dialectica

International Journal of Philosophy

Contents

EUROPEAN SOCIETY FOR ANALYTIC PHILOSOPHY

Vol 77 : 1
March 2023

dialectica

International Journal of Philosophy
Official Organ of the European Society of Analytic Philosophy

founded in 1947 by Gaston Bachelard, Paul Bernays and Ferdinand Gonseth

dialectica

vol 77
issue 1

March 2023

Contents

Is Somaliland a Country?

An Essay on Institutional Objects in the Social Sciences

JP Smit & Filip Buekens

Searle claims that his theory of institutional reality is particularly suitable as a theoretical scheme of individuation for work in the social sciences. We argue that this is not the case. The first problem with regulatory individuation is due to the familiar fact that institutional judgments have constrained revisability criteria. The second problem with regulatory individuation is due to the fact that institutions amend their declarative judgments based on the *inferential* (syntactic) properties of the judgments and in response to regulatory pressure, and not based on *descriptive* (semantic) properties and in response to matters of descriptive adequacy. These two problems imply that "regulatory kinds" (countries, borders, kings) will almost inevitably be disjunctive kinds that are ill-suited for scientific theorizing. This also explains why the law often makes odd pronouncements, e.g., calling ketchup a vegetable, considering an arm bent fifteen degrees to be straight, and not admitting that Somaliland is a country.

Somaliland is a democratically governed, autonomous region that maintains an independent police force, defends its borders, and issues currency in its own name.[1] Despite claims of statehood, it has not been officially recognized as a country by any state-level actors. Instead, it is considered an "autonomous region of Somalia." Scotland is a semi-autonomous region that neither controls nor defends its borders[2] and does not govern its own affairs to the degree that countries typically do. Despite being atypical in these respects, it is officially recognized as a country.

1 See "Why Somaliland is not a recognized state" in *The Economist*, November 1, 2015.

2 Or, at least, the borders that *are* defended are defended *qua* United Kingdom and not *qua* Scotland.

Atypical cases like Somaliland, Scotland, and others immediately raise the question as to the ontology of institutional objects like countries, presidents, money, borders, and traffic lights. The issue is particularly pressing among social scientists who study such phenomena. Suppose one is doing a cross-country comparative analysis of some social or economic trend. The trend does hold in Somalia (or the United Kingdom) but does not apply in Somaliland (or Scotland). In such a case, does Somaliland (or Scotland) serve as counter-examples, thus weakening any potential claim to generality? Or does Somaliland (or Scotland) not "count," hence not affecting the generality of any claim as to how widespread the trend actually is?[3]

The default answer one typically encounters when asking about what makes it the case that *X* is a country is that *X* is a country if, and only if, regulative bodies consider it a country. Of course, such regulative bodies have declared that ketchup is a vegetable,[4] that Microsoft is a person,[5] and that an arm bent 15 degrees is straight.[6] Botany, psychology, and mathematics have ignored these uses of "vegetable," "person," and "straight" to no ill effect. So why should we care what regulative bodies have to say when individuating the institutional world for the purposes of social science?

In this paper, we argue that social scientists should not feel compelled to individuate the social world in the same way that institutions do. The institutional use of language differs from the descriptive use proper to social science in at least two ways, and both serve to make the regulatory schemes of individuation used by institutions unsuited for descriptive work. The first bad consequence of regulatory individuation is due to the familiar fact that institutional judgments have constrained revisability criteria. This implies that the facts picked out by institutional judgments will almost inevitably be non-identical to the facts picked out by our best epistemic practices. The second

3 There are also more practical issues at stake. Somaliland, for instance, cannot receive state-level financial aid as such aid is earmarked for "countries" (Eubank 2015).

4 In 1981, the USDA's Food and Nutrition Service recommended that schools could comply with official nutritional regulations by crediting condiments as vegetables. Although ketchup was not specifically mentioned (pickle relish was mentioned as an example), it became known as the "Ketchup as a vegetable" controversy.

5 The doctrine of corporate personhood grants entities like corporations some of the rights and obligations normally reserved for actual people.

