ary relations of level n having, as arguments, entities of types $\beta_1,..., \beta_m$ respectively.

As with Church, let us define $(\alpha_1, ..., \alpha_m)/k < (\beta_1, ..., \beta_m)/n$ if $\alpha_1 = \beta_1,..., \alpha_m = \beta_m$ and $k<n$. Entities of type α are intended to include entities of types $<\alpha$. Axioms should express cumulativity.

An order is assigned to every type in the following way: the order of the type i is 0, the order of a type $(\beta_1, ..., \beta_m)/n$ is $N+n$, where N is the greatest of the orders of the types $\beta_1, ..., \beta_m$.

Variables and constants of the language should be typed accordingly, and formulas should be constructed on the basis of the type distinctions. An atomic formula is a sequence of symbols $F(t_1, ..., t_m)$, where F is a variable or a primitive constant of a type $(\beta_1, ..., \beta_m)/n$, for some n and $m \geq 1$, and $t_1, ..., t_m$ are variables or constants of the types $\beta_1, ..., \beta_m$ respectively. Non-atomic formulas are built in the standard way. Let us keep the name '*L*' for the language so typed.

Let A be a well-formed formula. It is convenient to adopt the notion of the order of a variable or a constant identified, as usual, with the order of its type, and the notion of order of a formula: the order of a formula A, abbreviated by ord(A), is max(h, k+1), where h is the greatest order of free variables and constants occurring in A, and k is the greatest order of bound variables occurring in A.

Then the schema for comprehension axioms, here numbered as above, is modified in the following way:

1.1 $\exists F\, \forall x_1, ..., \forall x_m\, (F(x_1, ..., x_m) \leftrightarrow A)$

where F is a variable of type $(\beta_1, ..., \beta_m)/n$, $x_1, ..., x_m$ are distinct variables of types $\beta_1, ..., \beta_m$ respectively, $ord(A) \leq ord(F)$, and F does not occur free in A.

The postulates for val are to be suitably modified. Let us focus only on the second postulate schema for val.

The modification of 2.2, for formulas intuitively expressing properties of individuals, is as follows:

2.2 $\exists F\, val^{n+1}([A], F)$ *Existence of semantic value*

where val^{n+1} has type (i, (i)/n)/1, A is a formula with at most one free variable of the type i, [A] is the constant assigned to A, $ord(A) \leq n$ and F is a 1-ary variable of type (i)/n not occurring free in A.

To see how the above derivation of a paradox cannot be carried on any more, let us assume that 'x' has type i and 'P' and 'P′' have type (i)/1, so order 1, and are written as P^1 and as P'^1. val^{1+1} has order 2 and val^{2+1} has order 3. T(v, x), i.e. $\exists P^1\, (val^{1+1}(v, P^1) \wedge P^1(x))$, has order 2. T, as a property, has order 2, indeed ≥2, according to the ramified version of 2.2, and, taking it as having order 2, is written as T^2. Similarly R, which is determined by $\sim\exists P^1\, (val^{1+1}(x, P^1) \wedge P^1(x))$. Let us write R^2 for it. The outcome of such a specification of orders in the above derivation of the paradox is as follows, where the first half is reproduced only up to a clearly invalid inferential step:

Assume $\sim\exists P^1\, (val^{1+1}(r, P^1) \wedge P^1(r))$, where r stands for $[\sim\exists P^1\, (val^{1+1}(x, P^1) \wedge P^1(x))]$. Then $\forall P^1\, (val^{1+1}(r, P^1) \rightarrow \sim P^1(r))$. By 2.2 there is a R^2 such that $(val^{2+1}(r, R^2)$. $val^{1+1}(r, R^2) \rightarrow \sim R^2(r)$ follows from $\forall P^1\, (val^{1+1}(r, P^1) \rightarrow \sim P^1)(r))$ by universal instantiation. Hence, by modus ponens, $\sim R^2(r)$.

Assume $\exists P^1\, (val^{1+1}(r, P^1) \wedge P^1(r))$, where r stands for $[\sim\exists P^1\, (val^{1+1}(x, P^1) \wedge P^1(x))]$. Then there is a P'^1 such that $val^{1+1}(r, P^1)$ and $P'^1(r)$. By 2.3 $(val^{1+1}(r, P'^1) \rightarrow \forall x\, (\sim\exists P^1\, (val^{1+1}(x, P^1) \wedge P^1(x)) \leftrightarrow P'^1(x))$. Thus, by modus ponens, $\forall x\, (\sim\exists P^1\, (val^{1+1}(x, P^1) \wedge P^1(x)) \leftrightarrow P'^1(x))$. Hence $\sim\exists P^1\, (val^{1+1}(r, P^1) \wedge P^1(r)) \leftrightarrow P'^1(r)$. Since $P'^1(r)$, $\sim\exists P^1\, (val^{1+1}(r, P^1) \wedge P^1(r))$.

While the first part of the derivation is invalid, nothing violates the ramification

distinctions in the latter part of the derivation. This provides a proof for $\sim\exists P^1 (val^{1+1}(r, P^1) \wedge P^1(r))$, shortly $\sim T^2(r, r)$ or $R^2(r)$. In fact, no property of order 1 can be expressed by the formula $\sim\exists P^1 (val^{1+1}(x, P^1) \wedge P^1(x))$.

However, as is well known, reducibility axioms, whose merits or defects will not be discussed here, allow us to replace the commitment to high-level entities with a commitment to lower-level entities. In particular, according the appropriate reducibility axiom, there is a R^1 such that

$$\forall x(R^1(x) \leftrightarrow R^2(x)).$$

It follows that $\forall P^1 (val^{1+1}(r, P^1) \rightarrow \sim P^1(r))$ admits the following instance:

$$val^{1+1}(r, R^1) \rightarrow \sim R^1(r).$$

The antecedent $val^{1+1}(r, R^1)$ of this implication cannot be asserted or consistently assumed. If $val^{1+1}(r, R^1)$ is assumed, the paradox is derivable by replacing R^2 with R^1 and the derivation just reduces the assumption to absurdum. Thus $\sim val^{1+1}(r, R^1)$.

