Reference:

David Chalmers, 2003, "Consciousness and its Place in Nature", in Stephen P. Stich, Ted A. Warfield (eds.) *The Blackwell Guide to Philosophy of Mind*, Blackwell, Pp. 102 – 142

الوعي ومكانه في الطبيعة

دافيد تشالمرز

5-1 مقدمة[1]

يتوافق الوعي بصورة غير مريحة مع تصورنا عن العالم الطبيعي. بالنسبة لأكثر التـصورات شيوعا عن الطبيعة، العالم الطبيعي هو عالم فيزيائي. ولكن بالنسبة لأكثر التصورات شـيوعا عـن الوعي، ليس من السهل أن نرى كيف يمكن أن يكون جزءا من العالم الفيزيائي. لذلك يبدو أنه لكـي نجد مكانا للوعي خلال النظام الطبيعي، يجب علينا إما أن نراجع تصورنا عن الوعي، أو أن نراجـع تصورنا عن الطبيعة.

في فلسفة القرن العشرين، هذه المعضلة طرحت بأكثر أشكالها حدة في عمل "سي دي بـرود" (C. D. Broad) 'العقل ومكانه في الطبيعة (1925)'. ظواهر العقل، بالنسبة لـ برود، هي ظـواهر الوعي. المشكلة المركزية هي تلك الخاصة بتموضع العقل بالنسبة للعالم الفيزيائي. مناقـشة بـرود التفصيلية، التي استنفذت كافة الجوانب، توجت بتصنيف مكون من سبعة عشر نظرة مختلفة للعلاقة بين ما هو عقلي وما هو فيزيائي[2]. في تصنيف برود، يمكن لوجهة النظر أن ترى 'العقلي' على أنه غير موجود ("delusive")، كشيء يمكن رده إلى ما هو أبسط منه (reducible)، كشيء منبثق من شيء أبسط منه (emergent)، أو كسمة مميزة لجوهر (a "differentiating" attribute). يمكن لما هو "فيزيائي" أن يرى بواحدة من نفس هذه الطرق الأربعة. لذلك تنتج مصفوفة من أربعة في أربعة. (المدخل السابع عشر ينتج من تقسيم برود للنظرة إلى العلاقة جوهر/ جوهر طبقا لـدخول جـوهر واحد أو اثنين في التصنيف). في النهاية، تبقى ثلاثة وجهات نظر قائمة: تلـك التـي تكـون فيهـا 'العقلانية' سمة منبثقة إما من جوهر فيزيائي أو من جوهر محايد، في حين أن الحالـة الأخيـرة، 'الفيزيائي' يمكن أن يكون إما منبثق أو غير موجود 'delusive'.

في هذا الفصل سوف آخذ نقطة البدء من برود، مقاربا مشكلة الوعي بواسـطة إسـتراتيجية التقسيم والاستبعاد. لن أتبنى مقولات برود: فقد تقدم فهمنـا للعلاقـة بـين المـادة والعقـل منـذ العشرينيات، ولسوف يكون لطيفا أن نفكر بأننا نملك فهما أفضل للقضايا الأكثر أهمية وخطورة. من

وجهة نظري، وجهات النظر الأكثر أهمية فيما يخص ميتافيزيقا الوعي يمكن تقسيمها إلى ستة أقسام مستقلة تماما. سوف أسمى الأنواع من "نوع "أ" حتى "نوع و" ("type A" through "type F") ثلاثة منهم (أ حتى ج) تتضمن عموما نظرات ردية (reductive)، تنظر إلى الوعي كعملية فيزيائية لا تتضمن الامتداد إلى الأنطولوجيا الفيزيائية. والثلاثة الأخر (د حتى و) تتضمن عموما نظرات غير ردية، والتي يتضمن الوعي فيها شيئا لا يمكن رده إلى الطبيعة، ويتطلب امتداد أو مراجعة تصورنا عن الأنطولوجيا الفيزيائية.

سوف تصاغ المناقشة على المستوى المجرد، معطيا نظرة عامة على المشهد الميتافيزيقي. بدلا من إشراك علم الوعي الإمبيريقي، أو تفصيلات النظريات الفلسفية للوعي، سوف أقوم بفحص بعض الفئات العامة التي يمكن أن تقع فيها نظريات الوعي. لن أتظاهر بالمحايدة في هذه المناقشة. أعتقد أن كل من وجهات النظر الردية غير صحيحة، في حين أن كل من وجهات النظر غير الردية تحمل قدرا من الوعد. لذلك فالجزء الأول من هذا الفصل يمكن أن ينظر إليه كامتداد للحجة ضد وجهات النظر الردية للوعي، في حين أن الجزء الثاني يمكن النظر إليه على أنه فحص إلى أين نذهب بعد ذلك.

2-5 المشكلة

كلمة "الوعي" تستخدم بطرق متعددة بشكل كبير. تستعمل أحيانا للتعبير عن القدرة على التمييز بين المؤثرات الحسية، أو لتقرير المعلومات، أو لمراقبة الحالات الداخلية، أو للسيطرة على السلوك. يمكن لنا أن نفكر في هذه الظواهر على أنها تضع "المشكلات السهلة" للوعي. هذه ظواهر مهمة، وهناك الكثير مما قد فهم عنها، ولكن المشكلات المرتبطة بتفسيرها لها طبيعة الألغاز وليس الأسرار. لا يبدوا من حيث المبدأ أن هناك مشكلة عميقة بالنسبة لفكرة أن منظومة فيزيائية يمكن أن تكون "واعية" بهذا المعنى، كما أنه ليس هناك عائق واضح لتفسير نهائي لهذه الظواهر تأسيسا على مصطلحات بيولوجية-عصبية أو حاسوبية.

المشكلة الصعبة في الوعي هي مشكلة الخبرة. الإنسان يمتلك خبرة ذاتية: هناك شيء ما يشبه المرور بها. يمكننا القول بأن موجود ما هو واع بهذا المعنى – أو كما يقال أحيانا أنه واع ظاهريا (Phenomenally consciousness) – عندما يكون هناك شيئا ما مماثل لأن تكون هذا الموجود. تكون الحالة العقلية واعية عندما يكون هناك شيئا ما مماثل للمرور بهذه الحالة. الحالات الواعية تتضمن حالات الخبرة الإدراكية، الإحساسات البدنية، الخيالات العقلية، الخبرة العاطفية، الفكرة الطارئة، وأكثر من ذلك. هناك شيئا ما يشبه لأن نرى لونا أخضر لامع، للإحساس بالألم

الحاد، لرؤية برج إيفل، للشعور بالإحباط العميق، أو للتفكير في أن المرء قد تأخر. كل من هذه الحالات له سمة ظاهرية (Phenomenal character)، مع صفات ظاهرية (Phenomenal properties)، أو كيفيات (qualia) تميز ما يشبهه أن تكون في مثل هذه الحالة[3].

ليس هناك خلاف على أن الخبرة مرتبطة بشكل كبير بالعمليات الفيزيائية في منظومات كمثل العقل. يبدو أن العمليات الفيزيائية تعطي الفرصة للخبرة في الظهور، على الأقل بمعنى أن إنتاج منظومة فيزيائية (مثل العقل) مع الخواص الفيزيائية الصحيحة ينتج بالضرورة حالات الخبرة. ولكن كيف ولماذا تؤدي العمليات الفيزيائية إلى ظهور الخبرة؟ لماذا لا تحدث هذه العمليات "في الظلام"، بدون أن تصاحبها عمليات الخبرة؟ هذا هو السر المركزي في الوعي.

ما الذي يجعل المشكلات السهلة سهلة؟ بالنسبة لهذه المشكلات، المهمة هي تفسير بعض الوظائف السلوكية أو المعرفية: أي، تفسير كيف تلعب بعض الأدوار السببية في المنظومة المعرفية، لتؤدي في النهاية إلى إنتاج السلوك. لكي نفسر طبيعة عمل هذه الوظيفة، لا نحتاج سوى إلى تحديد الآلية (mechanism) التي تلعب الدور المطلوب. وهناك سبب وجيه للاعتقاد أن الآليات العصبونية (neural) أو الحاسوبية (computational) يمكنها أن تلعب هذه الأدوار.

ما الذي يجعل المشكلة الصعبة صعبة؟ هنا، المهمة ليست تفسير الوظائف المعرفية والسلوكية: حتى ولو كان عند المرء تفسير لكل الوظائف المطلوبة بالقرب من الوعي – التمييز، التكامل، الوصول للمعلومات، التقرير، التحكم – سيظل هناك سؤال إضافي: لماذا يكون إجراء هذه الوظائف مصحوبا بالخبرة؟ بسبب ذلك، تبدو المشكلة الصعبة نوع مختلف من المشكلة، تتطلب نوع مختلفا من الحل.

الحل المطلوب للمشكلة الصعبة سوف يتضمن افتراض بخصوص العلاقة بين العمليات الفيزيائية والوعي، يفسر على أساس المبادئ الطبيعية كيف ولماذا تكون العمليات الفيزيائية مرتبطة بحالات الخبرة. التفسير الردي (reductive explanation) للوعي سوف يفسر ذلك كله على أساس مبادئ فيزيائية ليست لها، في ذاتها، أي علاقة بالوعي.[4] الحل المادي (materialist)، أو الفيزيائي، سوف يكون حلا ينظر فيه إلى الوعي ذاته على أنه عملية فيزيائية. الحل غير المادي (non-materialist)، أو غير الفيزيائي، سوف يكون حلا ينظر فيه إلى الوعي على أنه غير فيزيائي (حتى ولو كان مرتبطا بشكل وثيق بالعمليات الفيزيائية). الحل غير الردي (non-reductive solution) هو ذلك الذي سوف يكون فيه الوعي (أو مبادئ تتضمن الوعي) مقدما كجزء أساسي من عملية التفسير.

من الطبيعي أن يكون هناك أمل في أن يوجد حل مادي للمشكلة الصعبة وتفسير ردي للوعي، بمثل ما كان هناك تفسيرات ردية للعديد من الظواهر الأخرى في مجالات عديدة أخرى. ولكن الوعي يبدو أنه يقاوم التفسير المادي بطريقة لا تماثل الظواهر الأخرى. هذه المقاومة يمكن تركيزهـا فـي ثلاثة حجج ضد المادية، يمكن تلخيصها كما يلي.

5-3 ثلاثة حجج ضد المادية

5-3-1 الحجة التفسيرية (The explanatory argument)[5]،

الحجة الأولى تتأسس على التفرقة بين المشكلات السهلة والمشكلة الصعبة، كما تم توصـيفها أعلاه: المشكلات السهلة تخص تفسير الوظائف السلوكية والمعرفية، ولكن المشكلة الصعبة ليـست كذلك. يمكن للمرء أن يجادل بأنه بطبيعة التفسير الفيزيائي، التقريرات الفيزيائية تفسر "فقط" الوظيفة والبنية، حيث تكون البنيات المطلوب تفسيرها "مكانية – زمانية" (spatio-temporal)، وتقـوم تلـك الوظائف بأدوار سببية في إنتاج سلوك النظام. ويمكن للمرء أن يجادل كما هو أعلاه بـأن تفـسير البنيات والوظائف ليس كافيا لتفسير الوعي. إذا كان كذلك، ليس هناك تصور فيزيائي يمكنه أن يفسر الوعي.

يمكن أن نسمى ذلك الحجة التفسيرية (the explanatory argument):

(1) التصورات الفيزيائية تفسر على الأكثر البنية والوظيفة.
(2) تفسير البنية والوظيفة ليس كافيا لتفسير الوعي.

---

(3) ليس هناك تصورا فيزيائيا يمكنه تفسير الوعي.

إذا كان ذلك صحيحا، فإنه في حين أن التصورات الفيزيائية يمكنها أن تحل المشكلات الـسهلة (والتي لا تتضمن سوى تفسيرات للوظائف)، فهناك شيء إضافي نحتاجه لكي نحل المشكلة الصعبة. سوف يبدو أنه ليس هناك تفسيرا رديا للوعي يمكنه أن ينجح. وإذا أضفنا إلى ذلك المقدمة المنطقية القائلة بأن ما لا يمكننا فيزيائيا تفسيره ليس في ذاته فيزيائي (وهذا يمكن اعتباره خطـوة نهائيـة إضافية للحجة التفسيرية)، فتكون المادية خاطئة كتفسير للوعي، ويكون العالم الطبيعي متضمنا مـا هو أكثر من العالم الفيزيائي.

بطبيعة الحال الحجة من هذا النوع قابلة للنقاش. ولكن قبل فحص الطرق المختلفة للرد عليها، من المفيد فحص حجتان قريبتان أخريان تهدفان أيضا لإثبات أن المادية بخصوص الوعي خاطئة.

5-3-2 حجة إمكانية التصور (The conceivability argument)[6]،

طبقا لهذه الحجة، من المتصور أن يكون هناك منظومة متطابقة فيزيائيا مع موجود واع، ولكنها تفتقر على الأقل لبعض الحالات الواعية لهذا الموجود. مثل هذه المنظومة يمكن أن تكون "زومبي" (Zombie): منظومة متطابقة مع موجود واع ولكنها تفتقر كليا إلى الوعي. ويمكن أيضا أن تكون "مقلوبة" (invert)، بها بعض الخبرات الخاصة بالموجود الأصلي مستبدلة بخبرات مختلفة، أو "زومبي جزئي" (partial zombei) مع بعض الخبرات الغائبة، أو تجميعا من هذه الحالات المختلفة. هذه الأنظمة سوف تبدو متطابقة مع الموجود الواعي من منظور الشخص الثالث: على وجه الخصوص عملياتها المخية سوف تبدو جزيء-جزيء متطابقة مع الأصل، وسوف يكون سلوكها غير قابل للتمييز عنه. ولكن هناك أشياء سوف تكون مختلفة من وجهة نظر الشخص الأول. ما يشبه أن تكون "مقلوبا"، أو "زومبي جزئي" سوف يختلف عما يشبه أن تكون الموجود الأصلي. وليس هناك ما يشبه أن تكون "زومبي".