6 The rule states that "[a] ball is fairly delivered in respect of the arm if, once the bowler's arm has reached the level of the shoulder in the delivery swing, the elbow joint is not straightened partially or completely from that point until the ball has left the hand." Yet an arm that does bend up to 15 degrees is not considered to violate this rule. The current laws are available at https://www.lords.org/mcc/laws-of-cricket/.

bad consequence of regulatory individuation is due to the fact that institutions amend their declarative judgments based on the *syntactic* properties of the judgments and in response to regulatory pressure, not based on *semantic* properties and in response to matters of descriptive adequacy. This implies that "regulatory kinds" (countries, borders, kings) will almost inevitably be disjunctive kinds that are ill-suited for scientific theorizing. In making this argument, we reject the account of Searle (2005), whose position implies that social scientists should respect the individuation schemes of institutions.

1 Searle on Institutional Facts

John Searle, in a number of publications (1995, 2010), has defended an elegant view of institutional facts. The object of study is institutional objects, i.e., objects that serve some social purpose in virtue of having certain deontic powers (rights, duties, obligations). These deontic powers cannot be sufficiently explained by the intrinsic or natural properties of the object itself, but are the result of some institutional structure collectively endowing the object with such properties by *recognizing* it as having such properties.

A typical example is that of a president. A president is not a president in virtue of his or her physical or intrinsic properties, but in virtue of being *recognized* as a president by the governing institution of the country that he or she is president of. Paradigm cases of institutional objects also include countries, borders, driver's licenses, the playing field of a football game, and so on. Our social reality is filled with such objects, and we interact with them all the time.

Two aspects of Searle's view are of particular interest. First, he claims that institutional facts have the logical structure "*X* counts as *Y* in *C*" (1995, 28).[7] The *X*-term denotes the natural object, the *Y*-term is the institutional specification of the object, and the *C*-term denotes the context in which the institutional object has its function. In this way, Joe Biden counts as

7 A problem with this view is that the existence of some institutional facts does not seem to require the existence of anything for the *X*-term to denote. A paradigm case is money; most money does not exist in physical form but merely as account entries in bank ledgers. In response, Searle has stated that "*X* counts as *Y* in *C*" was only ever supposed to be a useful mnemonic that captures the core of his view (see Searle 2003). As "*X* counts as *Y* in *C*" is indeed a very useful mnemonic, and as nothing in the paper would be gained from using his later formulation, we stick with "*X* counts as *Y* in *C*." (On the topic of the ontological status of money, see Smit, Buekens and du Plessis 2016, where we argue that Searle's *X*-term can be interpreted as referring to an abstract object.)

doi: 10.48106/dial.v77.i1

the president of the United States at present, a specific line counts as the goal line during a game of football, and so on. Second, Searle claims that the recognition that is constitutive of the existence of an institutional fact is essentially *collective* recognition. Institutions are collectives, and the collective recognition by which they endow an object with deontic powers is irreducible[8] to individual recognition (1995, 24–25).[9]

Searle (2005) considers his view to have particular relevance for the social sciences. He introduces an article on the relevance of his view for economics (and social science in general) as follows:

> When I was an undergraduate at Oxford, we were taught economics almost as though it were a natural science. The subject matter of economics may be different from physics, but only in the way that the subject matter of chemistry or biology is different from physics. The actual results were presented to us as if they were scientific theories. So, when we learned that savings equals investment, it was taught in the same tone of voice as one teaches that force equals mass times acceleration. And we learned that rational entrepreneurs sell where marginal cost equals marginal revenue in the way that we once learned that bodies attract in a way that is directly proportional to their mass and inversely proportional to the square of the distance between them. At no point was it ever suggested that the reality described by economic theory was dependent on human beliefs and other attitudes in a way that was totally unlike the reality described by physics and chemistry. (Searle 2005, 1)

Searle sets up a basic distinction between the objects of the physical sciences and the objects of social sciences and advises social scientists to heed the fact that their objects are fundamentally unlike those of the physical sciences. The objects of social science are frequently institutional objects and, as

8 For a critique of this claim, see Smit, Buekens and du Plessis (2011, 2014), where we develop the incentive account of institutional facts. On our view, institutions can be fully understood in terms of incentives and actions, and the recognition of such incentives and actions need not be collective. The view is similar to Guala and Hindriks (2015; Hindriks and Guala 2015)—also see Guala (2016)—who account for institutions in terms of rules in game theoretical equilibria.

9 Searle, in recent years, has recognized that, in some cases, forms of collective institutional recognition may reduce to individual recognition (2010, 58).

such, should be understood as explained above, i.e., in terms of the collective recognition of objects as having certain deontic powers.