On the other hand, one might think that reducibility legitimizes the introduction of a constant TR^1 for a R^1 extensionally equivalent to R^2. Then the formula $TR^1(x)$ could be used to derive the paradox. However, it would be a natural reaction to conclude that such a constant for R^1 cannot belong to the language *L* and reducibility cannot legitimize its introduction in the language *L*.[5]

5. Removing Types and Connecting Principles for val and true-of

When types are removed, the postulates of the forms 2.1, 2.2 and 2.3 do not lose their interest. They can be appreciated because of the insights they provide into possible type-free treatments of semantic paradoxes, and more specifically because of their apparent relations with the Tarskian conditionals and the ways these conditionals might be restricted to avoid the paradoxes in a classical context.

[5] All this is more or less implicit in remarks first made by Church (1976) and Myhill (1979).

Let us state again 2.1, 2.2 and 2.3, but without any type distinctions except the distinction between individual and predicate variables:

2.1 $\forall v \forall F \forall G\ (val(v, F) \wedge val(v, G) \rightarrow \forall x\ (F(x) \leftrightarrow G(x))$ *Univocacy*
2.2 $\exists F\ val([A], F)$ *Existence of semantic value*

where A is a formula with at most one free variable, [A] is the constant assigned to A and F is a 1-ary predicate variable not occurring free in A.

2.3 $\forall F\ (val([A], F) \rightarrow \forall x\ (A \leftrightarrow F(x))$ *Semantic adequacy*

where A is a formula with at most one free variable, [A] is the constant assigned to A, F is a 1-ary predicate variable not occurring free in A and 'x' is the individual variable occurring free in A, if any.

With type distinctions omitted, the paradox here considered, like many other paradoxes of various kinds, is derivable and it is clear that semantic closeness, as expressed by 2.2 (*Existence of semantic value*) and 2.3 (*Semantic adequacy*) for the val predicates, is required to derive it.

However, 2.2 and 2.3 have very different roles. Giving up only 2.2 does not block the derivation of $\sim\exists P\ (val(r, P) \wedge P(r))$ from $\exists P\ (val(r, P) \wedge P(r))$, where r stands for $[\sim\exists P\ (val(x, P) \wedge P(x))]$, i.e., the proof of

(*) $\sim\exists P\ (val(r, P) \wedge P(r))$

or, equivalently,

$\forall P\ (val(r, P) \rightarrow \sim P(r))$.

If the postulates 2.2 are given up, and it is assumed that there is a R such that

(abs) $val(r, R) \wedge \forall x\ (R(x) \leftrightarrow \sim\exists P\ (val(x, P) \wedge P(x)))$

we get from $\forall P\ (val(r, P) \rightarrow \sim P(r))$

$val(r, R) \rightarrow \sim R(r)$

hence, from (abs)

$\sim R(r)$

so

$$\exists P\ (val(r, P) \wedge P(r))$$

against (*). Thus, by reduction

$$\sim\exists R\ (val(r, R) \wedge \forall x\ (R(x) \leftrightarrow \sim\exists P\ (val(x, P) \wedge P(x))).$$

All this can look very familiar because of the similarity with the Grelling paradox, which we will briefly take into account in the context of our reformulation of Church's theory.

As is well known, this paradox is obtained by considering the formula

$$\exists P\ (val(v, P) \wedge \sim P(v)).$$

Let us take 'het' as the constant for $\exists P\ (val(v, P) \wedge \sim P(v))$. It is easy to use 2.3 to derive $\sim\exists P\ (val(het, P) \wedge \sim P(het))$ from $\exists P\ (val(het, P) \wedge \sim P(het))$. On the other hand, suppose $\sim\exists P\ (val(het, P) \wedge \sim P(het))$. Then, for every P, if val(het, P), then P(het). By 2.2, for some P′, val(het, P′). Hence P′(het). By 2.3, $\forall x\ (\exists P\ (val(x, P) \wedge \sim P(x)) \leftrightarrow P'(x))$. Thus $\exists P\ (val(het, P) \wedge \sim P(het))$ by instantiation and propositional logic. As is the case above, if 2.3. is accepted and 2. 2 is given up, and it is assumed that there is a H such that

$$val(het, H) \wedge \forall x\ (H(x) \leftrightarrow \sim\exists P\ (val(x, P) \wedge \sim P(x)))$$

the derivation of a contradiction proves

$$\sim\exists H\ (val(het, H) \wedge \forall x\ (H(x) \leftrightarrow \sim\exists P\ (val(x, P) \wedge P(x))).$$

2.3 (*Semantic adequacy*) is resorted to in the proofs of both directions of the pseudo-Russell and of the Grelling paradox. It has a major role in expressing the characteristic features of the semantic relation val, whereas 2.2 (*Existence of semantic value*) just states a sort of universal applicability of val, even to formulas containing val, and hence implies a sort of semantic closure of the language which val belongs to.

The different roles of *Existence of semantic value* and *Semantic adequacy* are highlighted when they are connected with the Tarskian conditionals for the notion of true of. Let us take this notion as represented by T(v, w), according to the above adopted and here repeated definition:

$$T(v, w) =_{df} \exists F\ (val(v, F) \wedge F(w))$$

where F is a 1-ary relational variable. It is quite natural to assume that T fulfils the instances of the following schemata:

R $\quad T([A], x) \rightarrow A$
C $\quad A \rightarrow T([A], x)$

where [A] is a formula with at most one free variable and x is such a variable, if any. Recently, the corresponding principles for the truth predicate have been labeled respectively 'Release' and 'Capture'.[6] Here the letters 'R' and 'C' can be used in a similar sense.

It is very easy to derive R and C from 2.2 and 2.3:

Proof of R
Let us assume T([A], x), i.e., $\exists F\ (val([A], F) \wedge F(x))$. So, for some F′, $val([A], F') \wedge F'(x)$ and, by 2.3, $val([A], F') \rightarrow \forall x\ (A \leftrightarrow F'(x))$. Thus $A \leftrightarrow F'(x)$ and, since F′(x), then A.

Proof of C
Let us assume A. By 2.2, $\exists F\ val([A], F)$. So, for some F′, val([A], F′). By 2.3, $val([A], F') \rightarrow \forall x\ (A \leftrightarrow F'(x))$. Thus $A \leftrightarrow F'(x)$ follows. Since A by assumption, F′(x). Then $val([A], F') \wedge F'(x)$. So $\exists F\ (val([A], F) \wedge F(x))$, i.e., T([A], x).