هناك سبب محدود للاعتقاد بأن "الزومبيات" [جمع زومبي] يمكن أن توجد في العالم الطبيعي. ولكن الكثيرين يرون أنها على الأقل قابلة للتصور: يمكن بشكل متسق مع ذاته تصور "الزومبيات"، وليس هناك تناقض في الفكرة التي تكشف عن نفسها حتى ولو في التفكير. وكامتداد للفكرة، الكثير يعتقدون أن نفس الشيء ينطبق على عالم الزومبي (zombie world): عالم فيزيائي متطابق مع عالمنا، ولكن ليس فيه وعي. شيء مشابه ينطبق على "المقلوبين" والنسخ الأخرى.

من إمكانية تصور "الزومبيات"، يستند المدافعون عن هذه الحجة إمكانيتها الميتافيزيقية. "الزومبيات" ربما لا تكون ممكنة فيزيائيا: هي ربما لا تكون ممكنة الوجود في عالمنا، بقوانينه الطبيعية. لكن الحجة تقول بأن "الزومبيات" كان من الممكن أن تكون موجودة، ربما في كون من نوع مختلف جدا. على سبيل المثال، في بعض الأحيان يتم الاقتراح بأن الله كان يمكنه أن يخلق "عالم زومبي"، إذا اختار ذلك. من هنا، تم الاستنتاج بأن الوعي لا بد وأن لا يكون فيزيائيا. إذا كان ممكنا ميتافيزيقيا أن يكون هناك كون مطابق لنا فيزيائيا ولكن ينقصه الوعي، فإن الوعي لا بد وأن يكون مكونا إضافيا غير فيزيائيا لكوننا. إذا كان الله يمكنه أن يخلق "عالما زومبيا"، فإنه (كما يقول كريبك"Kripke") بعد خلق العمليات الفيزيائية في عالمنا، يجب أن يفعل شيئا إضافيا ليضمن أنه يحتوي على الوعي.

يمكننا أن نضع هذه الحجة في أبسط اشكالها كما يلي:

(1) يمكن تصور وجود "زومبيات".

(2) إذا كان من الممكن تصور وجود "الزومبيات"، فإنه يكون ممكنا ميتافيزيقيا وجود "الزومبيات".

(3) إذا ممكنا ميتافيزيقيا وجود "زومبيات"، فيكون الوعي غير فيزيائي.

(4) الوعي غير فيزيائي

نسخة أكثر عمومية ودقة لهذه الحجة تعود إلى "ف"، الفئة التي تمثل المشترك ما بين كل الحقائق الميكروفيزيائية عن الكون، و "ق"، حقيقة ظاهرية عشوائية عن الكون. (هنا "^" تمثل "و"، و"¬" تمثل "ليس".)

(1) ليس من الممكن تصور أن ف^ ¬ق.
(2) إذا كان من الممكن تصور أن ف^¬ق، يكون من الممكن ميتافيزيقيا أن ف^¬ق.
(3) إذا كان من الممكن ميتافيزيقيا تصور أن ف^¬ق، تكون المادية خاطئة.

(4) المادية خاطئة.

5-3-3 حجة المعرفة (The knowledge argument)[7]،

طبقا لحجة المعرفة، هناك حقائق عن الوعي لا يمكن استنتاجها من الحقائق الفيزيائية. يمكن لشخص ما أن يعرف كل الحقائق الفيزيائية، وأن يمتلك قدرة استنتاجية كاملة، ويبقى مع ذلك غير قادر على معرفة كل الحقائق عن الوعي على أساسها.

نسخة فرانك جاكسون القياسية لهذه الحجة تقدم بيانا واضحا. في هذه النسخة، ماري هي عالمة في علم الخلايا (neuroscientist) تعرف كل شيء هناك يمكن معرفته عن العمليات الفيزيائية المتعلقة بالرؤية اللونية. ولكن ماري قد شبت في حجرة من الأسود والأبيض (في نسخة بديلة هي مصابة بعمى الألوان[8]) ولم يتح لها أن تحصل على خبرة اللون الأحمر. بالرغم من كل معرفتها، يبدو أن هناك شيء مهم جدا عن الرؤية اللونية لا تعرفه ماري: هي لا تعرف ماذا يشبه أن ترى اللون الأحمر. حتى المعرفة الفيزيائية الكاملة والقدرة غير المحدودة على الاستنباط لا يمكنانها من معرفة ذلك. مؤخرا، إذا مرت بخبرة اللون الأحمر لأول مرة، سوف تتعلم حقيقة جديدة كانت تجهلها سابقا: سوف تتعلم ماذا يشبه أن ترى الأحمر.

نسخة جاكسون لهذه الحجة يمكن أن توضع كما يلي (هنا المقدمات تختص بمعرفة ماري عندما لم تكن تمتلك خبرة الأحمر):

(1) ماري تعرف كل الحقائق الفيزيائية.
(2) ماري لا تعرف كل الحقائق.

(3) الحقائق الفيزيائية لا تستنفد كل الحقائق.

يمكن للمرء أن يضع حجة المعرفة بشكل أكثر عمومية:

(1) هناك حقائق عن الوعي لا يمكن استخلاصها من الحقائق الفيزيائية.

(2) إذا كان هناك حقائق عن الوعي لا يمكن استخلاصها الحقائق الفيزيائية، فإن المادية تكون خاطئة.

(3) المادية خاطئة

5-3-4 صورة الحجة

هذه الأنواع الثلاث من الحجة مرتبطة ببعضها ارتباطا وثيقا. هي جميعها تؤسس فجوة معرفية (epistemic gap) بين المجالين الفيزيائي والظاهري. كل منها ينكر نوع معين من علاقة التقارب المعرفي بين المجالين: علاقة تتضمن ما يمكن معرفته، أو تصوره، أو تفسيره. على وجه الخصوص، كل منها ينكر نوع من معين من الاستتباع المعرفي (epistemic entailment) من الحقائق الفيزيائية (ف) إلى الحقائق الظاهرية (ق): استنباط (ق) من (ف)، أو القدرة على تفسير (ق) على أساس (ف)، أو القدرة على تصور (ق) بناء على التصور التأملي لـ (ف).

ربما يكون النوع الأكثر أساسية للاستتباع المعرفي هو الاستتباع القبلي (a priori)، أو التضمن (implication). في هذه الفكرة، (ف) تتضمن (ق) عندما تكون العلاقة المادية الشرطية (ق ⊃ ف) ثابتة قبليا؛ أي أنه، عندما يمكن للشخص أن يعرف أنه إذا كانت ف هي الحالة فإنه ق تكون هي الحالة، على أن يكون التبرير مستقل عن الخبرة. كل هذه الحجج الثلاثة أعلاه يمكن أن ترى على أنها تثبت وجهة النظر ضد الاستتباع القبلي لـ ق بواسطة ف . إذا كان من غير الممكن للشخص الذي يعرف فقط ف أن يستنبط ق (كما تقترح حجة المعرفة)، أو أنه يمكن للمرء أن يتصور عقلانيا ف بدون ق (كما تقترح حجة التصور)، فإنه يبدو أن ف لا تتضمن ق. الحجة التفسيرية يمكن النظر إليها على أنها دعوى بأن التضمن من ف إلى ق سوف يتطلب تحليلا وظيفيا للوعي، وأن تصور الوعي ليس تصورا وظيفيا.

بعد تأسيس هوة معرفية، تستمر هذه الحجج في الاستدلال على هوة أنطولوجية (ontological gap)، حيث تهتم الأنطولوجيا بطبيعة الأشياء في العالم. حجة التصور تستخلص من إمكانية التصور الإمكانية الميتافيزيقية؛ حجة المعرفة تستخلص من فشل الاستنباط الاختلاف في الحقائق؛ والحجة التفسيرية تستخلص من فشل التفسير الفيزيائي اللافيزيائية. يمكن للمرء القول بأن هذه الحجج تستخلص من فشل الاستتباع المعرفي الفشل في الاستتباع الأنطولوجي. النوع النموذجي (paradigmatic) للاستتباع الأنطولوجي هو اللزوم (necessitation): ف تستلزم ق عندما يكون الشرط المادي (ق⊃ف) لازم ميتافيزيقا، أو عندما يكون من المستحيل ميتافيزيقيا لـ ف أن تحصل بدون حصول ق. من المتفق عليه بشكل واسع أن المادية تتطلب أن تكون كل الحقائق لازمة من ف. لهذا إذا كانت هناك حقائق ظاهرية ق لا تتطلبها ف، فإن المادية تكون خاطئة.

يمكن لنا أن نسمي هذه الحجج *الحجج المعرفية ضد المادية* (epistemic arguments against materialism). ربما تعود الحجج المعرفية إلى حجج ديكارت ضد المادية (على الرغم أنها لها شكل مختلف قليلا)، وأنها قد أعطيت أول انتشار بحثي واسع في كتاب برود، والذي يحتوى على عناصر كل الحجج الثلاث أعلاه.[9]

الصورة العامة لحجة معرفية ضد المادية هي كما يلي:

(1) هناك هوة معرفية بين الحقائق الفيزيائية والظاهرية.
(2) إذا كانت هناك هوة معرفية بين الحقائق الفيزيائية والظاهرية، فإن هناك هوة أنطولوجية، والمادية تكون على خطأ.

(3) المادية على خطأ

بطبيعة الحال هذه الطريقة للنظر إلى الأشياء تبسط بشكل كبير الموضوع، وتجرد بعيدا الاختلافات بين الحجج.[10] نفس الشيء يصدق على التحليل الدقيق الخاص بالتضمن واللزوم. ومع ذلك، هذا التحليل يقدم عدسة نافعة يمكن من خلالها رؤية ما تشترك فيه هذه الحجج، ويمكن من خلالها تحليل الردود المختلفة لها.

هناك على وجه التقريب ثلاثة طرق يمكن من خلالها لأصحاب الاتجاه المادي مقاومة الحجج المعرفية. المادي من نوع أ (type-A materialist) ينكر وجود هذا النوع المشار إليه من الهوة المعرفية. المادي من نوع ب (type-B materialist) يقبل وجود هوة معرفية غير مغلقة، ولكنه ينكر وجود هوة أنطولوجية. والمادي من النوع ج (type-C materialist) يقبل وجود هوة معرفية عميقة، ولكنه يعتقد بأنها سوف تغلق في نهاية المطاف. فيما يلي، سوف أناقش كل هذه الاستراتيجيات الثلاث.

5-4 المادية من النوع أ (Type A Materialism)

طبقا للمادية من نوع أ، ليس هناك فجوة معرفية بين الحقائق الفيزيائية والحقائق الظاهرية، أو على الأقل، أي فجوة معرفية ظاهرة يمكن بسهولة إغلاقها. طبقا لوجهة النظر هذه، ليس من المتصور (أو على الأقل مع التفكير) أن يوجد نسخ من الموجودات الواعية التي ليس لها حالات، أو تمتلك حالات مقلوبة، واعية. عند وجهة النظر هذه، ليس هناك حقائق ظاهرية تجهلها ماري من حيث المبدأ من داخل حجرتها ذات الأبيض والأسود (عندما تترك الغرفة، سوف تكتسب فورا هذه الإمكانية). وطبقا لوجهة النظر هذه، مع إعمال التفكير ليس هناك "مشكلة صعبة" في تفسير الوعي تبقى بعد أن يحل المرء المشكلات السهلة لتفسير الوظائف المعرفية والسلوكية والبيئية المتعددة.[11]

explaining consciousness that remains once one has solved the easy problems of explaining the various cognitive, behavioral, and environmental functions.[11]

Type-A materialism sometimes takes the form of eliminativism, holding that consciousness does not exist, and that there are no phenomenal truths. It sometimes takes the form of analytic functionalism or logical behaviorism, holding that consciousness exists, where the concept of "consciousness" is defined in wholly functional or behavioral terms (e.g., where to be conscious might be to have certain sorts of access to information, and/or certain sorts of dispositions to make verbal reports). For our purposes, the difference between these two views can be seen as terminological. Both agree that we are conscious in the sense of having the functional capacities of access, report, control, and the like; and they agree that we are not conscious in any further (non-functionally defined) sense. The analytic functionalist thinks that ordinary terms such as "conscious" should be used in the first sort of sense (expressing a functional concept), while the eliminativist thinks that they should be used in the second. Beyond this terminological disagreement about the use of existing terms and concepts, the substance of the views is the same.

Some philosophers and scientists who do not explicitly embrace eliminativism, analytic functionalism, and the like are nevertheless recognizably type-A materialists. The characteristic feature of the type-A materialist is the view that on reflection there is nothing in the vicinity of consciousness that needs explaining over and above explaining the various functions: to explain these things is to explain everything in the vicinity that needs to be explained. The relevant functions may be quite subtle and complex, involving fine-grained capacities for access, self-monitoring, report, control, and their interaction, for example. They may also be taken to include all sorts of environmental relations. And the explanation of these functions will probably involve much neurobiological detail. So views that are put forward as rejecting functionalism on the grounds that it neglects biology or neglects the role of the environment may still be type-A views.

One might think that there is room in logical space for a view that denies even this sort of broadly functionalist view of consciousness, but still holds that there is no epistemic gap between physical and phenomenal truths. In practice, there appears to be little room for such a view, for reasons that I will discuss under type C, and there are few examples of such views in practice.[12] So I will take it for granted that a type-A view is one that holds that explaining the functions explains everything, and will class other views that hold that there is no unclosable epistemic gap under type C.

The obvious problem with type-A materialism is that it appears to deny the manifest. It is an uncontested truth that we have the various functional capacities of access, control, report, and the like, and these phenomena pose uncontested explananda (phenomena in need of explanation) for a science of consciousness. But in addition, it seems to be a further truth that we are conscious, and this phenomenon seems to pose a further explanandum. It is this explanandum that raises the interesting problems of consciousness. To flatly deny the further truth, or to deny without argument that there is a hard problem of consciousness over

and above the easy problems, would be to make a highly counterintuitive claim that begs the important questions. This is not to say that highly counterintuitive claims are always false, but they need to be supported by extremely strong arguments. So the crucial question is: are there any compelling *arguments* for the claim that, on reflection, explaining the functions explains everything?