Of particular importance to the current discussion is his claim that such objects can only exist for as long as they are represented as existing (1995, 13) and his claim that such objects should be understood as having a logical structure (1995, 22). This implies that if one is a social scientist and wishes to study borders, countries, or presidents, then one must understand one's area of study as pertaining to those things *recognized* to be borders, countries, and presidents. In other words, those things that exist in virtue of the collective acceptance of a declaration of the form "*X* counts as *Y* in *C*." This implies, although it is not explicitly stated by Searle, that the social scientist must individuate the institutional world as the institutions that create it do. For if it is constitutive of borders, countries, and presidents that they must be *recognized* to be these objects, then studying objects not so recognized is to not study borders, countries, and presidents at all.

2 Two Peculiarities of the Institutional Use of Language

2.1 *A Toy Example*

Below, we will argue that social scientists would be ill-served if they employed the individuation schemes used by institutions themselves. The argument is based on the fact that institutions use language in peculiar ways, and these ways make institutional standards of individuation ill-suited for the purposes of scientific description. This is not to say that there is any specific problem with describing institutions; rather, the claim is that the social scientist should not feel compelled to use the regulatory schemes of individuation adopted by institutions when describing institutional facts.

For purposes of exposition and illustration, it will be useful to have a toy example at our disposal. Suppose there is a village in the Scottish Highlands that has a cultural ritual called "Firecasting" that takes place annually on the first day of spring. They celebrate the end of winter and the reduced need for heating by letting each member of the village attempt to light a torch on fire, run to the firecasting line, and hurl it into a lake, extinguishing the flame within twenty seconds. Those who succeed get a medal (and receiving such a medal has significant prestige in the village).

doi: 10.48106/dial.v77.i1

In Firecasting, there is an umpire who keeps time, adjudicates whether a flame has been extinguished, etc. Every time a torch has been extinguished, the umpire proclaims, "Player *x* is a firecaster," i.e., a flame has been extinguished.

2.2 *First Peculiarity Constrained Revisability*

The umpire in Firecasting has to judge whether a specific state of affairs obtains, namely, whether the flame has been extinguished. This is a judgment that any spectator can also make. The umpire's judgment, however, counts in a way that the judgments of spectators do not. Consider the following judgment:

(1) John is a firecaster.

If a spectator makes a judgment by using (1), then this is a speech act of description that asserts that John extinguished the flaming torch. As it is a standard instance of description, it can be straightforwardly true or false. Call the use of the institutional term "firecaster" in a speech act of description the *descriptive use* of the term.

If (1) is used by the umpire, however, the situation is different. The umpire's use of (1) is *based on* his assessment of whether the flame has been extinguished, yet his speech act is that of declaration. His speech act has the function of creating a certain institutional fact, namely, the institutional fact that John is a firecaster. In Searle's terminology, such a judgment has certain deontic consequences, namely, that the player is entitled to be awarded a medal by the village. Call such a use of the institutional term "firecasted" the *regulative use*.[10]

Note that if the umpire makes a mistake in adjudging whether John has extinguished the flame and erroneously declares that he is a firecaster, then the descriptive content of (1) is false, yet the regulative content of (1) can still be affirmed. This is so that, even if the umpire makes a mistake, the deontic consequences of his judgment will still obtain, i.e., John will still be entitled to the medal. What the umpire commits himself (and the village) to through the speech act of declaration is, above all else, that John is entitled to receive the medal. The umpire's judgment might be *based on* whether the descriptive

10 Of course, (1) can also be used in a third way: as a *report* of an umpiring judgment. This use, while also descriptive, is distinct from the descriptive use in the main text and need not trouble us here.

content of (1) obtains, yet what is affirmed by the umpire in making his judgment is something else.[11]

It is a staple of the literature on the philosophy of law (see, for instance, Hart 1961) that institutional declarations cannot be revised in light of future evidence in the same way that descriptive judgments can. Even if the umpire and the village see conclusive evidence that John did not extinguish the torch, they may choose to "let the judgment stand," i.e., to remain committed to enforcing the deontic consequences of the original regulative judgment. While the village may choose to explicitly adopt regulative rules that do allow for the revision of prior judgments, there is nothing inherently irrational about not doing so, as affirming the regulative content of (1) does not logically commit them to any specific position as to the truth value of the descriptive content of (1). In this way, the regulative judgment contrasts sharply with the descriptive judgment, as it is a *sine qua non* of descriptive practice that such judgments are always revisable in light of future evidence.