It should be noted that 2.2 (*Existence of semantic value*) is utilized only in the proof of C, whereas 2.3 (*Semantic adequacy*) has a role in both the proof of R and the proof of C.

R and C together provide the Tarskian biconditionals. We should note that, in addition to R, there is also a restricted version of the Tarskian biconditionals that does not depend on 2.2. One such version is stated by Kripke for the closed off interpretation of the truth predicate T at the minimal fixed point of his semantic hierarchy (Kripke, 1975) and is informally introduced by Parsons (1974). Kripke's restricted version of the Tarskian biconditionals is stated for

[6] See, as an example, Beall and Glanzberg (2011). Indeed, Beall and Glanzberg (2011) do not introduce R and C with '→', but with '|–' taken as a place-holder for—they say—"a range of different logical notions, each of which will provide some notion of valid inference in some logical theory". Thus our R and C, stated above as valid classic implications, are specific principles instantiating their schemata.

an *undefined* truth predicate of sentences of a first order interpreted language as follows:

K $\quad (T([A]) \vee T([\sim A])) \rightarrow (A \leftrightarrow T([A]))$

val 2-nary predicates of type (i, n/0), where n/0 is a type of propositions, can be introduced in a Churchian language, so that T predicates, specific for sentences, can be defined in the following natural way:

$$T(v) =_{df} \exists p\ (val(v, p) \wedge p)$$

where v is an individual variable and p is a propositional variable of type n/0. If propositions are taken to be of the same type, only one T predicate specific for sentences is defined in the above way. However, we can dispense with propositions and propositional variables, and go on with our true-of predicate expressed by T(v, w) as previously defined and state K as:

K′ $\quad (T([A], x) \vee T([\sim A], x) \rightarrow (A \leftrightarrow T([A], x))$

K′ is easily derivable from 2.3.

Proof
Let us assume $T([A], x) \vee T([\sim A], x)$, i.e., $\exists F\ (val([A], F) \wedge F(x)) \vee \exists F\ (val([\sim A], F) \wedge F(x))$. Let us prove that each disjunct entails $A(x) \leftrightarrow T([A], x)$, i.e., $A \leftrightarrow \exists F\ (val([A], F) \wedge F(x))$:

1. $\exists F\ (val([A], F) \wedge F(x))$ (hyp.)
   By propositional logic, $A \rightarrow \exists F\ (val([A], F) \wedge F(x))$. On the other hand, by hyp., for some F′, val([A], F′) and F′(x). By 2.3 $A \leftrightarrow F'(x)$. So A. Thus $\exists F\ (val([A], F) \wedge F(x)) \rightarrow A$.
2. $\exists F\ (val([\sim A], F) \wedge F(x))$ (hyp.)
   For some F′, val([~A], F′) and F′(x). By 2.3 $\sim A \leftrightarrow F'(x)$; thus ~A. By propositional logic, $A \rightarrow \exists F\ (val([A], F) \wedge F(x))$. Concerning the other direction, assume $\exists F\ (val([A], F) \wedge F(x))$. Then for some F′, val([A], F′) and F′(x). By 2.3 $A \leftrightarrow F'(x)$; thus A, against the previous derivation of ~A. Thus $\sim\exists F\ (val([A], F) \wedge F(x))$. By propositional logic, $\exists F\ (val([A], F) \wedge F(x)) \rightarrow A$.

Vice versa, is 2.3 (*Semantic adequacy*) derivable from K′? It is, if the following principle is adopted:

NEG $\quad \mathrm{val}([A], F) \rightarrow \exists G\,(\mathrm{val}([\sim A], G) \wedge \forall x\,(G(x) \leftrightarrow \sim F(x)))$.

Loosely speaking, NEG says that if A expresses F, then ~A expresses ~F. It allows the following derivation of 2.3 from K′.

Proof
Let us assume val([A], F*) (hyp.). It will be proved that $A \leftrightarrow F^*(x)$, where x is the only variable occurring in A, if any.

1. Suppose $F^*(x)$. Then, by hyp., $\exists F\,(\mathrm{val}([A], F) \wedge F(x))$, i.e., T([A], x). By K′, A. Thus $F^*(x) \rightarrow A$.
2. Suppose A. Let us take into account the cases T([A], x), T([~A], x), $\sim T([A], x) \wedge \sim T([\sim A], x)$.

2a. T([A], x), i.e. $\exists F\,(\mathrm{val}([A], F) \wedge F(x))$. So, for some F′, val([A], F′) and F′(x). By hyp. and 2.1, $F^*(x) \leftrightarrow F'(x)$, hence F*(x).

2b. T([~A], x). By K′ and supposition 2, T([A], x). So, for some F′, val([A], F′) and F′(x). By 2.1, $F^*(x) \leftrightarrow F'(x)$, hence F*(x).

2c. $\sim T([A], x) \wedge \sim T([\sim A], x)$. So (i) $\forall F\,(\mathrm{val}([A], F) \rightarrow \sim F(x))$ and (ii) $\forall F\,(\mathrm{val}([\sim A], F) \rightarrow \sim F(x))$. By hyp. and (i), $\sim F^*(x)$. By hyp. and NEG, for some G, val([~A], G) and $G(x) \leftrightarrow \sim F^*(x)$. By (ii), ~G(x). Hence, F*(x). Thus, by reductio, hyp. is false.

It follows that

$\mathrm{val}([A], F) \rightarrow (F(x) \rightarrow A)$
$\mathrm{val}([A], F) \rightarrow (A \rightarrow F(x))$

whence 2.3.

Thus, under the quite natural assumption NEG, K′ is equivalent to *Semantic adequacy*. Since *Semantic adequacy* entails R, it follows that Kripke (1975) is committed to R.[7]

As well known, Kripke's attitude towards R and C is not Tarskian. The difference can be clearly accounted for within Church's framework.

Let us suppose that A has no free variable, i.e., A is a sentence and x is an individual variable. Then

R $\quad T([A], x) \rightarrow A$

says that if A expresses a property true of an individual, then A (or A is intuitively true), and

[7] The endorsement of R by Kripke (1975) was first, in a different way, shown by Feferman (1984, pp. 101-102).