Type-A materialists often argue by analogy. They point out that in other areas of science, we accept that explaining the various functions explains the phenomena, so we should accept the same here. In response, an opponent may well accept that in other domains the functions are all we need to explain. In explaining life, for example, the only phenomena that present themselves as needing explanation are phenomena of adaptation, growth, metabolism, reproduction, and so on, and there is nothing else that even calls out for explanation. But the opponent holds that the case of consciousness is different and possibly unique, precisely because there is something else, phenomenal experience, that calls out for explanation. The type-A materialist must either deny even the appearance of a further explanandum, which seems to deny the obvious, or accept the apparent disanalogy and give further substantial arguments for why, contrary to appearances, only the functions need to be explained.

At this point, type-A materialists often press a different sort of analogy, holding that at various points in the past, thinkers held that there was an analogous epistemic gap for other phenomena, but that these turned out to be physically explained. For example, Dennett (1996) suggests that a vitalist might have held that there was a further "hard problem" of life over and above explaining the biological function, but that this would have been misguided.

On examining the cases, however, the analogies do not support the type-A materialist. Vitalists typically *accepted*, implicitly or explicitly, that the biological functions in question were what needed explaining. Their vitalism arose because they thought that the functions (adaptation, growth, reproduction, and so on) would not be physically explained. So this is quite different from the case of consciousness. The disanalogy is very clear in the case of Broad. Broad was a vitalist about life, holding that the functions would require a non-mechanical explanation. But at the same time, he held that in the case of life, unlike the case of consciousness, the only evidence we have for the phenomenon is behavioral, and that "being alive" means exhibiting certain sorts of behavior. Other vitalists were less explicit, but very few of them held that something more than the functions needed explaining (except consciousness itself, in some cases). If a vitalist had held this, the obvious reply would have been that there is no reason to believe in such an explanandum. So there is no analogy here.[13]

So these arguments by analogy have no force for the type-A materialist. In other cases, it was always clear that structure and function exhausted the apparent explananda, apart from those tied directly to consciousness itself. So the type-A materialist needs to address the apparent further explanandum in the case of consciousness head on: either flatly denying it, or giving substantial arguments to dissolve it.

Some arguments for type-A materialists proceed indirectly, by pointing out the unsavory metaphysical or epistemological consequences of rejecting the view: e.g., that the rejection leads to dualism, or to problems involving knowledge of consciousness.[14] An opponent will either embrace the consequences or deny that they are consequences. As long as the consequences are not completely untenable, then for the type-A materialist to make progress, this sort of argument needs to be supplemented by a substantial direct argument against the further explanandum.

Such direct arguments are surprisingly hard to find. Many arguments for type-A materialism end up presupposing the conclusion at crucial points. For example, it is sometimes argued (e.g., Rey 1995) that there is no reason to postulate qualia, since they are not needed to explain behavior; but this argument presupposes that only behavior needs explaining. The opponent will hold that qualia are an explanandum in their own right. Similarly, Dennett's (1991) use of "heterophenomenology" (verbal reports) as the primary data to ground his theory of consciousness appears to rest on the assumption that these reports are what need explaining, or that the only "seemings" that need explaining are dispositions to react and report.

One way to argue for type-A materialism is to argue that there is some intermediate X such that (i) explaining functions suffices to explain X, and (ii) explaining X suffices to explain consciousness. One possible X here is *representation*: it is often held both that conscious states are representational states, representing things in the world, and that we can explain representation in functional terms. If so, it may seem to follow that we can explain consciousness in functional terms. On examination, though, this argument appeals to an ambiguity in the notion of representation. There is a notion of *functional representation*, on which P is represented roughly when a system responds to P and/or produces behavior appropriate for P. In this sense, explaining functioning may explain representation, but explaining representation does not explain consciousness. There is also a notion of *phenomenal representation*, on which P is represented roughly when a system has a conscious experience as if P. In this sense, explaining representation may explain consciousness, but explaining functioning does not explain representation. Either way, the epistemic gap between the functional and the phenomenal remains as wide as ever. Similar sorts of equivocation can be found with other X's that might be appealed to here, such as "perception" or "information."

Perhaps the most interesting arguments for type-A materialism are those that argue that we can give a physical explanation of our *beliefs* about consciousness, such as the belief that we are conscious, the belief that consciousness is a further explanandum, and the belief that consciousness is non-physical. From here it is argued that once we have explained the belief, we have done enough to explain, or to explain away, the phenomenon (e.g., Clark 2000, Dennett forthcoming). Here it is worth noting that this only works if the beliefs themselves are functionally analyzable; Chalmers (2002a) gives reason to deny this. But even if one accepts that beliefs are ultimately functional, this claim then reduces to the claim that explaining

our dispositions to talk about consciousness (and the like) explains everything. An opponent will deny this claim: explaining the dispositions to report may remove the third-person warrant (based on observation of others) for accepting a further explanandum, but it does not remove the crucial first-person warrant (from one's own case). Still, this is a strategy that deserves extended discussion.

At a certain point, the debate between type-A materialists and their opponents usually comes down to intuition: most centrally, the intuition that consciousness (in a non-functionally defined sense) exists, or that there is something that needs to be explained (over and above explaining the functions). This claim does not gain its support from argument, but from a sort of observation, along with rebuttal of counterarguments. The intuition appears to be shared by the large majority of philosophers, scientists, and others; and it is so strong that to deny it, a type-A materialist needs exceptionally powerful arguments. The result is that even among materialists, type-A materialists are a distinct minority.

5.5 Type-B Materialism[15]

According to type-B materialism, there is an epistemic gap between the physical and phenomenal domains, but there is no ontological gap. According to this view, zombies and the like are conceivable, but they are not metaphysically possible. On this view, Mary is ignorant of some phenomenal truths from inside her room, but nevertheless these truths concern an underlying physical reality (when she leaves the room, she learns old facts in a new way). And on this view, while there is a hard problem distinct from the easy problems, it does not correspond to a distinct ontological domain.

The most common form of type-B materialism holds that phenomenal states can be *identified* with certain physical or functional states. This identity is held to be analogous in certain respects (although perhaps not in all respects) with the identity between water and H_2O, or between genes and DNA.[16] These identities are not derived through conceptual analysis, but are discovered empirically: the concept *water* is different from the concept H_2O, but they are found to refer to the same thing in nature. On the type-B view, something similar applies to consciousness: the concept of consciousness is distinct from any physical or functional concepts, but we may discover empirically that these refer to the same thing in nature. In this way, we can explain why there is an epistemic gap between the physical and phenomenal domains, while denying any ontological gap. This yields the attractive possibility that we can acknowledge the deep epistemic problems of consciousness while retaining a materialist worldview.

Although such a view is attractive, it faces immediate difficulties. These difficulties stem from the fact that the character of the epistemic gap with consciousness seems to differ from that of epistemic gaps in other domains. For a start, there do not seem to be analogs of the epistemic arguments above in the cases of water,

genes, and so on. To explain genes, we merely have to explain why systems function a certain way in transmitting hereditary characteristics; to explain water, we have to explain why a substance has a certain objective structure and behavior. Given a complete physical description of the world, Mary would be able to deduce all the relevant truths about water and about genes, by deducing which systems have the appropriate structure and function. Finally, it seems that we cannot coherently conceive of a world physically identical to our own, in which there is no water, or in which there are no genes. So there is no epistemic gap between the *complete* physical truth about the world and the truth about water and genes that is analogous to the epistemic gap with consciousness.

(Except, perhaps, for epistemic gaps that derive from the epistemic gap for consciousness. For example, perhaps Mary could not deduce or explain the perceptual *appearance* of water from the physical truth about the world. But this would just be another instance of the problem we are concerned with, and so cannot help the type-B materialist.)

So it seems that there is something unique about the case of consciousness. We can put this by saying that while the identity between genes and DNA is empirical, it is not *epistemically primitive*: the identity is itself deducible from the complete physical truth about the world. By contrast, the type-B materialist must hold that the identification between consciousness and physical or functional states is epistemically primitive: the identity is not deducible from the complete physical truth. (If it were deducible, type-A materialism would be true instead.) So the identity between consciousness and a physical state will be a sort of primitive principle in one's theory of the world.

Here, one might suggest that something has gone wrong. Elsewhere, the only sort of place that one finds this sort of primitive principle is in the fundamental laws of physics. Indeed, it is often held that this sort of primitiveness – the inability to be deduced from more basic principles – is the mark of a fundamental law of nature. In effect, the type-B materialist recognizes a principle that has the epistemic status of a fundamental law, but gives it the ontological status of an identity. An opponent will hold that this move is more akin to theft than to honest toil: elsewhere, identifications are grounded in explanations, and primitive principles are acknowledged as fundamental laws.

It is natural to suggest that the same should apply here. If one acknowledges the epistemically primitive connection between physical states and consciousness as a fundamental law, it will follow that consciousness is distinct from any physical property, since fundamental laws always connect distinct properties. So the usual standard will lead to one of the non-reductive views discussed in the second half of this chapter. By contrast, the type-B materialist takes an observed connection between physical and phenomenal states, unexplainable in more basic terms, and suggests that it is an identity. This suggestion is made largely in order to preserve a prior commitment to materialism. Unless there is an independent case for primitive identities, the suggestion will seem at best ad hoc and mysterious, and at worst incoherent.

A type-B materialist might respond in various ways. First, some (e.g., Papineau 1993) suggest that identities do not *need* to be explained, so are always primitive. But we have seen that identities in other domains can at least be *deduced* from more basic truths, and so are not primitive in the relevant sense. Secondly, some (e.g., Block and Stalnaker 1999) suggest that even truths involving water and genes cannot be deduced from underlying physical truths. This matter is too complex to go into here (see Chalmers and Jackson 2001 for a response[17]), but one can note that the epistemic arguments outlined at the beginning suggest a very strong disanalogy between consciousness and other cases. Thirdly, some (e.g., Loar 1990/1997) acknowledge that identities involving consciousness are unlike other identities by being epistemically primitive, but seek to explain this uniqueness by appealing to unique features of the concept of consciousness. This response is perhaps the most interesting, and I will return to it.

There is another line that a type-B materialist can take. One can first note that an *identity* between consciousness and physical states is not strictly required for a materialist position. Rather, one can plausibly hold that materialism about consciousness simply requires that physical states *necessitate* phenomenal states, in that it is metaphysically impossible for the physical states to be present while the phenomenal states are absent or different. That is, materialism requires that entailments $P \supset Q$ be necessary, where P is the complete physical truth about the world and Q is an arbitrary phenomenal truth.

At this point, a type-B materialist can naturally appeal to the work of Kripke (1980), which suggests that some truths are necessarily true without being a priori. For example, Kripke suggests that "water is H_2O" is necessary – true in all possible worlds – but not knowable a priori. Here, a type-B materialist can suggest that $P \supset Q$ may be a Kripkean a posteriori necessity, like "water is H_2O" (though it should be noted that Kripke himself denies this claim). If so, then we would *expect* there to be an epistemic gap, since there is no a priori entailment from P to Q, but at the same time there will be no ontological gap. In this way, Kripke's work can seem to be just what the type-B materialist needs.

Here, some of the issues that arose previously arise again. One can argue that in other domains, necessities are not epistemically primitive. The necessary connection between water and H_2O may be a posteriori, but it can itself be deduced from a complete physical description of the world (one can deduce that water is identical to H_2O, from which it follows that water is necessarily H_2O). The same applies to the other necessities that Kripke discusses. By contrast, the type-B materialist must hold that the connection between physical states and consciousness is epistemically primitive, in that it cannot be deduced from the complete physical truth about the world. Again, one can suggest that this sort of primitive necessary connection is mysterious and ad hoc, and that the connection should instead be viewed as a fundamental law of nature.

I will discuss further problems with these necessities in the next section. But here, it is worth noting that there is a sense in which any type-B materialist position gives up on reductive explanation. Even if type-B materialism is true, we

cannot give consciousness the same sort of explanation that we give genes and the like, in purely physical terms. Rather, our explanation will always require explanatorily primitive principles to bridge the gap from the physical to the phenomenal. The *explanatory* structure of a theory of consciousness, on such a view, will be very much unlike that of a materialist theory in other domains, and very much like the explanatory structure of the non-reductive theories described below. By labeling these principles identities or necessities rather than laws, the view may preserve the letter of materialism; but by requiring primitive bridging principles, it sacrifices much of materialism's spirit.

5.6 The Two-Dimensional Argument Against Type-B Materialism

As discussed above, the type-B materialist holds that zombie worlds and the like are conceivable (there is no contradiction in $P\neg Q$) but are not metaphysically possible. That is, $P\supset Q$ is held to be an a posteriori necessity, akin to such a posteriori necessities as "water is H_2O." We can analyze this position in more depth by taking a closer look at the Kripkean cases of a posteriori necessity. This material is somewhat technical (hence the separate section) and can be skipped if necessary on a first reading.

It is often said that in Kripkean cases, conceivability does not entail possibility: it is conceivable that water is not H_2O (in that it is coherent to suppose that water is not H_2O), but it is not possible that water is not H_2O. But at the same time, it seems that there is *some* possibility in the vicinity of what one conceives. When one conceives that water is not H_2O, one conceives of a world W (the XYZ-world) in which the watery liquid in the oceans is not H_2O, but XYZ, say. There is no reason to doubt that the XYZ-world is metaphysically possible. If Kripke is correct, the XYZ-world is not correctly described as one in which water is XYZ. Nevertheless, this world is relevant to the truth of "water is XYZ" in a slightly different way, which can be brought out as follows.