In fact, most real-world sports do not, except in extreme cases, allow such later reviews of umpiring decisions. Mistakes inevitably happen, and sports fans everywhere use the institutional term itself in a descriptive way in order to register their disagreement with the referee. Consider judgments like "That was never a strike!" "He was miles offside!" "It pitched outside leg stump!" and so on. In such cases, the utterer uses a *non-institutional*, descriptive standard for applying the terms "strike," "offside," and "outside." Here, the institutional term is used in order to voice disagreement with the factual basis of an umpiring decision (and also to draw attention to the unfairness of the deontic consequences of such a decision).

The phenomenon of *constrained revisability* is found in all institutional settings. While institutional judgments can sometimes be overruled—i.e., appealed in various ways—such revisability is constrained in a way that that open-ended, epistemic inquiry is not. For example, legal systems in a wide variety of countries recognize a principle of "double jeopardy," whereby an accused cannot be retried for an offense that they have already been acquitted of. This remains so even if definitive evidence of prior guilt is produced and

11 This distinction between the basis for an institutional judgment and its deontic consequences was first set out in Ransdell (1971, 388). I am departing from his terminology (he distinguishes between the "connotation" of a term and its "import"), but this departure should not be taken to imply any difference of substance. Ásta draws a similar distinction between "base properties" and "conferred properties" (2018).

doi: 10.48106/dial.v77.i1

no one believes that the descriptive judgment underlying the institutional declaration was accurate.

Some legal systems do allow various, tightly restricted exceptions to this principle. In general, though, the revisability of the legal declaration is constrained in a way that commitment to the underlying descriptive claim is not. The distinction between the descriptive use and the regulative use of legal terms is again well-recognized in our ordinary discourse. Consider judgments like "Andy Dufresne was innocent,"[12] "OJ Simpson was guilty," "Jimmy Saville was a criminal," and so on.[13]

The fact that (1) can express distinct speech acts with distinct criteria of revisability means that the denotation of the descriptive use and the regulative use of a term can diverge. In the case of the game Firecasting, the denotation of the descriptive use of the term "firecasters" will include those who succeeded in extinguishing a flame. The denotation of the regulative use will include all those who were *adjudged* to have extinguished a flame. If a scientist were to study firecasting, then the nature of the study might force her to take the distinction seriously and use one or the other criterion. In this way, if the scientist were tasked with determining what physical characteristics allow one to firecast, then she would be ill-served by the regulative use. This is because, if a number of serious umpiring errors have occurred in the history of Firecasting, any law-like generalization that the scientist seeks to uncover will be much more likely to apply to those who *actually* achieved the feat of extinguishing a flame and not merely those adjudged to have done so. The denotation of the regulative use of "firecasting" will almost inevitably be non-identical to the denotation of the descriptive use of "firecasting"; the denotation of the former will be more heterogeneous with regards to physical characteristics (as it includes both those who succeeded and those who did not) and, as such, less likely to be the object of useful law-like generalizations of the required type.

The opposite is likely to be true for the historian of the game. The historian who writes about the stars of the game is implicitly, and correctly, writing about the regulative use when writing about those *recognized* to have fire-

12 The protagonist of the Stephen King novella *Rita Hayworth and the Shawshank Redemption*—later made into the film *The Shawshank Redemption*—who was convicted of a crime he did not commit.

13 The disagreement need not take the form of a factual disagreement but can also be used to express disagreement with the normative judgment behind an institutional judgment. Few people would consider Nelson Mandela a "terrorist," despite the fact that he used to be on the US terrorist watch list.

casted. Here, the main interest lies in those falling under the denotation of the regulative use, and as such, reports of prior regulative use are appropriate to the study. As this is the main topic of interest, any law-like regularities that the typical sports historian seeks are supposed to concern those who were *adjudged* to have firecasted.

The same distinction applies to the academic study of less frivolous matters. Consider a criminologist who aims to make discoveries about the causes of crime in order to determine how law-breaking can be prevented. Here, the interest is likely to lie in determining what causes people to break the law. Breaking the law, of course, is not the same thing as being *adjudged* to have broken the law. In this manner, the criminologist would feel vindicated if their theory applied to someone who did commit a crime but was never caught or convicted. In the same way, they would be untroubled if their theory did not apply to someone who was wrongfully convicted. When we express an interest in preventing crime, we are typically not expressing an interest in having fewer of the people who break the law caught but in having the law broken less. This makes the *descriptive use* of the term "criminal" the one appropriate to such a study.