C $A \rightarrow T([A], x)$

says that if A (or A is intuitively true), A expresses a property true of an individual.

Surely this is not the meaning assigned to Tarskian conditionals either by Tarski or Kripke. However, R and C are together contradictory, like the intuitive conditionals taken into account by Tarski and Kripke. Their derivation from the more basic Churchian principles 2.2. and 2.3 suggest the following account of the different ways in which Tarski and Kripke pursue the goal of avoiding the semantic paradoxes.

Their different approaches may be traced to different imaginary choices concerning the validity of the val principles. Assuming that Kripke's final move is the closing off, our fiction is as follows. Both Tarski and Kripke endorse 2.1 (*Univocacy*) without any limitation. Both require 2.3 (*Semantic adequacy*) to hold. However, Tarski also requires that 2.2 (*Existence of semantic value*) be generally valid, in order for his biconditionals for truth to hold. So, to avoid the contradiction, he has to restrict the range of formulas upon which the val relation, and therefore truth, is defined. Kripke appears to give up the general validity of 2.2 (*Existence of semantic value*), while allowing—consistently with the closing off—that the val relation is defined on formulas having no semantic value. It follows that the antecedent of 2.3 (*Semantic adequacy*) is false for some formulas and not all Tarskian biconditionals can be asserted.

6. A Final Look: How to Move in the Direction of Field's Recent Approach

The predicate T of a Liar sentence is usually undefined. Kripke's Liar form is $\forall x\,(P(x) \rightarrow \sim T(x))$, where P(x) is a syntactic predicate uniquely satisfied by the code of $\forall x\,(P(x) \rightarrow \sim T(x))$. A shorter standard form is $\sim T(l)$, where l is a term whose value is the formula $\sim T(l)$. When T(x) is understood as saying that x expresses a property true of any individual, Kripke's semantic construction can be conceived as a way of systematically identifying the sentences expressing a universal property. However, the specific kind of values that may be assigned to sentences is not so relevant. The outcome matters: an interpretation of T is reached at the minimal fixed point such that some sentences do not come out as true and their negations either. A Liar sentence is such a sentence. By means of the closing off it can be acknowledged as intuitively true but is not true according to the fixed point interpretation of T. Thus, as em-

phasized by Field, both the Liar and the negation that the Liar is true should be asserted.

This awkward result essentially depends on the non-validity of bivalence —which, using the Churchian true-of predicate, is expressed by means of the schema $T([A], x) \vee T([\sim A], x)$—and on the final preservation of classical logic. In our Churchian framework bivalence follows from *Existence of semantic value*, *Semantic adequacy*, NEG and Excluded Middle. For, by *Existence of semantic value*, both $A \rightarrow \exists F\ val([A], F)$ and $\sim A \rightarrow \exists F\ val([\sim A], F)$. Then, given A, by *Semantic adequacy*, $\exists F\ (val([A], F) \wedge F(x))$, and, given ~A, by *Semantic adequacy* and NEG, $\exists F\ (val([\sim A], F) \wedge \sim F(x))$. By Excluded Middle, $\exists F\ (val([A], F) \wedge F(x)) \vee \exists F\ (val([\sim A], F) \wedge \sim F(x))$, i.e. $T([A], x) \vee T([\sim A], x)$. From the outlined Churchian perspective, Kripke's approach amounts to giving up *Existence of semantic value*, in agreement with his informal introductory remarks arguing for the compatibility of linguistic legitimacy and lack of semantic value.

However, another possible move is giving up Excluded Middle. Then acknowledging the lack of designated value is not a sufficient reason for asserting the negation and a correlate move is to introduce a conditional allowing the restatement of Tarskian biconditionals in such a way that they come out true. This is the perspective that Field develops in great detail. Of course, the abandonment of classical logic and—we should add—the rejection of ontological commitment to semantic values locate it completely outside a Churchian framework.[8]

REFERENCES

BEALL, J.C. and GLANZBERG, M. (2011). Liar Paradox. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, http://plato.stanford.edu/archives/spr2011/entries/liar-paradox/.

CHURCH, A. (1976). Comparison of Russell's Resolution of the Semantical Antinomies with that of Tarski. *The Journal of Symbolic Logic*, **41**, pp. 747-760.

CHURCH, A. (1984). Comparison of Russell's Resolution of the Semantical Antinomies with that of Tarski. In R. L. Martin (Ed.), *Recent Essays on Truth and the Liar Paradox*. Oxford: Oxford University Press, pp. 289-306.

COCCHIARELLA, N. (1980). The Development of the Theory of Logical Types and

[8] Comments of various people have been useful for the final version of this article. I especially thank Matteo Plebani and Jiří Raclavský.

the Notion of Logical Subject in Russell's Early Philosophy. *Synthese*, **45**, pp. 71-115.

FEFERMAN, S. (1984). Toward useful type-free theories. I. *The Journal of Symbolic Logic*, **49**, pp. 75-111.

FIELD, H. (2008). *Saving Truth From Paradox*. Oxford: Oxford University Press.

KRIPKE, S. (1975). Outline of a Theory of Truth. *Journal of Philosophy*, **72**, pp. 690-716.

MYHILL, J. (1979). A Refutation of an Unjustified Attack on the Axiom of Reducibility. In G. W. Roberts (Ed.), *Bertrand Russell Memorial Volume*. London: Allen &Unwin, pp. 81-90.

PARSONS, C. (1974). The Liar Paradox, *Journal of Philosophical Logic*, **3**, pp. 381-412.

RUSSELL, B. (1908). Mathematical Logic as Based on the Theory of Types. *American Journal of Mathematics*, **30**; reprinted in R. C. Marsh (Ed.), *Logic and Knowledge*. London: Allen & Unwin, 1956, pp. 59-102.

TARSKI, A. (1935). The Concept of Truth in Formalized Languages. In A. Tarski, *Logic, Semantics, Metamathematics*. Oxford: Clarendon Press. 2nd ed. Indianapolis: Hackett Publishing Company, 1983, pp.152-278.

WHITEHEAD, A N. and RUSSELL, B. (1910-1913). *Principia Mathematica*. Cambridge: Cambridge University Press.