One can say that the XYZ-world could *turn out* to be actual, in that for all we know a priori, the actual world is just like the XYZ-world. And one can say that *if* the XYZ-world turns out to be actual, it will turn out that water is XYZ. Similarly: if we hypothesize that the XYZ-world is actual, we should rationally conclude on that basis that water is not H_2O. That is, there is a deep *epistemic* connection between the XYZ-world and "water is not H_2O." Even Kripke allows that it is *epistemically possible* that water is not H_2O (in the broad sense that this is not ruled out a priori). It seems that the epistemic possibility that the XYZ-world is actual is a specific instance of the epistemic possibility that water is not H_2O.

Here, we adopt a special attitude to a world W. We think of W as an epistemic possibility: as a way the world might actually be. When we do this, we consider W *as actual*. When we think of W as actual, it may make a given sentence S true or

false. For example, when thinking of the XYZ-world as actual, it makes "water is not H_2O" true. This is brought out in the intuitive judgment that if W turns out to be actual, it will turn out that water is not H_2O, and that the epistemic possibility that W is actual is an instance of the epistemic possibility that water is H_2O.

By contrast, one can also consider a world W *as counterfactual*. When we do this, we acknowledge that the character of the actual world is already fixed, and we think of W as a counterfactual way things might have been but are not. If Kripke is right, then if the watery stuff *had been* XYZ, XYZ would nevertheless not have been water. So when we consider the XYZ-world as counterfactual, it does not make "water is not H_2O" true. Considered as counterfactual, we describe the XYZ-world in light of the actual-world fact that water is H_2O, and we conclude that XYZ is not water but merely watery stuff. These results do not conflict: they simply involve two different ways of considering and describing possible worlds. Kripke's claims consider *counterfactual* evaluation of worlds, whereas the claims in the previous paragraph concern the *epistemic* evaluation of worlds.

One can formalize this using *two-dimensional semantics*.[18] We can say that if W considered as actual makes S true, then W *verifies* S, and that if W considered as counterfactual makes S true, then W *satisfies* S. Verification involves the epistemic evaluation of worlds, whereas satisfaction involves the counterfactual evaluation of worlds. Correspondingly, we can associate S with different *intensions*, or functions from worlds to truth values. The *primary* (or epistemic) intension of *S* is a function that is true at a world W iff W verifies S, and the *secondary* (or subjunctive) intension is a function that is true at a world W if W satisfies S. For example, where S is "water is not H_2O," and W is the XYZ-world, we can say that W verifies S but W does not satisfy S; and we can say that the primary intension of S is true at W, but the secondary intension of S is false at W.

With this in mind, one can suggest that when a statement S is conceivable – that is, when its truth cannot be ruled out a priori – then there is some world that verifies S, or equivalently, there is some world at which S's primary intension is true. This makes intuitive sense: when S is conceivable, S represents an epistemic possibility. It is natural to suggest that corresponding to these epistemic possibilities are specific worlds W, such that when these are considered *as* epistemic possibilities, they verify S. That is, W is such that intuitively, if W turns out to be actual, it would turn out that S.

This model seems to fit all of Kripke's cases. For example, Kripke holds that it is an a posteriori necessity that heat is the motion of molecules. So it is conceivable in the relevant sense that heat is not the motion of molecules. Corresponding to this conceivable scenario is a world W in which heat sensations are caused by something other than the motion of molecules. W represents an epistemic possibility: and we can say that if W turns out to be actual, it will turn out that heat is not the motion of molecules. The same goes in many other cases. The moral is that these Kripkean phenomena involve two different ways of thinking of possible worlds, with just one underlying space of worlds.

If this principle is applied to the case of type-B materialism, trouble immediately arises. As before, let P be the complete physical truth about the world, and let Q be a phenomenal truth. Let us say that S is conceivable when the truth of S is not ruled out a priori. Then one can mount an argument as follows:[19]

(1) P∧¬Q is conceivable.
(2) If P∧¬Q is conceivable, then a world verifies P∧¬Q.
(3) If a world verifies P∧¬Q, then a world satisfies P∧¬Q or type-F monism is true.
(4) If a world satisfies P∧¬Q, materialism is false.

---

(5) Materialism is false or type-F monism is true.

The type-B materialist grants premise (1): to deny this would be to accept type-A materialism. Premise (2) is an instance of the general principle discussed above. Premise (4) can be taken as definitive of materialism. As for premise (3): in general one cannot immediately move from a world verifying S to a world satisfying S, as the case of "water is H_2O" (and the XYZ-world) suggests. But in the case of P∧¬Q, a little reflection on the nature of P and Q takes us in that direction, as follows.

First, Q. Here, it is plausible that if W verifies "there is consciousness," then W satisfies "there is consciousness," and vice versa. This corresponds to the Kripkean point that in the case of consciousness, there is no distinction analogous to that between water itself and mere watery stuff. To put it intuitively, if W verifies "there is consciousness," it contains something that at least *feels* conscious, and if something *feels* conscious, it *is* conscious. One can hold more generally that the primary and secondary intensions of our core phenomenal concepts are the same (see Chalmers 2002a). It follows that if world W verifies ¬Q, W satisfies ¬Q. (This claim is not required for the argument to go through, but it is plausible and makes things more straightforward.)

Second, P. A type-B materialist might seek to evade the argument by arguing that while W verifies P, it does not satisfy P. On reflection, the only way this might work is as follows. If a world verifies P, it must have at least the *structure* of the actual physical world. The only reason why W might not satisfy P is that it lacks the intrinsic properties underlying this structure in the actual world. (On this view, the primary intension of a physical concept picks out whatever property plays a certain role in a given world, and the secondary intension picks out the actual intrinsic property across all worlds.) If this difference in W is responsible for the absence of consciousness in W, it follows that consciousness in the actual world is not necessitated by the structural aspects of physics, but by its underlying intrinsic nature. This is precisely the position I call type-F monism, or "panprotopsychism." Type-F monism is an interesting and important position, but it is much more radical than type-B materialism as usually conceived, and I count it as a different position. I will defer discussion of the reasoning and of the resulting position until later.

It follows that premise (4) is correct. If a world verifies P∧¬Q, then either a world satisfies P∧¬Q, or type-F monism is true. Setting aside type-F monism for now, it follows that the physical truth about our world does not necessitate the phenomenal truth, and materialism is false.

This conclusion is in effect a consequence of (i) the claim that P∧¬Q is conceivable (in the relevant sense), (ii) the claim that when S is conceivable, there is a world that verifies S, and (iii) some straightforward reasoning. A materialist might respond by denying (i), but that is simply to deny the relevant epistemic gap between the physical and the phenomenal, and so to deny type-B materialism. I think there is little promise for the type-B materialist in denying the reasoning involved in (iii). So the only hope for the type-B materialist is to deny the central thesis (ii).[20]

To do this, a type-B materialist could deny the coherence of the distinction between verification and satisfaction, or accept that the distinction is coherent but deny that thesis (ii) holds even in the standard Kripkean cases, or accept that thesis (ii) holds in the standard Kripkean cases but deny that it holds in the special case of consciousness. The first two options deserve exploration, but I think they are ultimately unpromising, as the distinction and the thesis appear to fit the Kripkean phenomena very well. Ultimately, I think a type-B materialist must hold that the case of consciousness is special, and that the thesis that holds elsewhere fails here.

On this view, the a posteriori necessities connecting the physical and phenomenal domains are much stronger than those in other domains in that they are verified by all worlds. Elsewhere, I have called these unusual a posteriori necessities *strong necessities*, and have argued that there is no good reason to believe they exist. As with explanatorily primitive identities, they appear to be primitive facts postulated in an ad hoc way, largely in order to save a theory, with no support from cases elsewhere. Further, one can argue that this view leads to an underlying *modal dualism*, with independent primitive domains of logical and metaphysical possibility; and one can argue that this is unacceptable.

Perhaps the most interesting response from a type-B materialist is to acknowledge that strong necessities are unique to the case of consciousness, and to try to explain this uniqueness in terms of special features of our conceptual system. For example, Christopher Hill (1997) has argued that one can predict the epistemic gap in the case of consciousness from the fact that physical concepts and phenomenal concepts have different conceptual roles. Brian Loar (1990/1997) has appealed to the claim that phenomenal concepts are recognitional concepts that lack contingent modes of presentation. Joseph Levine (2000) has argued that phenomenal concepts have non-ascriptive modes of presentation. In response, I have argued (Chalmers 1999) that these responses do not work, and that there are systematic reasons why they cannot work.[21] But it is likely that further attempts in this direction will be forthcoming. This remains one of the key areas of debate on the metaphysics of consciousness.

Overall, my own view is that there is little reason to think that explanatorily primitive identities or strong necessities exist. There is no good *independent*

reason to believe in them: the best reason to postulate them is to save materialism, but in the context of a debate over whether materialism is true this reasoning is uncompelling, especially if there are viable alternatives. Nevertheless, further investigation into the key issues underlying this debate is likely to be philosophically fruitful.

5.7 Type-C Materialism

According to type-C materialism, there is a deep epistemic gap between the physical and phenomenal domains, but it is closable in principle. On this view, zombies and the like are conceivable for us now, but they will not be conceivable in the limit. On this view, it currently seems that Mary lacks information about the phenomenal, but in the limit there would be no information that she lacks. And on this view, while we cannot see now how to solve the hard problem in physical terms, the problem is solvable in principle.

This view is initially very attractive. It appears to acknowledge the deep explanatory gap with which we seem to be faced, while at the same time allowing that the apparent gap may be due to our own limitations. There are different versions of the view. Nagel (1974) has suggested that just as the Presocratics could not have understood how matter could be energy, we cannot understand how consciousness could be physical, but a conceptual revolution might allow the relevant understanding. Churchland (1997) suggests that even if we cannot now imagine how consciousness could be a physical process, that is simply a psychological limitation on our part that further progress in science will overcome. Van Gulick (1993) suggests that conceivability arguments are question-begging, since once we have a good explanation of consciousness, zombies and the like will no longer be conceivable. McGinn (1989) has suggested that the problem may be unsolvable by humans because of deep limitations in our cognitive abilities, but that it nevertheless has a solution in principle.

One way to put the view is as follows. Zombies and the like are prima facie conceivable (for us now, with our current cognitive processes), but they are not *ideally* conceivable (under idealized rational reflection). Or we could say: phenomenal truths are deducible in principle from physical truths, but the deducibility is akin to that of a complex truth of mathematics: it is accessible in principle (perhaps accessible a priori), but is not accessible to us now, perhaps because the reasoning required is currently beyond us, or perhaps because we do not currently grasp all the required physical truths. If this is so, then it will appear to us that there is a gap between physical processes and consciousness, but there will be no gap in nature.

Despite its appeal, I think that the type-C view is inherently unstable. Upon examination, it turns out either to be untenable, or to collapse into one of the other views on the table. In particular, it seems that the view must collapse into

a version of type-A materialism, type-B materialism, type-D dualism, or type-F monism, and so is not ultimately a distinct option.

One way to hold that the epistemic gap might be closed in the limit is to hold that in the limit, we will see that explaining the functions explains everything, and that there is no further explanandum. It is at least coherent to hold that we currently suffer from some sort of conceptual confusion or unclarity that leads us to believe that there is a further explanandum, and that this situation could be cleared up by better reasoning. I will count this position as a version of type-A materialism, not type-C materialism: it is obviously closely related to standard type-A materialism (the main difference is whether we have yet had the relevant insight), and the same issues arise. Like standard type-A materialism, this view ultimately stands or falls with the strength of (actual and potential) first-order arguments that dissolve any apparent further explanandum.

Once type-A materialism is set aside, the potential options for closing the epistemic gap are highly constrained. These constraints are grounded in the nature of physical concepts, and in the nature of the concept of consciousness. The basic problem has already been mentioned. First: physical descriptions of the world characterize the world in terms of structure and dynamics. Secondly: from truths about structure and dynamics, one can deduce only further truths about structure and dynamics. And thirdly: truths about consciousness are not truths about structure and dynamics. But we can take these steps one at a time.

First, a microphysical description of the world specifies a distribution of particles, fields, and waves in space and time. These basic systems are characterized by their spatio-temporal properties, and properties such as mass, charge, and quantum wave function state. These latter properties are ultimately defined in terms of spaces of states that have a certain abstract structure (e.g., the space of continuously varying real quantities, or of Hilbert space states), such that the states play a certain causal role with respect to other states. We can subsume spatio-temporal descriptions and descriptions in terms of properties in these formal spaces under the rubric of *structural* descriptions. The state of these systems can change over time in accord with dynamic principles defined over the relevant properties. The result is a description of the world in terms of its underlying spatio-temporal and formal structure, and dynamic evolution over this structure.

Some type-C materialists hold we do not yet have a complete physics, so we cannot know what such a physics might explain. But here we do not need to have a complete physics: we simply need the claim that physical descriptions are in terms of structure and dynamics. This point is general across physical theories. Such novel theories as relativity, quantum mechanics, and the like may introduce new structures, and new dynamics over those structures, but the general point (and the gap with consciousness) remains.

A type-C materialist might hold that there could be new physical theories that go beyond structure and dynamics. But given the character of physical explanation, it is unclear what sort of theory this could be. Novel physical properties are postulated for their potential in explaining existing physical phenomena, themselves

characterized in terms of structure and dynamics, and it seems that structure and dynamics always suffice here. One possibility is that instead of postulating novel properties, physics might end up appealing to consciousness itself, in the way that some theorists hold that quantum mechanics does. This possibility cannot be excluded, but it leads to a view on which consciousness is itself irreducible, and is therefore to be classed in a non-reductive category (type D or type F).

There is one appeal to a "complete physics" that should be taken seriously. This is the idea that current physics characterizes its underlying properties (such as mass and charge) in terms of abstract structures and relations, but it leaves open their intrinsic natures. On this view, a complete physical description of the world must also characterize the intrinsic properties that ground these structures and relations; and once such intrinsic properties are invoked, physics will go beyond structure and dynamics, in such a way that truths about consciousness may be entailed. The relevant intrinsic properties are unknown to us, but they are knowable in principle. This is an important position, but it is precisely the position discussed under type F, so I defer discussion of it until then.