The criminologist could, of course, decide to try and find out what distinguishes those convicted of committing a crime from those not so convicted. Here, the regulative use of "criminal" would be appropriate to the study. Note, however, that we have some reason to believe that in typical cases, law-like generalizations are more likely to apply to the descriptive use. The denotation of the descriptive use of criminal would include all those who broke the law. The denotation of the regulative use would include those who broke the law and were convicted and those who did not break the law and were convicted, while excluding those who did break the law and were not convicted. As the latter category is individuated in terms of a more heterogeneous mix of properties, one would suspect, *ceteris paribus*, that the descriptive use of the term "criminal" would be more suitable for obtaining law-like generalization. Simply put, it will typically be easier to obtain general truths among a group whose members were rightfully classified as belonging to the group than among a group that includes a mixture of those correctly and incorrectly classified as members of the relevant group.

The above reasoning implies that the constrained revisability of regulative judgments sometimes gives the social scientist a good reason to, despite using the terms used by some specific institution itself, reject the individuation scheme of the institution. This is so as, firstly, the ultimate goal of the inquiry

doi: 10.48106/dial.v77.i1

(i.e., crime prevention) may demand it. Secondly, the descriptive use of the institutional term will be more suitable for law-like generalization and, so, more useful to social science.

2.3 *Second Peculiarity Institutional Judgments Are Amended Based on Their Inferential (Syntactic) Properties, Not Descriptive (Semantic) Properties*

Suppose that the village that practices Firecasting notices that players are sometimes prevented from hurling the torch by other players kicking them just as they are about to hurl it and, in so doing, making it less likely that the player throws across the line. They wish to make such behavior pointless and, so, announce that players who are kicked as they are about to throw will receive a medal anyway, even if their throw did not cross the line.

The required rule change can be made in two distinct ways. Prior to the rule change, the relevant rules of Firecasting are as follows:

(1) A player x has firecasted if, and only if, x is adjudged to have extinguished the flaming torch by hurling it into the ocean within the context of the game of Firecasting.
(2) Firecasters are entitled to receive a medal from the village.

The first way to change the rule so as to award those who were kicked just prior to throwing would be to amend the definition of firecasting so that those who were kicked also "count" as firecasters. This option is analogous to a "penalty try" in rugby. If a rugby player is illegitimately prevented from scoring a regular try, the referee may award a so-called "penalty try" to the team prevented from scoring. A rugby team awarded a penalty try is awarded five points in the same way that a team that scores a regular try is awarded five points. In this way, the penalty try "still counts," despite the fact that the attacking team was prevented from scoring a regular try. In the same spirit, the village can amend (1) as follows:

(1*) A player x has firecasted if, and only if, x is adjudged to have extinguished the flaming torch by hurling it into the ocean or x is adjudged to have been kicked prior to hurling the flaming torch within the context of a game of firecasting.

The village, however, need not amend the definition of "firecasting" in order to secure the result that those who are kicked in order to prevent them from firecasting still receive a medal. They can leave (1) intact and simply amend (2) so that it states that those who are kicked also receive a medal from the village. In this way, (2) can be amended as follows:

(2*) Firecasters and those who were adjudged to have been kicked just prior to hurling the flaming torch within the context of a game of Firecasting are entitled to receive a medal from the village.

(2*) directly regulates the result of the person attempting to firecast being kicked, whereas (1*) does indirectly by changing the concept of "firecasting." Yet these two ways of amending the rules are equivalent; both of the above rule changes would have the effect that those who were kicked receive a medal. The change can be made in distinct ways, as, in the context of enforcement of such rules, the rules of Firecasting constitute a set of *premises* that can be amended in distinct ways so as to, in conjunction with judgments about an instance of the game, imply the statement that some player who has been kicked should receive a medal. In other words, the aim of the village, when amending the rules of Firecasting, is to appropriately link the following two statements concerning some specific instance of the game.

(3) Player *K* was adjudged to have been kicked prior to hurling the flaming torch within the context of a game of Firecasting.
(4) Player *K* is entitled to receive a medal from the village.

(3) is a specific judgment concerning some specific instance of the game of Firecasting, and (4) is the regulatory response to what was adjudged to have happened in some such specific instance of the game. The village aims to formulate rules that, in conjunction with (3), imply (4). The combination of (1*), (2), and (3) implies (4), and the combination of (2*) and (3) also implies (4). In this way, the fact that the rules can be amended in distinct ways reflects no more than the fact that the same conclusion can follow from distinct sets of premises.