L&PS – Logic & Philosophy of Science
Vol. XII, No. 1, 2014, pp. 91-101

A Note on Confirmation and Matthew Properties

William Roche
Department of Philosophy
Texas Christian University
e-mail: w.roche@tcu.edu

ABSTRACT. There are numerous (Bayesian) confirmation measures in the literature. Festa provides a formal characterization of a certain class of such measures. He calls the members of this class "incremental measures". Festa then introduces six rather interesting properties called "Matthew properties" and puts forward two theses, hereafter "T1" and "T2", concerning which of the various extant incremental measures have which of the various Matthew properties. Festa's discussion is potentially helpful with the problem of measure sensitivity. I argue, that, while Festa's discussion is illuminating on the whole and worthy of careful study, T1 and T2 are strictly speaking incorrect (though on the right track) and should be rejected in favor of two similar but distinct theses.

KEYWORDS: Confirmation, Festa, Matthew Properties, Problem of Measure Sensitivity.

1. Introduction

There are numerous (Bayesian) confirmation measures in the literature. Festa (2012) provides a formal characterization of a certain class of such measures.[1]

[1] All references to Festa are to Festa (2012).

Published by L&PS – Logic and Philosophy of Science
http://www2.units.it/~episteme – ISSN: 1826-1043

He calls the members of that class "incremental measures".[2] Each of the following is an incremental measure:[3]

$$c_d(H, E) = p(H|E) - p(H)$$

$$c_{lr}(H, E) = \frac{p(E|H)}{p(E|\neg H)}$$

$$c_r(H, E) = \frac{p(H|E)}{p(H)}.$$

Festa then introduces six rather interesting properties called "Matthew properties" and puts forward two theses, hereafter "T1" and "T2", concerning which of the various extant incremental measures have which of the various Matthew properties.

No two of the three measures c_d, c_{lr}, and c_r are ordinally equivalent to each other (i.e., impose the same ordering on any two ordered pairs of proposi-

[2] Below, in Section 2, I explain Festa's formal characterization of the class of incremental measures. It is worth noting now, though, that *incremental* measures are more than just *relevance* measures, where a measure c is a relevance measure just in case there is a neutral point n such that $c(H,E) > / = / < n$ iff $p(H|E) > / = / < p(H)$. (This characterization of relevance measures is adapted from Fitelson 1999.) Consider the following (well known) measures:

$c_C(H,E) = p(H \wedge E) - p(H)p(E)$

$c_M(H,E) = p(E|H) - p(E)$

$c_N(H,E) = p(E|H) - p(E|\neg H)$

$c_S(H,E) = p(H|E) - p(H|\neg E)$.

Each of these measures is a relevance measure as defined above. But none of them is an incremental measure as characterized by Festa. This is just as it should be, it seems, if an incremental measure is understood as a measure of the amount of increase in H's probability due to E, for on each of c_C, c_M, c_N, and c_s there can be cases where $p(H_1|E_1) > p(H_2|E_2)$ while $p(H_1) < p(H_2)$ and yet the degree to which E_1 confirms H_1 is less than the degree to which E_2 confirms H_2. This allows that there are conceptions of confirmation distinct from the incremental conception (where confirmation is a matter of the amount of increase in H's probability due to E) and in terms of which c_C, c_M, c_N, and c_s are best understood. See Hajek and Joyce (2008) and Joyce (1999, Ch. 6, sec. 6.4) for relevant discussion.

[3] The subscripts in these measures, along with the subscripts in the measures set out below in Section 2, are identical to the subscripts used by Festa.

tions).[4] This is prima facie problematic in that each of the three measures has some intuitive plausibility and yet certain results in confirmation theory involving one of the measures do not carry over to (at least one of) the other two measures. This is the problem of measure sensitivity.[5]

Festa's discussion is potentially helpful with this problem. Suppose one of the various Matthew properties is compelling in that any adequate incremental measure should have that property. Suppose it follows from T1 and T2 that, say, c_d has the property in question but neither c_{lr} nor c_r does. Then c_{lr} and c_r should be rejected as inadequate (as incremental measures). This would serve to narrow down the field of potentially adequate incremental measures and thus constitute progress towards solving the problem of measure sensitivity.

It turns out, however, that, while Festa's discussion is illuminating on the whole and worthy of careful study, T1 and T2 are strictly speaking incorrect (though on the right track). In Section 2, I set out the various incremental measures under consideration along with T1 and T2. In Section 3, I argue that T1 and T2 should be rejected in favor of two similar but distinct theses. In Section 4, I conclude.

2. Festa's Two Theses

Festa characterizes the class of incremental measures in terms of the following properties (or conditions):[6]

Initial and Final Probability Dependence (IFPD): $c(H,E)$ is a function of $p(H|E)$ and $p(H)$.

Final Probability Incrementality (FPI): Suppose $p(H_1) = p(H_2)$. Then $c(H_1,E_1) > / < c(H_2,E_2)$ if and only if $p(H_1|E_1) > / < p(H_2|E_2)$.

[4] Measures c and c^* are ordinally equivalent to each other just in case, for any ordered pairs of propositions $\langle H, E\rangle$ and $\langle H', E'\rangle$, the following holds: $c(H, E) > / = / < c(H', E')$ iff $c^*(H, E) > / = / < c^*(H', E')$.

[5] See Brössel (2013) and Fitelson (1999) for helpful discussion of the problem of measure sensitivity.

[6] It should be understood throughout the discussion that the propositions involved in the various probabilities are "p-normal" in that they have nonextreme unconditional probabilities (i.e., unconditional probabilities less than one and greater than zero).

Initial Probability Incrementality (IPI): Suppose $0 < p(H_1|E_1) = p(H_2|E_2) < 1$. Then $c(H_1,E_1) > / < c(H_2,E_2)$ if and only if $p(H_1) < / > p(H_2)$. Suppose $p(H_1|E_1) = p(H_2E_2) = 0$ or $p(H_1|E_1) = p(H_2|E_2) = 1$. Then (a) $c(H_1,E_1) \geq c(H_2,E_2)$ if $p(H_1) < p(H_2)$ and (b) $c(H_1,E_1) \leq c(H_2,E_2)$ if $p(H_1) > p(H_2)$.