Secondly, what can be inferred from this sort of description in terms of structure and dynamics? A low-level microphysical description can entail all sorts of surprising and interesting macroscopic properties, as with the emergence of chemistry from physics, of biology from chemistry, or more generally of complex emergent behaviors in complex systems theory. But in all these cases, the complex properties that are entailed are nevertheless structural and dynamic: they describe complex spatio-temporal structures and complex dynamic patterns of behavior over those structures. So these cases support the general principle that, from structure and dynamics, one can infer only structure and dynamics.

A type-C materialist might suggest there are some truths that are not themselves structural-dynamical that are nevertheless implied by a structural-dynamical description. It might be argued, perhaps, that truths about *representation* or *belief* have this character. But as we saw earlier, it seems clear that any sense in which these truths are implied by a structural-dynamic description involves a tacitly functional sense of representation or of belief. This is what we would expect: if claims involving these can be seen (on conceptual grounds) to be true *in virtue* of a structural-dynamic descriptions holding, the notions involved must themselves be structural-dynamic, at some level.

One might hold that there is some intermediate notion X, such that truths about X hold in virtue of structural-dynamic descriptions, and truths about consciousness hold in virtue of X. But as in the case of type-A materialism, either X is functionally analyzable (in the broad sense), in which case the second step fails, or X is not functionally analyzable, in which case the first step fails. This is brought out clearly in the case of representation: for the notion of functional representation, the first step fails, and for the notion of phenomenal representation, the second step fails. So this sort of strategy can only work by equivocation.

Thirdly, does explaining or deducing complex structure and dynamics suffice to explain or deduce consciousness? It seems clearly not, for the usual reasons. Mary

could know from her black-and-white room all about the spatio-temporal structure and dynamics of the world at all levels, but this will not tell her what it is like to see red. For any complex macroscopic structural or dynamic description of a system, one can conceive of that description being instantiated without consciousness. And explaining structure and dynamics of a human system is only to solve the easy problems, while leaving the hard problems untouched. To resist this last step, an opponent would have to hold that explaining structure and dynamics *thereby* suffices to explain consciousness. The only remotely tenable way to do this would be to embrace type-A materialism, which we have set aside.

A type-C materialist might suggest that instead of leaning on dynamics (as a type-A materialist does), one could lean on structure. Here, spatio-temporal structure seems very unpromising: to explain a system's size, shape, position, motion, and so on is clearly not to explain consciousness. A final possibility is leaning on the structure present in conscious states themselves. Conscious states have structure: there is both internal structure within a single complex conscious state, and there are patterns of similarities and differences between conscious states. But this structure is a distinctively *phenomenal* structure, quite different in kind from the spatio-temporal and formal structure present in physics. The structure of a complex phenomenal state is not spatio-temporal structure (although it may involve the representation of spatio-temporal structure), and the similarities and differences between phenomenal states are not formal similarities and differences, but differences between specific phenomenal characters. This is reflected in the fact that one can conceive of any spatio-temporal structure and formal structure without any associated phenomenal structure; one can know about the first without knowing about the second; and so on. So the epistemic gap is as wide as ever.

The basic problem with any type-C materialist strategy is that epistemic implication from A to B requires some sort of *conceptual hook* by virtue of which the condition described in A can satisfy the conceptual requirements for the truth of B. When a physical account implies truths about life, for example, it does so in virtue of implying information about the macroscopic functioning of physical systems, of the sort required for life: here, broadly functional notions provide the conceptual hook. But in the case of consciousness, no such conceptual hook is available, given the structural-dynamic character of physical concepts, and the quite different character of the concept of consciousness.

Ultimately, it seems that any type-C strategy is doomed for familiar reasons. Once we accept that the concept of consciousness is not itself a functional concept, and that physical descriptions of the world are structural-dynamic descriptions, there is simply no conceptual room for it to be implied by a physical description. So the only room left is to hold that consciousness is a broadly functional concept after all (accepting type-A materialism), to hold that there is more in physics than structure and dynamics (accepting type-D dualism or type-F monism), or to hold that the truth of materialism does not require an implication from physics to consciousness (accepting type-B materialism).[22] So in the end, there is no separate space for the type-C materialist.

5.8 Interlude

Are there any other options for the materialist? One further option is to reject the distinctions on which this taxonomy rests. For example, some philosophers, especially followers of Quine (1951), reject any distinction between conceptual truth and empirical truth, or between the a priori and the a posteriori, or between the contingent and the necessary. One who is sufficiently Quinean might therefore reject the distinction between type-A and type-B materialism, holding that talk of epistemic implication and/or modal entailment is ungrounded, but that materialism is true nevertheless. We might call such a view type-Q materialism. Still, even on this view, similar issues arise. Some Quineans hold that explaining the functions explains everything (Dennett may be an example); if so, all the problems of type-A materialism arise. Others hold that we can postulate identities between physical states and conscious states in virtue of the strong isomorphic connections between them in nature (Paul Churchland may be an example); if so, the problems of type-B materialism arise. Others may appeal to novel future sorts of explanation; if so, the problems of type-C materialism arise. So the Quinean approach cannot avoid the relevant problems.

Leaving this sort of view aside, it looks like the only remotely viable options for the materialist are type-A materialism and type-B materialism. I think that other views are either ultimately unstable, or collapse into one of these (or the three remaining options).[23] It seems to me that the costs of these views – denying the manifest explanandum in the first case, and embracing primitive identities or strong necessities in the second case – suggest very strongly that they are to be avoided unless there are no viable alternatives.

So the residual question is whether there are viable alternatives. If consciousness is not necessitated by physical truths, then it must involve something ontologically novel in the world: to use Kripke's metaphor, after fixing all the physical truths, God had to do more work to fix all the truths about consciousness. That is, there must be ontologically fundamental features of the world over and above the features characterized by physical theory. We are used to the idea that some features of the world are fundamental: in physics, features such as spacetime, mass, and charge are taken as fundamental and not further explained. If the arguments against materialism are correct, these features from physics do not exhaust the fundamental features of the world: we need to expand our catalog of the world's basic features.

There are two possibilities here. First, it could be that consciousness is itself a fundamental feature of the world, like spacetime and mass. In this case, we can say that phenomenal properties are fundamental. Secondly, it could be that consciousness is not itself fundamental, but is necessitated by some more primitive fundamental feature X that is not itself necessitated by physics. In this case, we might call X a *protophenomenal* property, and we can say that protophenomenal properties are fundamental. I will typically put things in terms of the first possibility

for ease of discussion, but the discussion that follows applies equally to the second. Either way, consciousness involves something novel and fundamental in the world.

The question then arises: how do these novel fundamental properties relate to the already acknowledged fundamental properties of the world, namely those invoked in microphysics? In general, where there are fundamental properties, there are fundamental laws. So we can expect that there will be some sort of fundamental principles – psychophysical laws – connecting physical and phenomenal properties. Like the fundamental laws of relativity or quantum mechanics, these psychophysical laws will not be deducible from more basic principles, but instead will be taken as primitive.

But what is the character of these laws? An immediate worry is that the microphysical aspects of the world are often held to be causally closed, in that every microphysical state has a microphysical sufficient cause. How are fundamental phenomenal properties to be integrated with this causally closed network?

There seem to be three main options for the non-reductionist here. First, one could deny the causal closure of the microphysical, holding that there are causal gaps in microphysical dynamics that are filled by a causal role for distinct phenomenal properties: this is type-D dualism. Secondly, one could accept the causal closure of the microphysical and hold that phenomenal properties play no causal role with respect to the physical network: this is type-E dualism. Thirdly, one could accept that the microphysical network is causally closed, but hold that phenomenal properties are nevertheless integrated with it and play a causal role, by virtue of constituting the intrinsic nature of the physical: this is type-F monism.

In what follows, I will discuss each of these views. The discussion is necessarily speculative in certain respects, and I do not claim to establish that any one of the views is true or completely unproblematic. But I do aim to suggest that none of them has obvious fatal flaws, and that each deserves further investigation.

5.9 Type-D Dualism

Type-D dualism holds that microphysics is not causally closed, and that phenomenal properties play a causal role in affecting the physical world.[24] On this view, usually known as *interactionism*, physical states will cause phenomenal states, and phenomenal states cause physical states. The corresponding psychophysical laws will run in both directions. On this view, the evolution of microphysical states will not be determined by physical principles alone. Psychophysical principles specifying the effect of phenomenal states on physical states will also play an irreducible role.

The most familiar version of this sort of view is Descartes's substance dualism (hence D for Descartes), on which there are separate interacting mental and physical substances or entities. But this sort of view is also compatible with a

property dualism, on which there is just one sort of substance or entity with both physical and phenomenal fundamental properties, such that the phenomenal properties play an irreducible role in affecting the physical properties. In particular, the view is compatible with an "emergentist" view such as Broad's, on which phenomenal properties are ontologically novel properties of physical systems (not deducible from microphysical properties alone), and have novel effects on microphysical properties (not deducible from microphysical principles alone). Such a view would involve basic principles of "downward" causation of the mental on the microphysical (hence also D for downward causation).

It is sometimes objected that distinct physical and mental states could not interact, since there is no causal nexus between them. But one lesson from Hume and from modern science is that the same goes for any fundamental causal interactions, including those found in physics. Newtonian science reveals no causal nexus by which gravitation works, for example; rather, the relevant laws are simply fundamental. The same goes for basic laws in other physical theories. And the same, presumably, applies to fundamental psychophysical laws: there is no need for a causal nexus distinct from the physical and mental properties themselves.

By far the most influential objection to interactionism is that it is incompatible with physics. It is widely held that science tells us that the microphysical realm is causally closed, so that there is no room for mental states to have any effects. An interactionist might respond in various ways. For example, it could be suggested that although no experimental studies have revealed these effects, none has ruled them out. It might further be suggested that physical theory allows any number of basic *forces* (four as things stand, but there is always room for more), and that an extra force associated with a mental field would be a reasonable extension of existing physical theory. These suggestions would invoke significant revisions to physical theory, so are not to be made lightly; but one could argue that nothing rules them out.

By far the strongest response to this objection, however, is to suggest that far from ruling out interactionism, contemporary physics is positively encouraging to the possibility. On the standard formulation of quantum mechanics, the state of the world is described by a wave function, according to which physical entities are often in a superposed state (e.g., in a superposition of two different positions), even though superpositions are never directly observed. On the standard dynamics, the wave function can evolve in two ways: linear evolution by the Schrödinger equation (which tends to produce superposed states), and non-linear *collapses* from superposed states into non-superposed states. Schrödinger evolution is deterministic, but collapse is non-deterministic. Schrödinger evolution is constantly ongoing, but on the standard formulation, collapses occur only occasionally, on measurement.

The collapse dynamics leaves a door wide open for an interactionist interpretation. Any physical non-determinism might be held to leave room for non-physical effects, but the principles of collapse do much more than that. Collapse is supposed to

occur on measurement. There is no widely agreed definition of what a measurement is, but there is one sort of event that everyone agrees is a measurement: observation by a conscious observer. Further, it seems that no purely physical criterion for a measurement can work, since purely physical systems are governed by the linear Schrödinger dynamics. As such, it is natural to suggest that a measurement is precisely a conscious observation, and that this conscious observation causes a collapse.

The claim should not be too strong: quantum mechanics does not force this interpretation of the situation onto us, and there are alternative interpretations of quantum mechanics on which there are no collapses, or on which measurement has no special role in collapse.[25] Nevertheless, quantum mechanics appears to be perfectly *compatible* with such an interpretation. In fact, one might argue that if one were to design elegant laws of physics that allow a role for the conscious mind, one could not do much better than the bipartite dynamics of standard quantum mechanics: one principle governing deterministic evolution in normal cases, and one principle governing non-deterministic evolution in special situations that have a prima facie link to the mental.

Of course such an interpretation of quantum mechanics is controversial. Many physicists reject it precisely because it is dualistic, giving a fundamental role to consciousness. This rejection is not surprising, but it carries no force when we have independent reason to hold that consciousness may be fundamental. There is some irony in the fact that philosophers reject interactionism on largely physical grounds[26] (it is incompatible with physical theory), while physicists reject an interactionist interpretation of quantum mechanics on largely philosophical grounds (it is dualistic). Taken conjointly, these reasons carry little force, especially in light of the arguments against materialism elsewhere in this chapter.

This sort of interpretation needs to be formulated in detail to be assessed.[27] I think the most promising version of such an interpretation allows conscious states to be correlated with the total quantum state of a system, with the extra constraint that conscious states (unlike physical states) can never be superposed. In a conscious physical system such as a brain, the physical and phenomenal states of the system will be correlated in a (non-superposed) quantum state. Upon observation of a superposed system, then Schrödinger evolution at the moment of observation would cause the observed system to become correlated with the brain, yielding a resulting superposition of brain states and so (by psychophysical correlation) a superposition of conscious states. But such a superposition cannot occur, so one of the potential resulting conscious states is somehow selected (presumably by a non-deterministic dynamic principle at the phenomenal level). The result is that (by psychophysical correlation) a definite brain state and a definite state of the observed object are also selected. The same might apply to the connection between consciousness and non-conscious processes in the brain: when superposed non-conscious processes threaten to affect consciousness, there will be some sort of selection. In this way, there is a causal role for consciousness in the physical world.

(Interestingly, such a theory may be empirically testable. In quantum mechanics, collapse theories yield predictions slightly different from no-collapse theories, and different hypotheses about the location of collapse yield predictions that differ from each other, although the differences are extremely subtle and are currently impossible to measure. If the relevant experiments can one day be performed, some outcomes would give us strong reason to accept a collapse theory, and might in turn give us grounds to accept a role for consciousness. As a bonus, this could even yield an empirical criterion for the presence of consciousness.)

There are any number of further questions concerning the precise formulation of such a view, its compatibility with physical theory more generally (e.g., relativity and quantum field theory), and its philosophical tenability (e.g., does this view yield the sort of causal role that we are inclined to think consciousness must have). But at the very least, it cannot be said that physical theory immediately rules out the possibility of an interactionist theory. Those who make this claim often raise their eyebrows when a specific theory such as quantum mechanics is mentioned; but this is quite clearly an inconsistent set of attitudes. If physics is supposed to rule out interactionism, then careful attention to the detail of physical theory is required.

All this suggests that there is at least room for a viable interactionism to be explored, and that the most common objection to interactionism has little force. Of course it does not entail that interactionism is true. There is much that is attractive about the view of the physical world as causally closed, and there is little direct evidence from cognitive science of the hypothesis that behavior cannot be wholly explained in terms of physical causes. Still, if we have independent reason to think that consciousness is irreducible, and if we wish to retain the intuitive view that consciousness plays a causal role, then this is a view to be taken very seriously.

5.10 Type-E Dualism

Type-E dualism holds that phenomenal properties are ontologically distinct from physical properties, and that the phenomenal has no effect on the physical.[28] This is the view usually known as *epiphenomenalism* (hence type-E): physical states cause phenomenal states, but not vice versa. On this view, psychophysical laws run in one direction only, from physical to phenomenal. The view is naturally combined with the view that the physical realm is causally closed: this further claim is not essential to type-E dualism, but it provides much of the motivation for the view.

As with type-D dualism, type-E dualism is compatible with a substance dualism with distinct physical and mental substances or entities, and is also compatible with a property dualism with one sort of substance or entity and two sorts of property. Again, it is compatible with an emergentism such as Broad's, on which mental properties are ontologically novel emergent properties of an underlying entity, but in this case although there are emergent qualities, there is no emergent downward causation.

Type-E dualism is usually put forward as respecting both consciousness and science: it simultaneously accommodates the anti-materialist arguments about consciousness and the causal closure of the physical. At the same time, type-E dualism is frequently rejected as deeply counterintuitive. If type-E dualism is correct, then phenomenal states have no effect on our actions, physically construed. For example, a sensation of pain will play no causal role in my hand's moving away from a flame; my experience of decision will play no causal role in my moving to a new country; and a sensation of red will play no causal role in my producing the utterance "I am experiencing red now." These consequences are often held to be obviously false, or at least unacceptable.

Still, the type-E dualist can reply that there is no direct *evidence* that contradicts their view. Our evidence reveals only regular connections between phenomenal states and actions, so that certain sorts of experience are typically followed by certain sorts of action. Being exposed to this sort of constant conjunction produces a strong *belief* in a causal connection (as Hume pointed out in another context); but it is nevertheless compatible with the absence of a causal connection. Indeed, it seems that if epiphenomenalism *were* true, we would have exactly the same evidence, and be led to believe that consciousness has a causal role for much the same reasons. So if epiphenomenalism is otherwise coherent and acceptable, it seems that these considerations do not provide strong reasons to reject it.[29]

Another objection holds that if consciousness is epiphenomenal, it could not have evolved by natural selection. The type-E dualist has a straightforward reply, however. On the type-E view, there are fundamental psychophysical laws associating physical and phenomenal properties. If evolution selects appropriate physical properties (perhaps involving physical or informational configurations in the brain), then the psychophysical laws will ensure that phenomenal properties are instantiated, too. If the laws have the right form, one can even expect that, as more complex physical systems are selected, more complex states of consciousness will evolve. In this way, physical evolution will carry the evolution of consciousness along with it as a sort of by-product.

Perhaps the most interesting objections to epiphenomenalism focus on the relation between consciousness and representations of consciousness. It is certainly at least strange to suggest that consciousness plays no causal role in my utterances of "I am conscious." Some have suggested more strongly that this rules out any *knowledge* of consciousness. It is often held that if a belief about X is to qualify as knowledge, the belief must be caused in some fashion by X. But if consciousness does not affect physical states, and if beliefs are physically constituted, then consciousness cannot cause beliefs. And even if beliefs are not physically constituted, it is not clear how epiphenomenalism can accommodate a causal connection between consciousness and belief.

In response, an epiphenomenalist can deny that knowledge always requires a causal connection. One can argue on independent grounds that there is a stronger connection between consciousness and beliefs about consciousness: consciousness plays a role in *constituting* phenomenal concepts and phenomenal beliefs. A red

experience plays a role in constituting a belief that one is having a red experience, for example. If so, there is no causal distance between the experience and the belief. And one can argue that this immediate connection to experience and belief allows for the belief to be justified. If this is right, then epiphenomenalism poses no obstacle to knowledge of consciousness.

A related objection holds that my zombie twin would produce the same reports (e.g., "I am conscious"), caused by the same mechanisms, and that his reports are unjustified; if so, my own reports are unjustified. In response, one can hold that the true bearers of justification are beliefs, and that my zombie twin and I have *different* beliefs, involving different concepts, because of the role that consciousness plays in constituting my concepts but not the zombie's. Further, the fact that we produce isomorphic reports implies that a third-person observer might not be any more justified in believing that I am conscious than that the zombie is conscious, but it does not imply a difference in first-person justification. The first-person justification for my belief that I am conscious is not grounded in any way in my reports but rather in my experiences themselves, experiences that the zombie lacks.

I think that there is no knock-down objection to epiphenomenalism here. Still, it must be acknowledged that the situation is at least odd and counterintuitive. The oddness of epiphenomenalism is exacerbated by the fact that the relationship between consciousness and reports about consciousness seems to be something of a lucky coincidence, on the epiphenomenalist view. After all, if psychophysical laws are independent of physical evolution, then there will be possible worlds where physical evolution is the same as ours but the psychophysical laws are very different, so that there is a radical mismatch between reports and experiences. It seems lucky that we are in a world whose psychophysical laws match them up so well. In response, an epiphenomenalist might try to make the case that these laws are somehow the most "natural" and are to be expected; but there is at least a significant burden of proof here.

Overall, I think that epiphenomenalism is a coherent view without fatal problems. At the same time, it is an inelegant view, producing a fragmented picture of nature, on which physical and phenomenal properties are only very weakly integrated in the natural world. And of course it is a counterintuitive view that many people find difficult to accept. Inelegance and counterintuitiveness are better than incoherence; so if good arguments force us to epiphenomenalism as the most coherent view, then we should take it seriously. But at the same time, we have good reason to examine other views very carefully.

5.11 Type-F Monism

Type-F monism is the view that consciousness is constituted by the intrinsic properties of fundamental physical entities: that is, by the categorical bases of fundamental physical dispositions.[30] On this view, phenomenal or protophenomenal

properties are located at the fundamental level of physical reality, and, in a certain sense, underlie physical reality itself.

This view takes its cue from Bertrand Russell's discussion of physics in *The Analysis of Matter* (1927). Russell pointed out that physics characterizes physical entities and properties by their relations to one another and to us. For example, a quark is characterized by its relations to other physical entities, and a property such as mass is characterized by an associated dispositional role, such as the tendency to resist acceleration. At the same time, physics says nothing about the intrinsic nature of these entities and properties. Where we have relations and dispositions, we expect some underlying intrinsic properties that ground the dispositions, characterizing the entities that stand in these relations.[31] But physics is silent about the intrinsic nature of a quark, or about the intrinsic properties that play the role associated with mass. So this is one metaphysical problem: what are the intrinsic properties of fundamental physical systems?

At the same time, there is another metaphysical problem: how can phenomenal properties be integrated with the physical world? Phenomenal properties seem to be intrinsic properties that are hard to fit in with the structural/dynamic character of physical theory; and arguably, they are the only intrinsic properties of which we have direct knowledge. Russell's insight was that we might solve both these problems at once. Perhaps the intrinsic properties of the physical world are themselves phenomenal properties. Or perhaps the intrinsic properties of the physical world are not phenomenal properties, but nevertheless constitute phenomenal properties: that is, perhaps they are protophenomenal properties. If so, then consciousness and physical reality are deeply intertwined.

This view holds the promise of integrating phenomenal and physical properties very tightly in the natural world. Here, nature consists of entities with intrinsic (proto)phenomenal qualities standing in causal relations within a spacetime manifold. Physics as we know it emerges from the relations between these entities, whereas consciousness as we know it emerges from their intrinsic nature. As a bonus, this view is perfectly compatible with the causal closure of the microphysical, and indeed with existing physical laws. The view can retain the *structure* of physical theory as it already exists; it simply supplements this structure with an intrinsic nature. And the view acknowledges a clear causal role for consciousness in the physical world: (proto)phenomenal properties serve as the ultimate categorical basis of all physical causation.

This view has elements in common with both materialism and dualism. From one perspective, it can be seen as a sort of materialism. If one holds that physical terms refer not to dispositional properties but the underlying intrinsic properties, then the protophenomenal properties can be seen as physical properties, thus preserving a sort of materialism. From another perspective, it can be seen as a sort of dualism. The view acknowledges phenomenal or protophenomenal properties as ontologically fundamental, and it retains an underlying duality between structural-dispositional properties (those directly characterized in physical theory) and intrinsic protophenomenal properties (those responsible for consciousness).

One might suggest that while the view arguably fits the letter of materialism, it shares the spirit of anti-materialism.

In its protophenomenal form, the view can be seen as a sort of neutral monism: there are underlying neutral properties X (the protophenomenal properties), such that the X properties are simultaneously responsible for constituting the physical domain (by their relations) and the phenomenal domain (by their collective intrinsic nature). In its phenomenal form, it can be seen as a sort of idealism, such that mental properties constitute physical properties, although these need not be mental properties in the mind of an observer, and they may need to be supplemented by causal and spatio-temporal properties in addition. One could also characterize this form of the view as a sort of panpsychism, with phenomenal properties ubiquitous at the fundamental level. One could give the view in its most general form the name *panprotopsychism*, with either protophenomenal or phenomenal properties underlying all of physical reality.

A type-F monist may have one of a number of attitudes to the zombie argument against materialism. Some type-F monists may hold that a complete physical description must be expanded to include an intrinsic description, and may consequently deny that zombies are conceivable. (We only think we are conceiving of a physically identical system because we overlook intrinsic properties.) Others could maintain that existing physical concepts refer via dispositions to those intrinsic properties that ground the dispositions. If so, these concepts have different primary and secondary intensions, and a type-F monist could correspondingly accept conceivability but deny possibility: we misdescribe the conceived world as physically identical to ours, when in fact it is just structurally identical.[32] Finally, a type-F monist might hold that physical concepts refer to dispositional properties, so that zombies are both conceivable and possible, and the intrinsic properties are not physical properties. The differences between these three attitudes seem to be ultimately terminological rather than substantive.

As for the knowledge argument, a type-F monist might insist that for Mary to have complete physical knowledge, she would have to have a description of the world involving concepts that directly characterize the intrinsic properties; if she had this (as opposed to her impoverished description involving dispositional concepts), she might thereby be in a position to know what it is like to see red. Regarding the explanatory argument, a type-F monist might hold that physical accounts involving intrinsic properties can explain more than structure and function. Alternatively, a type-F monist who sticks to dispositional physical concepts will make responses analogous to one of the other two responses above.

The type-F view is admittedly speculative, and it can sound strange at first hearing. Many find it extremely counterintuitive to suppose that fundamental physical systems have phenomenal properties: e.g., that there is something it is like to be an electron. The protophenomenal version of the view rejects this claim, but retains something of its strangeness: it seems that any properties responsible for constituting consciousness must be strange and unusual properties, of a sort that we might not expect to find in microphysical reality. Still, it is not

clear that this strangeness yields any strong objections. Like epiphenomenalism, the view appears to be compatible with all our evidence, and there is no direct evidence against it. One can argue that if the view were true, things would appear to us just as they in fact appear. And we have learned from modern physics that the world is a strange place: we cannot expect it to obey all the dictates of common sense.

One might also object that we do not have any conception of what protophenomenal properties might be like, or of how they could constitute phenomenal properties. This is true, but one could suggest that this is merely a product of our ignorance. In the case of familiar physical properties, there were principled reasons (based on the character of physical concepts) for denying a constitutive connection to phenomenal properties. Here, there are no such principled reasons. At most, there is ignorance and absence of a connection. Of course it would be very desirable to form a positive conception of protophenomenal properties. Perhaps we can do this indirectly, by some sort of theoretical inference from the character of phenomenal properties to their underlying constituents; or perhaps knowledge of the nature of protophenomenal properties will remain beyond us. Either way, this is no reason to reject the truth of the view.[33]

There is one sort of principled problem in the vicinity, pointed out by William James (1890: ch. 6). Our phenomenology has a rich and specific structure: it is unified, bounded, differentiated into many different aspects, but with an underlying homogeneity to many of the aspects, and appears to have a single subject of experience. It is not easy to see how a distribution of a large number of individual microphysical systems, each with their own protophenomenal properties, could somehow add up to this rich and specific structure. Should one not expect something more like a disunified, jagged collection of phenomenal spikes?

This is a version of the *combination problem* for panpsychism (Seagar 1995), or what Stoljar (2001) calls the *structural mismatch* problem for the Russellian view (see also Foster 1991: 119–30). To answer it, it seems that we need a much better understanding of the *compositional* principles of phenomenology: that is, the principles by which phenomenal properties can be composed or constituted from underlying phenomenal properties, or protophenomenal properties. We have a good understanding of the principles of physical composition, but no real understanding of the principles of phenomenal composition. This is an area that deserves much close attention: I think it is easily the most serious problem for the type-F monist view. At this point, it is an open question whether or not the problem can be solved.

Some type-F monists appear to hold that they can avoid the combination problem by holding that phenomenal properties are the intrinsic properties of *high-level* physical dispositions (e.g., those involved in neural states), and need not be constituted by the intrinsic properties of microphysical states (hence they may also deny panprotopsychism). But this seems to be untenable: if the low-level network is causally closed and the high-level intrinsic properties are not

constituted by low-level intrinsic properties, the high-level intrinsic properties will be epiphenomenal all over again, for familiar reasons. The only way to embrace this position would seem to be in combination with a denial of microphysical causal closure, holding that there are fundamental dispositions above the microphysical level, which have phenomenal properties as their grounds. But such a view would be indistinguishable from type-D dualism.[34] So a distinctive type-F monism will have to face the combination problem directly.