Equineutrality (E): Suppose $p(H_1|E_1) = p(H_1)$ and $p(H_2|E_2) = p(H_2)$. Then $c(H_1,E_1) = c(H_2,E_2)$.

The class of incremental measures is defined as the class of measures having each of IFPD, FPI, IPI, and E.

It turns out that many extant confirmation measures are members of the class of incremental measures. Festa considers, in addition to c_d, c_{lr}, and c_r, the following:

$$c_{r^*}(H,E) = \frac{p(H|E) - p(H)}{p(H|E) + p(H)}$$

$$c_{or}(H,E) = \frac{o(H|E)}{o(H)} \text{ where } o(H|E) = \frac{p(H|E)}{p(\neg H|E)} \text{ and } o(H) = \frac{p(H)}{p(\neg H)}$$

$$c_G(H,E) = \frac{p(H|E) - p(H)}{1 - p(H)}$$

$$c_z = \begin{cases} \dfrac{p(H|E) - p(H)}{1 - p(H)} \text{ if } p(H|E) \geq p(H) \\ \dfrac{p(H|E) - p(H)}{p(H)} \text{ if } p(H|E) < p(H) \end{cases}$$

$$c_{So}(H,E) = \frac{\log[p(H|E)/p(H)]}{-\log[p(H)]}$$

$$c_{Pl}(H,E) = \frac{p(H|E) - p(H)}{p(H|E) + p(H) - p(H|E)p(H)}$$

$$c_{hP}(H,E) = \frac{p(H|E) - p(H)}{p(H|E) + p(H) + p(H|E)p(H)}$$

$$c_{\pi}(H,E) = \frac{p(H|E) - p(H)}{p(H|E) + p(H) + \pi p(H|E)p(H)} \text{ where } -2 \leq \pi \leq \infty$$

$$c_{\alpha}(H,E) = \frac{p(H|E) + \alpha p(H)p(H|E)}{p(H) + \alpha p(H)p(H|E)} \text{ where } -1 \leq \alpha \leq \infty$$

$$c_{db}(H,E) = \frac{p(H|E) - p(H)}{p(H|E)p(H)}$$

$$c_{KO}(H,E) = \frac{p(E|H) - p(E|\neg H)}{p(E|H) + p(E|\neg H)}$$

$$c_{P}(H,E) = \frac{p(E|H) - p(E)}{p(E|H) + p(E) - p(H \wedge E)}$$

$$c_{Ku}(H,E) = \frac{p(E|H)}{p(E)}$$

Some of the sixteen measures under consideration are ordinally equivalent to each other: c_r is ordinally equivalent to each of c_{Ku} and c_{r*}; c_{lr} is ordinally equivalent to each of c_{or} and c_{KO}; c_P is ordinally equivalent to c_{Pl}.[7] This is significant in that if one measure is ordinally equivalent to another, then the one has a given Matthew property just in case the other too has that property. I thus want to set aside c_{Ku}, c_{r*}, c_{or}, c_{KO}, and c_{Pl} and focus on c_r, c_{lr}, and c_P along with the remaining eight measures.[8]

Take some incremental measure c. Since c has IFPD, it follows that $c(H, E)$ is a function of $p(H|E)$ and $p(H)$. But:

$$p(H,E) = \frac{p(E|H)}{p(E)}\, p(H).$$

So $c(H,E)$ is a function of $Q(H,E) = p(E|H)/p(E)$ and $p(H)$, where, following Festa, $Q(H,E)$ is H's predictive success with respect to E.

[7] It is straightforward to verify that $c_r(H,E) = c_{Ku}(H,E)$, $c_{r*}(H,E) = [c_r(H,E) - 1] / [c_r(H,E) + 1]$ where $[n - 1] / [n + 1]$ is an increasing function of n for $n \geq 0$, $c_{lr}(H,E) = c_{or}(H,E)$, $c_{KO}(H,E) = [c_{lr}(H,E) - 1] / [c_{lr}(H,E) + 1]$ where, again, $[n - 1] / [n + 1]$ is an increasing function of n for $n \geq 0$, and $c_P(H,E) = c_{Pl}(H,E)$.

[8] There is thus no mention of c_{Ku}, c_{r*}, c_{or}, c_{KO}, and c_{Pl} in T1 and T2 as formulated below.

It follows, as Festa notes, that each of the incremental measures under consideration can be restated in terms of $Q(H,E)$, hereafter "Q", and $p(H)$. $c_d(H,E)$, for example, can be restated as $p(H)[Q - 1]$. This can be seen as follows:

$$\begin{aligned} c_d(H,E) &= p(H|E) - p(H) \\ &= \frac{p(E|H)}{p(E)}\, p(H) - p(H) \\ &= p(H)[Q - 1] \end{aligned}$$

Festa provides a "Q-function" for each of the incremental measures under consideration.

I can now state the six Matthew properties introduced by Festa. They can be put as follows:

> *Matthew Independence for Positive Confirmation (MIP)*: For any $Q > 1$, if Q is held fixed, then $c(H,E)$ is held fixed and thus is independent of $p(H)$.
>
> *Matthew Effect for Positive Confirmation (MEP)*: For any $Q > 1$, if Q is held fixed, then $c(H,E)$ is an increasing function of $p(H)$.
>
> *Reverse Matthew Effect for Positive Confirmation (RMP)*: For any $Q > 1$, if Q is held fixed, then $c(H,E)$ is a decreasing function of $p(H)$.
>
> *Matthew Independence for Disconfirmation (MID)*: For any $Q < 1$, if Q is held fixed, then $c(H,E)$ is held fixed and thus is independent of $p(H)$.
>
> *Matthew Effect for Disconfirmation (MED)*: For any $Q < 1$, if Q is held fixed, then $c(H,E)$ is a decreasing function of $p(H)$.
>
> *Reverse Matthew Effect for Disconfirmation (RMD)*: For any $Q < 1$, if Q is held fixed, then $c(H,E)$ is an increasing function of $p(H)$.

Recall that (following Festa) Q is H's predictive success with respect to E. MIP can be glossed: for any degree of predictive success greater than 1, $c(H,E)$ is independent of H's prior probability. MEP, in turn, can be glossed: for any degree of predictive success greater than 1, the greater is H's prior probability, the greater is $c(H,E)$. And so on for RMP, MID, MED, and RMD.