Overall, type-F monism promises a deeply integrated and elegant view of nature. No one has yet developed any sort of detailed theory in this class, and it is not yet clear whether such a theory can be developed. But at the same time, there appear to be no strong reasons to reject the view. As such, type-F monism is likely to provide fertile grounds for further investigation, and it may ultimately provide the best integration of the physical and the phenomenal within the natural world.

5.12 Conclusions

Are there any other options for the non-reductionist? There are two views that may not fit straightforwardly into the categories above.

First, some non-materialists hold that phenomenal properties are ontologically wholly distinct from physical properties, that microphysics is causally closed, but that phenomenal properties play a causal role with respect to the physical nevertheless. One way this might happen is by a sort of causal overdetermination: physical states causally determine behavior, but phenomenal states cause behavior at the same time. Another is by causal mediation: it might be that in at least some instances of microphysical causation from A to B, there is actually a causal connection from A to the mind to B, so that the mind enters the causal nexus without altering the structure of the network. And there may be further strategies here. We might call this class type-O dualism (taking overdetermination as a paradigm case). These views share much of the structure of the type-E view (causally closed physical world, distinct phenomenal properties), but escapes the charge of epiphenomenalism. The special causal setups of these views may be hard to swallow, and they share some of the same problems as the type-E view (e.g., the fragmented view of nature, and the "lucky" psychophysical laws), but this class should nevertheless be put on the table as an option.[35]

Second, some non-materialists are *idealists* (in a Berkeleyan sense), holding that the physical world is itself constituted by the conscious states of an observing agent. We might call this view type-I monism. It shares with type-F monism the property that phenomenal states play a role in constituting physical reality, but on the type-I view this happens in a very different way: not by having separate "microscopic" phenomenal states underlying each physical state, but rather by having physical states constituted holistically by a "macroscopic" phenomenal

mind. This view seems to be non-naturalistic in a much deeper sense than any of the views above, and in particular seems to suffer from an absence of causal or explanatory closure in nature: once the natural explanation in terms of the external world is removed, highly complex regularities among phenomenal states have to be taken as unexplained in terms of simpler principles. But again, this sort of view should at least be acknowledged.

As I see things, the best options for a non-reductionist are type-D dualism, type-E dualism, or type-F monism: that is, interactionism, epiphenomenalism, or panprotopsychism. If we acknowledge the epistemic gap between the physical and the phenomenal, and we rule out primitive identities and strong necessities, then we are led to a disjunction of these three views. Each of the views has at least some promise, and none has clear fatal flaws. For my part, I give some credence to each of them. I think that in some ways the type-F view is the most appealing, but this sense is largely grounded in aesthetic considerations whose force is unclear.

The choice between these three views may depend in large part on the development of specific theories within these frameworks. Especially for the type-D view and type-F view, further theoretical work is crucial in assessing the theories (e.g., in explicating quantum interactionism, or in understanding phenomenal composition). It may also be that the empirical science of consciousness will give some guidance. As the science progresses, we will be led to infer simple principles that underlie correlations between physical and phenomenal states. It may be that these principles turn out to point strongly toward one or the other of these views: e.g., if simple principles connecting microphysical states to phenomenal or protophenomenal states can do the explanatory work, then we may have reason to favor a type-F view, while if the principles latch onto the physical world at a higher level, then we may have reason to favor a type-D or type-E view. And if consciousness has a specific pattern of effects on the physical world, as the type-D view suggests, then empirical studies ought in principle to be able to find these effects, although perhaps only with great difficulty.

Not everyone will agree that each of these views is viable. It may be that further examination will reveal deep problems with some of these views. But this further examination needs to be performed. There has been little critical examination of type-F views to date, for example; we have seen that the standard arguments against type-D views carry very little weight; and while arguments against type-E views carry some intuitive force, they are far from making a knock-down case against the views. I suspect that even if further examination reveals deep problems for some views in this vicinity, it is very unlikely that all such views will be eliminated.

In any case, this gives us some perspective on the mind–body problem. It is often held that even though it is hard to see how materialism could be true, materialism *must* be true, since the alternatives are unacceptable. As I see it, there are at least three prima facie acceptable alternatives to materialism on the table, each of which is compatible with a broadly naturalistic (even if not materialistic)

worldview, and none of which has fatal problems. So given the clear arguments against materialism, it seems to me that we should at least tentatively embrace the conclusion that one of these views is correct. Of course all of the views discussed in this chapter need to be developed in much more detail, and examined in light of all relevant scientific and philosophical developments, in order to be comprehensively assessed. But as things stand, I think that we have good reason to suppose that consciousness has a fundamental place in nature.

Notes

1 This chapter is an overview of issues concerning the metaphysics of consciousness. Much of the discussion in this chapter (especially the first part) recapitulates discussion in Chalmers (1995; 1996; 1997), although it often takes a different form, and sometimes goes beyond the discussion there. I give a more detailed treatment of many of the issues discussed here in the works cited in the bibliography.

2 The taxonomy is in the final chapter, chapter 14, of Broad's book (set out on pp. 607–11, and discussed until p. 650). The dramatization of Broad's taxonomy as a 4 × 4 matrix is illustrated on Andrew Chrucky's website devoted to Broad, at http://www.ditext.com/broad/mpn14.html#t.

3 On my usage, qualia are simply those properties that characterize conscious states according to what it is like to have them. The definition does not build in any further substantive requirements, such as the requirement that qualia are intrinsic or non-intentional. If qualia are intrinsic or non-intentional, this will be a substantive rather than a definitional point (so the claim that the properties of consciousness are non-intrinsic or that they are wholly intentional should not be taken to entail that there are no qualia). Phenomenal properties can also be taken to be properties of individuals (e.g., people) rather than of mental states, characterizing aspects of what it is like to be them at a given time; the difference will not matter much for present purposes.

4 Note that I use "reductive" in a broader sense than it is sometimes used. Reductive explanation requires only that high-level phenomena can be explained wholly in terms of low-level phenomena. This is compatible with the "multiple realizability" of high-level phenomena in low-level phenomena. For example, there may be many different ways in which digestion could be realized in a physiological system, but one can nevertheless reductively explain a system's digestion in terms of underlying physiology. Another subtlety concerns the possibility of a view on which consciousness can be explained in terms of principles which do not make appeal to consciousness but cannot themselves be physically explained. The definitions above count such a view as neither reductive nor non-reductive. It could reasonably be classified either way, but I will generally assimilate it with the non-reductive class.

5 A version of the explanatory argument as formulated here is given in Chalmers (1995). For related considerations about explanation, see Levine (1983) on the "explanatory gap" and Nagel (1974). See also the papers in Shear (1997).

6 Versions of the conceivability argument are put forward by Campbell (1970), Kirk (1974), Kripke (1980), Bealer (1994), and Chalmers (1996), among others. Important predecessors include Descartes's conceivability argument about disembodiment, and Leibniz's "mill" argument.

7 Sources for the knowledge argument include Nagel (1974), Maxwell (1968), Jackson (1982), and others. Predecessors of the argument are present in Broad's discussion of a "mathematical archangel" who cannot deduce the smell of ammonia from physical facts (1925: 70–1), and Feigl's discussion of a "Martian superscientist" who cannot know what colors look like and what musical tones sound like (1967[1958]: 64, 68, 140).

8 This version of the thought experiment has a real life exemplar in Knut Nordby, a Norwegian sensory biologist who is a rod monochromat (lacking cones in his retina for color vision), and who works on the physiology of color vision. See Nordby (1990).

9 For limited versions of the conceivability argument and the explanatory argument, see Broad (1925: 614–15). For the knowledge argument, see pp. 70–2, where Broad argues that even a "mathematical archangel" could not deduce the smell of ammonia from microscopic knowledge of atoms. Broad is arguing against "mechanism," which is roughly equivalent to contemporary materialism. Perhaps the biggest lacuna in Broad's argument, to contemporary eyes, is any consideration of the possibility that there is an epistemic but not an ontological gap.

10 For a discussion of the relationship between the conceivability argument and the knowledge argument, see Chalmers (1996 and 2002b).

11 Type-A materialists include Ryle (1949), Lewis (1988), Dennett (1991), Dretske (1995), Rey (1995), and Harman (1990).

12 Two specific views may be worth mentioning: (1) Some views (e.g., Dretske 1995) deny an epistemic gap while at the same time denying functionalism, by holding that consciousness involves not just functional role but also causal and historical relations to objects in the environment. I count these as type-A views: we can view the relevant relations as part of functional role, broadly construed, and exactly the same considerations arise. (2) Some views (e.g., Strawson 2000 and Stoljar 2001) deny an epistemic gap not by functionally analyzing consciousness but by expanding our view of the physical base to include underlying intrinsic properties. These views are discussed under type-F (sectn 5.11).

13 In another analogy, Churchland (1996) suggests that someone in Goethe's time might have mounted analogous epistemic arguments against the reductive explanation of "luminescence." But on a close look, it is not hard to see that the only further explanandum that could have caused doubts here is the *experience* of seeing light (see Chalmers 1997). This point is no help to the type-A materialist, since this explanandum remains unexplained.

14 For an argument from unsavory metaphysical consequences, see White (1986). For an argument from unsavory epistemological consequences, see Shoemaker (1975). The metaphysical consequences are addressed in the second half of this chapter. The epistemological consequences are addressed in Chalmers 2002a.

15 Type-B materialists include Levine (1983), Loar (1990/1997), Papineau (1993), Tye (1995), Lycan (1996), Hill (1997), Block and Stalnaker (1999), and Perry (2001).

16 In certain respects, where type-A materialism can be seen as deriving from the logical behaviorism of Ryle and Carnap, type-B materialism can be seen as deriving from the identity theory of Place and Smart. The matter is complicated, however, by the fact that the early identity theorists advocated "topic-neutral" (functional) analyses of phenomenal properties, suggesting an underlying type-A materialism.

17 Block and Stalnaker (1999) argue against deducibility in part by arguing that there is usually no explicit conceptual analysis of high-level terms such as "water" in microphysical terms, or in any other terms that could ground an a priori entailment from microphysical truths to truths about water. In response, Chalmers and Jackson (2001) argue that explicit conceptual analyses are not required for a priori entailments, and that there is good reason to believe that such entailments exist in these cases.

18 Two-dimensional semantic frameworks originate in the work of Stalnaker (1978), Evans (1979), and Kaplan (1989). The version used in these arguments is somewhat different: for discussion of the differences, see Chalmers (forthcoming).

19 This is a slightly more formal version of an argument in Chalmers (1996: 131–6). It is quite closely related to Kripke's modal argument against the identity theory, though different in some important respects. The central premise 2 can be seen as a way of formalizing Kripke's claim that where there is "apparent contingency," there is some misdescribed possibility in the background. The argument can also be seen as a way of formalizing a version of the "dual property" objection attributed to Max Black by Smart (1959), and developed by Jackson (1979) and White (1986). Related applications of the two-dimensional framework to questions about materialism are given by Jackson (1994) and Lewis (1994).

20 I have passed over a few subtleties here. One concerns the role of indexicals: to handle claims such as "I am here," primary intensions are defined over *centered worlds*: worlds with a marked individual and time, corresponding to indexical "locating information" about one's position in the world. This change does not help the type-B materialist, however. Even if we supplement P with indexical locating information I (e.g., telling Mary about her location in the world), there is as much of an epistemic gap with Q as ever; so $P \wedge I \wedge \neg Q$ is conceivable. And given that there is a centered world that verifies $P \wedge I \wedge \neg Q$, one can see as above that either there is a world satisfying $P \wedge \neg Q$, or type-F monism is true.

21 Hill (1997) tries to explain away our modal intuitions about consciousness in cognitive terms. Chalmers (1999) responds that any modal intuition might be explained in cognitive terms (a similar argument could "explain away" our intuition that there might be red squares), but that this has no tendency to suggest that the intuition is incorrect. If such an account tells us that modal intuitions about consciousness are unreliable, the same goes for all modal intuitions. What is really needed is not an explanation of our modal intuitions about consciousness, but an explanation of why these intuitions in particular should be unreliable.

Loar (1990/1997) attempts to provide such an explanation in terms of the unique features of phenomenal concepts. He suggests that (1) phenomenal concepts are recognitional concepts ("*that* sort of thing"); that (2) like other recognitional concepts, they can co-refer with physical concepts that are cognitively distinct; and that (3) unlike other recognitional concepts, they lack contingent modes of presentation (i.e., their primary and secondary intensions coincide). If (2) and (3) both hold (and if we assume that physical concepts also lack contingent modes of presentation), then a phenomenal-physical identity will be a strong necessity in the sense above. In response, Chalmers (1999) argues that (2) and (3) cannot both hold. The co-reference of other recognitional concepts with theoretical concepts is *grounded* in their contingent modes of presentation; in the absence of such modes of presentation, there is no reason to think that these concepts can co-refer. So accepting (3) undercuts any support for (2).

Chalmers (1999) also argues that by assuming that physical properties can have phenomenal modes of presentation non-contingently, Loar's account is in effect presupposing rather than explaining the relevant strong necessities.

22 Of those mentioned above as apparently sympathetic with type-C materialism, I think McGinn is ultimately a type-F monist, Nagel is either a type-B materialist or a type-F monist, and Churchland is either a type-B materialist or a type-Q materialist (below).

23 One might ask about specific reductive views, such as representationalism (which identifies consciousness with certain representational states), and higher-order thought theory (which identifies consciousness with the objects of higher-order thoughts). How these views are classified depends on how a given theorist regards the representational or higher-order states (e.g., functionally definable or not) and their connection to consciousness (e.g., conceptual or empirical). Among representationalists, I think that Harman (1990) and Dretske (1995) are type-A materialists, while Tye (1995) and Lycan (1996) are type-B materialists. Among higher-order thought theorists, Carruthers (2000) is clearly a type-B materialist, while Rosenthal (1997) is either type-A or type-B. One could also in principle hold non-materialist versions of each of these views.

24 Type-D dualists include Popper and Eccles (1977), Sellars (1981), Swinburne (1986), Foster (1991), Hodgson (1991), and Stapp (1993).