Why are the six Matthew properties named "Matthew" properties? Festa (referencing Kuipers 2000) writes:

> Kuipers ... introduces the concept of *Matthew effect for confirmation* just w.r.t. [MEP restricted to cases where H logically implies E]. In fact, [MEP restricted to cases where H logically implies E] "may be seen as a methodological version of the so-called Matthew effect, according to which the rich profit more than the poor" ..., in agreement with the sentence—made famous by the Gospel according to St. Matthew—that "unto every one that hath shall be given". (p. 95, emphasis original)

MEP implies that if two hypotheses have the same predictive success (greater than 1) with respect to some piece of evidence, and if initially the two hypotheses had different probabilities, then the hypothesis that initially had the higher probability (the "richer" of the two hypotheses initially) is more strongly confirmed by ("profits" more from) the evidence than does the hypothesis that initially had the lower probability (the "poorer" of the two hypotheses initially). Thus the name "Matthew Effect for Positive Confirmation" and, for consistency, the names of the remaining five properties.

It is clear that MIP, MEP, and RMP are pairwise mutually inconsistent in that any measure having one of them lacks each of the other two. It is also clear that the same is true with respect to MID, MED, and RMD. But which measures have which properties?

T1 and T2 are meant to answer this question. They can be put like this:

T1
- a c_r has MIP and MID.
- b c_d, c_G, c_{So}, c_{lr}, and c_P have MEP and MED.
- c c_z has MEP and MID.

T2
- a c_{hP} and c_{db} have RMP and RMD.
- b c_π has MEP and MED when $\pi < 0$, has MIP and MID when $\pi = 0$, and has RMP and RMD when $\pi > 0$.
- c c_α has MEP and MED when $\alpha < 0$, has MIP and MID when $\alpha = 0$, and has RMP and RMD when $\alpha > 0$.

T1 and T2, I take it, are meant to follow straightforwardly from the various Q-functions provided by Festa. Recall that the Q-function for $c_d(H, E)$ is $p(H)[Q - 1]$. If $Q > 1$ and Q is held fixed, it follows that $p(H)[Q - 1]$ is an increasing function of $p(H)$. If $Q < 1$ and Q is held fixed, it follows that $p(H)[Q - 1]$ is a decreasing function of $p(H)$. So, just as T1b implies, c_d has MEP and MED.

It turns out, however, that not all is right with T1 and T2. Some modifications are needed.

3. Two Replacement Theses

Suppose E entails $\neg H$ so that $p(H|E) = 0 = p(E|H)$. Then $p(H|E)/p(H) = 0$ regardless of $p(H)$. But $Q = p(E|H)/p(E) = p(H|E)/p(H)$. So $Q = 0$ regardless of $p(H)$. Suppose c is an incremental measure such that $c(H,E)$ takes the minimum value (for c) in any case where E entails $\neg H$. Then it is not true that for any $Q < 1$, if Q is held fixed, then $c(H,E)$ is a decreasing function of $p(H)$, and it is not true that for any $Q < 1$, if Q is held fixed, then $c(H,E)$ is an increasing function of $p(H)$. So c has neither MED nor RMD.

This spells trouble for T1 and T2. Suppose E entails $\neg H$. Then it follows that:[9]

$$c_{So}(H,E) = \frac{\log[p(H|E)\ /p(H)]}{-\log[p(H)]}$$

$$= \frac{-\infty}{-\log[p(H)]}$$

$$= -\infty$$

$$c_{lr}(H,E) = \frac{p(E|H)}{p(E|\neg H)}$$

$$= \frac{0}{p(E|\neg H)}$$

$$= 0$$

$$c_{P}(H,E) = \frac{p(E|H) - p(E)}{p(E|H) + p(E) - p(H \wedge E)}$$

$$= \frac{0 - p(E)}{0 + p(E) - 0}$$

$$= -1$$

[9] c_{So} can be understood as having the range $(-\infty, 1]$. See Shogenji (2012, p. 37) and Atkinson (2012, p. 53). But then, as the only plausible candidate value for $c_{So}(H,E)$ to take when $p(H|E) = 0$ is $-\infty$, it follows that $c_{So}(H,E)$ is undefined when $p(H|E) = 0$. This is less than ideal, it seems, since there should be a degree of confirmation even when $p(H|E) = 0$. It seems preferable to understand

$$c_{hP}(H,E) = \frac{p(H|E) - p(H)}{p(H|E) + p(H) + p(H|E)p(H)}$$
$$= \frac{0 - p(H)}{0 + p(H) + 0}$$
$$= -1$$

$$c_{\pi}(H,E) = \frac{p(H|E) - p(H)}{p(H|E) + p(H) + \pi p(H|E)p(H)}$$
$$= \frac{0 - p(H)}{0 + p(H) + 0}$$
$$= -1$$

$$c_{\alpha}(H,E) = \frac{p(H|E) + \alpha p(H)p(H|E)}{p(H) + \alpha p(H)p(H|E)}$$
$$= \frac{0 + 0}{p(H) + 0}$$
$$= 0$$

Note that $c_{\pi}(H,E) = -1$ and $c_{\alpha}(H,E) = 0$ regardless of the values specified for π and α respectively. It follows that c_{So}, c_{lr}, and c_P do not have MED, that c_{hP} does not have RMD, that c_{π} does not have MED when $\pi < 0$ and does not have RMD when $\pi > 0$, and that c_{α} does not have MED when $\alpha < 0$ and does not have RMD when $\alpha > 0$. So T1b is incorrect and each of T2a, T2b, and T2c is incorrect. So T1 and T2 are incorrect.

T1 and T2, though, are on the right track. They can be replaced by the following:

T1* a c_r has MIP and MID.
b c_d and c_G have MEP and MED.
c c_{So}, c_{lr}, and c_P have MEP but do not have MID, MED, or RMD; c_{So}, c_{lr}, and c_P have MED in the special case where $1 > Q > 0$.
d c_z has MEP and MID.

c_{So} as having the range $[-\infty, 1]$ and as taking the value $-\infty$ when $p(H|E) = 0$. Atkinson and Shogenji (personal communication) agree that c_{So} should be understood as having the range $[-\infty, 1]$.