25 No-collapse interpretations include Bohm's "hidden-variable" interpretations, and Everett's "many-worlds" (or "many-minds") interpretation. A collapse interpretation that does not invoke measurement is the Ghirardi-Rimini-Weber interpretation (with random occasional collapses). Each of these interpretations requires a significant revision to the standard dynamics of quantum mechanics, and each is controversial, although each has its benefits (see Albert 1993 for discussion of these and other interpretations). It is notable that there seems to be no remotely tenable interpretation that preserves the standard claim that collapses occur upon measurement, except for the interpretation involving consciousness.

26 I have been as guilty of this as anyone, setting aside interactionism in Chalmers (1996) partly for reasons of compatibility with physics. I am still not especially inclined to endorse interactionism, but I now think that the argument from physics is much too glib. Three further reasons for rejecting the view are mentioned in Chalmers (1996). First, if consciousness is to make an interesting qualitative difference to behavior, this requires that it act non-randomly, in violation of the probabilistic requirements of quantum mechanics. I think there is something to this, but one could bite the bullet on non-randomness in response, or one could hold that even a random causal role for consciousness is good enough. Secondly, I argued that denying causal closure yields no special advantage, as a view with causal closure can achieve much the same effect via type-F monism. Again there is something to this, but the type-D view does have the significant advantage of avoiding the type-F view's "combination problem." Thirdly, it is not clear that the collapse interpretation yields the *sort* of causal role for consciousness that we expect it to have. I think that this is an important open question that requires detailed investigation.

27 Consciousness-collapse interpretations of quantum mechanics have been put forward by Wigner (1961), Hodgson (1991), and Stapp (1993). Only Stapp goes into much detail, with an interesting but somewhat idiosyncratic account that goes in a direction different from that suggested above.

28 Type-E dualists include Huxley (1874), Campbell (1970), Jackson (1982), and Robinson (1988).

29 Some accuse the epiphenomenalist of a double standard: relying on intuition in making the case against materialism, but going counter to intuition in denying a causal role for consciousness. But intuitions must be assessed against the background of reasons and evidence. To deny the relevant intuitions in the anti-materialist argument (in particular, the intuition of a further explanandum) appears to contradict the available first-person evidence; but denying a causal role for consciousness appears to be compatible on reflection with all our evidence, including first-person evidence.

30 Versions of type-F monism have been put forward by Russell (1927), Feigl (1967[1958]), Maxwell (1979), Lockwood (1989), Chalmers (1996), Griffin (1998), Strawson (2000), and Stoljar (2001).

31 There is philosophical debate over the thesis that all dispositions have a categorical basis. If the thesis is accepted, the case for type-F monism is particularly strong, since microphysical dispositional must have a categorical basis, and we have no independent characterization of that basis. But even if the thesis is rejected, type-F monism is still viable. We need only the thesis that microphysical dispositions *may* have a categorical basis to open room for intrinsic properties here.

32 Hence type-F monism is the sort of "physicalism" that emerges from the loophole mentioned in the two-dimensional argument against type-B materialism. The only way a "zombie world" W could satisfy the primary intension but not the secondary intension of P is for it to share the dispositional structure of our world but not the underlying intrinsic microphysical properties. If this difference is responsible for the lack of consciousness in W, then the intrinsic microphysical properties in our world are responsible for constituting consciousness. Maxwell (1979) exploits this sort of loophole in replying to Kripke's argument.

Note that such a W must involve either a different corpus of intrinsic properties from those in our world, or no intrinsic properties at all. A type-F monist who holds that the only coherent intrinsic properties are protophenomenal properties might end up denying the conceivability of zombies, even under a structural-functional description of their physical state – for reasons very different from those of the type-A materialist.

33 McGinn (1989) can be read as advocating a type-F view, while denying that we can know the nature of the protophenomenal properties. His arguments rests on the claim that these properties cannot be known either through perception or through introspection. But this does not rule out the possibility that they might be known through some sort of inference to the best explanation of (introspected) phenomenology, subject to the additional constraints of (perceived) physical structure.

34 In this way, we can see that type-D views and type-F views are quite closely related. We can imagine that if a type-D view is true and there are microphysical causal gaps, we could be led through physical observation alone to postulate higher-level entities to fill these gaps – "psychons," say – where these are characterized in wholly structural/dispositional terms. The type-D view adds to this the suggestion that psychons have an intrinsic phenomenal nature. The main difference between the type-D view and the type-F view is that the type-D view involves fundamental causation above the microphysical level. This will involve a more radical view of physics, but it might have the advantage of avoiding the combination problem.

35 Type-O positions are advocated by Lowe (1996), Mills (1996), and Bealer (forthcoming).

References

Albert, D. Z. (1993). *Quantum Mechanics and Experience*. Cambridge, MA: Harvard University Press.

Bealer, G. (1994). "Mental Properties." *Journal of Philosophy*, 91: 185–208.

Bealer, G. (forthcoming). "Mental Causation."

Block, N. and Stalnaker, R. (1999). "Conceptual Analysis, Dualism, and the Explanatory Gap." *Philosophical Review*, 108: 1–46.

Broad, C. D. (1925). *The Mind and its Place in Nature*. London: Routledge and Kegan Paul.

Campbell, K. K. (1970). *Body and Mind*. London: Doubleday.

Carruthers, P. (2000). *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge: Cambridge University Press.

Chalmers, D. J. (1995). "Facing up to the Problem of Consciousness." *Journal of Consciousness Studies*, 2: 200–19. Reprinted in Shear (1997). http://consc.net/papers/facing.html.

—— (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

—— (1997). "Moving Forward on the Problem of Consciousness." *Journal of Consciousness Studies*, 4: 3–46. Reprinted in Shear (1997). http://consc.net/papers/moving.html.

—— (1999). "Materialism and the Metaphysics of Modality." *Philosophy and Phenomenological Research*, 59: 473–93. http://consc.net/papers/modality.html.

—— (2002a). "The Content and Epistemology of Phenomenal Belief." In Q. Smith and A. Jokic (eds.), *Consciousness: New Philosophical Essays*. Oxford: Oxford University Press. http://consc.net/papers/belief.html.

—— (2002b). "Does Conceivability Entail Possibility?" In T. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility*. Oxford: Oxford University Press. http://consc.net/papers/conceivability.html.

—— (forthcoming). "The Foundations of Two-Dimensional Semantics." http://consc.net/papers/foundations.html.

Chalmers, D. J. and Jackson, F. (2001). "Conceptual Analysis and Reductive Explanation." *Philosophical Review*, 110: 315–61. http://consc.net/papers/analysis.html.

Churchland, P. M. (1996). "The Rediscovery of Light." *Journal of Philosophy*, 93: 211–28.

Churchland, P. S. (1997). "The Hornswoggle Problem." In Shear (1997).

Clark, A. (2000). "A Case Where Access Implies Qualia?" *Analysis*, 60: 30–8.

Dennett, D. C. (1991). *Consciousness Explained*. Boston, MA: Little, Brown.

—— (1996). "Facing Backward on the Problem of Consciousness." *Journal of Consciousness Studies*, 3: 4–6.

—— (forthcoming). "The Fantasy of First-Person Science." http://ase.tufts.edu/cogstud/papers/chalmersdeb3dft.htm.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Evans, G. (1979). "Reference and Contingency." *The Monist*, 62: 161–89.

Feigl, H. (1967[1958]). "The 'Mental' and the 'Physical'." *Minnesota Studies in the Philosophy of Science*, 2: 370–497. Reprinted (with a postscript) as *The "Mental" and the "Physical"*. University of Minnesota Press.

Foster, J. (1991). *The Immaterial Self: A Defence of the Cartesian Dualist Conception of the Mind*. Oxford: Oxford University Press.

Griffin, D. R. (1998). *Unsnarling the World-Knot: Consciousness, Freedom, and the Mind-Body Problem*. Berkeley: University of California Press.

Harman, G. (1990). "The Intrinsic Quality of Experience." *Philosophical Perspectives*, 4: 31–52.

Hill, C. S. (1997). "Imaginability, Conceivability, Possibility, and the Mind–Body Problem." *Philosophical Studies*, 87: 61–85.

Hodgson, D. (1991). *The Mind Matters: Consciousness and Choice in a Quantum World*. Oxford: Oxford University Press.

Huxley, T. (1874). "On the Hypothesis that Animals are Automata, and its History." *Fortnightly Review*, 95: 555–80. Reprinted in *Collected Essays*. London, 1893.

Jackson, F. (1979). "A Note on Physicalism and Heat." *Australasian Journal of Philosophy*, 58: 26–34.

——(1982). "Epiphenomenal Qualia." *Philosophical Quarterly*, 32: 127–36.

——(1994). "Finding the Mind in the Natural World." In R. Casati, B. Smith, and G. White (eds.), *Philosophy and the Cognitive Sciences*. Vienna: Holder-Pichler-Tempsky.

James, W. (1890). *The Principles of Psychology*. Henry Holt and Co.

Kaplan, D. (1989). "Demonstratives." In J. Almog, J. Perry, and H. Wettstein (eds.), *Themes from Kaplan*. New York: Oxford University Press.

Kirk, R. (1974). "Zombies vs Materialists." *Proceedings of the Aristotelian Society (Supplementary Volume)*, 48: 135–52.

Kripke, S. A. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Levine, J. (1983). "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly*, 64: 354–61.

——(2000). *Purple Haze: The Puzzle of Conscious Experience*. Cambridge, MA: MIT Press.

Lewis, D. (1988). "What Experience Teaches." *Proceedings of the Russellian Society* (University of Sydney).

——(1994). "Reduction of Mind." In S. Guttenplan (ed.), *Companion to the Philosophy of Mind*. Oxford: Blackwell.

Loar, B. (1990/1997). "Phenomenal States." *Philosophical Perspectives*, 4: 81–108. Revised edition in N. Block, O. Flanagan, and G. Güzeldere (eds.), *The Nature of Consciousness*. Cambridge, MA: MIT Press.

Lockwood, M. (1989). *Mind, Brain, and the Quantum*. Oxford: Oxford University Press.

Lowe, E. J. (1996). *Subjects of Experience*. Cambridge: Cambridge University Press.

Lycan, W. G. (1996). *Consciousness and Experience*. Harvard, MA: MIT Press.

Maxwell, G. (1979). "Rigid Designators and Mind–Brain Identity." *Minnesota Studies in the Philosophy of Science*, 9: 365–403.

Maxwell, N. (1968). "Understanding Sensations." *Australasian Journal of Philosophy*, 46: 127–45.

McGinn, C. (1989). "Can We Solve the Mind–Body Problem?" *Mind*, 98: 349–66.

Mills, E. (1996). "Interactionism and Overdetermination." *American Philosophical Quarterly*, 33: 105–15.

Nagel, T. (1974). "What Is It Like To Be a Bat?" *Philosophical Review*, 83: 435–50.
Nordby, K. (1990). "Vision in a Complete Achromat: A Personal Account." In R. Hess, L. Sharpe, and K. Nordby (eds.), *Night Vision: Basic, Clinical, and Applied Aspects*. Cambridge: Cambridge University Press.
Papineau, D. (1993). "Physicalism, Consciousness, and the Antipathetic Fallacy." *Australasian Journal of Philosophy*, 71: 169–83.
Perry, J. (2001). *Knowledge, Possibility, and Consciousness*. Cambridge, MA: MIT Press.
Popper, K. and Eccles, J. (1977). *The Self and Its Brain: An Argument for Interactionism*. New York: Springer.
Quine, W. V. (1951). "Two Dogmas of Empiricism." *Philosophical Review*, 60: 20–43.
Rey, G. (1995). "Toward a Projectivist Account of Conscious Experience." In T. Metzinger (ed.), *Conscious Experience*. Paderborn: Ferdinand Schöningh.
Robinson, W. S. (1988). *Brains and People: An Essay on Mentality and its Causal Conditions*. Philadelphia: Temple University Press.
Rosenthal, D. M. (1997). "A Theory of Consciousness." In N. Block, O. Flanagan, and G. Güzeldere (eds.), *The Nature of Consciousness*. Cambridge, MA: MIT Press.
Russell, B. (1927). *The Analysis of Matter*. London: Kegan Paul.
Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson and Co.
Seagar, W. (1995). "Consciousness, Information and Panpsychism." *Journal of Consciousness Studies*, 2.
Sellars, W. (1981). "Is Consciousness Physical?" *The Monist*, 64: 66–90.
Shear, J. (ed.) (1997). *Explaining Consciousness: The Hard Problem*. Cambridge, MA: MIT Press.
Shoemaker, S. (1975). "Functionalism and Qualia." *Philosophical Studies*, 27: 291–315.
Smart, J. J. C. (1959). "Sensations and Brain Processes." *Philosophical Review*, 68: 141–56.
Stalnaker, R. (1978). "Assertion." In P. Cole (ed.), *Syntax and Semantics: Pragmatics, Vol. 9*. New York: Academic Press.
Stapp, H. (1993). *Mind, Matter, and Quantum Mechanics*. Berlin: Springer-Verlag.
Stoljar, D. (2001). "Two Conceptions of the Physical." *Philosophy and Phenomenological Research*, 62: 253–81.
Strawson, G. (2000). "Realistic Materialist Monism." In S. Hameroff, A. Kaszniak, and D. Chalmers (eds.), *Toward a Science of Consciousness III*. Cambridge, MA: MIT Press.
Swinburne, R. (1986). *The Evolution of the Soul*. Oxford: Oxford University Press.
Tye, M. (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: MIT Press.
Van Gulick, R. (1993). "Understanding the Phenomenal Mind: Are We All Just Armadillos?" In M. Davies and G. Humphreys (eds.), *Consciousness: Philosophical and Psychological Aspects*. Oxford: Blackwell.
White, S. (1986). "Curse of the Qualia." *Synthese*, 68: 333–68.
Wigner, E. P. (1961). "Remarks on the Mind–Body Question." In I. J. Good (ed.), *The Scientist Speculates*. London: Basic Books.