T2* a c_{db} has RMP and RMD.

b c_{hP} has RMP but does not have MID, MED, or RMD; c_{hP} has RMD in the special case where $1 > Q > 0$.

c c_{π} has MEP but does not have MID, MED, or RMD when $\pi < 0$; c_{π} has MIP and MID when $\pi = 0$; c_{π} has RMP but does not have MID, MED, or RMD when $\pi > 0$; c_{π} has MED when $\pi < 0$ in the special case where $1 > Q > 0$; c_{π} has RMD when $\pi > 0$ in the special case where $1 > Q > 0$.

d c_{α} has MEP but does not have MID, MED, or RMD when $\alpha < 0$; c_{α} has MIP and MID when $\alpha = 0$; c_{α} has RMP but does not have MID, MED, or RMD when $\alpha > 0$; c_{α} has MED when $\alpha < 0$ in the special case where $1 > Q > 0$; c_{α} has RMD when $\alpha > 0$ in the special case where $1 > Q > 0$.

T1* and T2* differ from T1 and T2 only with respect to cases where E entails $\neg H$ and thus $Q = 0$.

Some of the measures referred to in T1 and T2 have a maximum value and take that value in any case where E entails H. Consider c_{So} for example. If E entails H so that $p(H|E) = 1$, it follows that $c_{So}(H,E)$ takes its maximum value of 1 regardless of H's prior probability. Why is it that T1 and T2 run into trouble in the case where E entails $\neg H$ but do not run into trouble in the case where E entails H?

Return to the case where E entails $\neg H$. The key here is that Q equals 0 regardless of H's prior probability. This means that Q can be held fixed while $p(H)$ increases or decreases. This in turn means that if $c(H,E)$ takes the minimum value (for c) in any case where E entails $\neg H$, then there can be cases where E entails $\neg H$, Q is held fixed, and $c(H,E)$ remains constant at the minimum value while $p(H)$ decreases, in which case c does not have MED, and there can be cases where E entails $\neg H$, Q is held fixed, and $c(H,E)$ remains constant at the minimum value while $p(H)$ increases, in which case c does not have RMD. Things are different in the case where E entails H. Suppose c is an incremental measure such that $c(H,E)$ takes the maximum value (for c) in any case where E entails H. Suppose E entails H so that $p(H|E) = 1$. Then $1/p(H) = p(H|E)/p(H) = p(E|H)/p(E) = Q$. But then if Q is held fixed, it follows that $p(H)$ too is held fixed. Hence there can be no cases where E entails H, Q is held fixed, and $c(H, E)$ remains constant at the maximum value while $p(H)$ increases or decreases. So no case where E entails H could show that c does not have MEP or that c does not have RMP.

4. Conclusion

T1 and T2 are incorrect in some of what they imply with respect to cases where E entails $\neg H$ and thus $Q = 0$. They should be rejected in favor of T1* and T2*. The way is now clear for confirmation theorists to focus on which, if any, of the various Matthew properties are compelling.[10] By doing so confirmation theorists can perhaps use T1* and T2* to narrow down the field of potentially adequate incremental measures and so make progress towards solving the problem of measure sensitivity.

Acknowledgments. Thanks to an anonymous reviewer for helpful comments on an earlier version of the paper.

REFERENCES

ATKINSON, D. (2012). Confirmation and Justification: A Commentary on Shogenji's Measure. *Synthese*, **184**, pp. 49-61.

BRÖSSEL, P. (2013). The Problem of Measure Sensitivity Redux. *Philosophy of Science*, **80**, pp. 378-397.

FESTA, R. (2012). "*For unto every one that hath shall be given*". Matthew Properties for Incremental Confirmation. *Synthese*, **184**, pp. 89-100.

FITELSON, B. (1999). The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity. *Philosophy of Science*, **66**, pp. S362-S378.

HAJEK, A. and JOYCE, J. (2008). Confirmation. In S. Psillos and M. Curd (Eds.), *The Routledge Companion to Philosophy of Science*. London: Routledge, pp. 115-128.

JOYCE, J. (1999). *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

KUIPERS, T. (2000). *From Instrumentalism to Constructive Realism: On Some Relations between Confirmation, Empirical Progress, and Truth Approximation*. Dordrecht: Kluwer.

SHOGENJI, T. (2012). The Degree of Epistemic Justification and the Conjunction Fallacy. *Synthese*, **184**, pp. 29-48.

[10] Festa (sec. 3.3.2) suggests that at least in some cases RMP is compelling.

L&PS – Logic & Philosophy of Science
Information on the Journal

Aims and contents

L&PS – Logic and Philosophy of Science is an on-line philosophical journal sponsored by the Department of Humanistic Studies of the University of Trieste (Italy). The journal promotes both theoretical and historical research in the philosophy of science and logic, without excluding any particular cultural perspective.

Topics welcomed by the journal include:

- the theory of scientific knowledge and the analysis of the general methodological problems of science (such as scientific discovery, causation, scientific inference, induction and probability, the structure of scientific theories and their relations with empirical data);
- the methodological and foundational problems of the different sciences, from the natural, to the biomedical, to the social sciences;
- the problems related to the historical development of logic, in all its branches, and to the role of logical methods both in the general methodology of science and in the foundations of empirical and mathematical sciences;
- the philosophical problems raised by the development of the cognitive sciences and the philosophy of mind, with particular attention to those results that are relevant for the analysis of scientific practice;
- the epistemological problems related to Artificial Intelligence, robotics, virtual reality, and artificial life;
- the problems in the sociology and the history of science that are relevant to the philosophical investigation of science;
- the problems related to the ethics of science;
- the questions related to the historical and conceptual development of the philosophy of science and logic;
- the problems of the philosophy of language, with particular attention to those results that are relevant for logic and philosophy of science.

Information for the authors

Papers submitted to the journal must be written either in Italian or in English, and must be accompanied by a short summary in English (and also in Italian for the articles written in Italian). All papers will be evaluated by anonymous referees.

In order to promote critical discussion and exchange among scholars, the journal is willing to publish reports on work in progress, to be submitted and evaluated according to the criteria already mentioned above.