In the above case, the first way of changing the rules amounted to changing the definition of "firecasting," whereas the second amounted to changing the statement of rewards given out by the village. The regulatory equivalence of these changes in our toy example is a phenomenon that applies to law in general. When we wish to amend the law in order to secure a specific

doi: 10.48106/dial.v77.i1

consequence, there will always be distinct ways of doing so, and the only criteria for choosing whether to amend a definition or amend some statement of penalties or awards is, where rational, pragmatic.

Legal language turns out to be holistic in an almost Quinean way (Quine 1951). The law is holistic in two distinct ways. First, there is no one correct way to change the law so as to secure some regulatory response. Second, the list of claims we call definitions has no special status that prevents them from being changed so as to secure the desired regulatory response.

The fact that the law is holistic serves to explain why the law often uses perfectly ordinary terms in peculiar ways. The claim that a company is a "person" is just a tool to secure a regulatory response concerning the legal liability of the members of a corporation; the claim that ketchup is a "vegetable" is a tool to effectively lower the legally mandated nutritional requirements for school lunches. In the same way, the claim that an arm bent 15 degrees is "straight" is a tool to secure the result that cricketers may bowl with a slightly bent arm.

Cases of atypical use of terms like "ketchup," "person," "straight," and the like serve to demonstrate something important. When the lawmaker changes the law, it has no overriding reason to respect the semantics of the term (as used in non-legal contexts). Rather, specific statements only matter inasmuch as they help to, in conjunction with other statements, secure the desired regulatory response when the law is applied. Such regulatory responses in specific instances can be represented as the conclusions of arguments that have legal statements among their premises. As the overriding factor governing the formulation and emendation of laws is the regulatory response to which it gives rise, this implies that the overriding factor governing the emendation of statements within a system of law is *the role of such statements in facilitating inference*. This, in turn, implies that we can expect changes to the law to end up radically changing the denotation of terms that were originally used in a perfectly ordinary sense. In the final analysis, this is due to the fact that institutional judgments are amended in virtue of their inferential (*syntactic*) properties and not their descriptive (*semantic*) properties.

The first peculiarity of language that was noted was that constrained revisability meant that the denotation of the regulative use of a term would not exactly coincide with the denotation of the descriptive condition based on which the term is applied. The second problem, however, is much more basic and would apply even if judges never made mistakes. Lawmakers will change the content of perfectly ordinary terms in order to secure regulatory consequences. This implies that the legal system will tend towards a scheme

of individuation designed to serve regulatory and not descriptive purposes. This much is obvious enough, but it has the less commonly understood consequence that institutional judgments will be amended based on their syntactic properties and not on their semantic properties.

This implies that technical terms introduced for some regulative purpose (like "firecasting") are not constrained so as to include only relevantly similar elements under their denotation. Furthermore, even when the regulative term is taken from ordinary language (like "straight," "vegetable," etc.), the term will tend to start out being used as legal terms in their familiar sense but will often end up including unlike objects in the same category. In this way, "regulatory kinds" will end up, if judged against a standard of descriptive adequacy, becoming disjunctive kinds that are ill-suited to scientific theorizing.

No one would expect mathematicians to do useful work while treating an arm bent 15 degrees as straight, and no one would expect a botanist to employ the term "vegetable" so as to include ketchup. These cases, however, are just the tip of the iceberg that serve to make the general phenomenon visible. The physical sciences pay no attention to regulative bodies when individuating the world, as such bodies are simply involved in another kind of activity altogether. In the same way, there is no reason for the social scientist to consider herself uniquely encumbered by, and beholden to, a schema of individuation that does not serve her purposes.

Note that the point concerning ordinary terms being introduced into law is not merely that such regulatory terms "change their meaning." The problem, rather, is that such changes occur due to inferential (syntactic) considerations. Legal changes may be phrased as changes in definition (as when the definition of "firecasting" is amended so as to include being kicked) or as changes in regulation (such as when those kicked during firecasting also receive a reward). Whether these changes are phrased as one or the other change has little, if anything, to do with descriptive adequacy, and so we end up with categories that mix unlike things together, i.e., "regulatory kinds" become disjunctive kinds.

To illustrate the above point, consider the difference between the descriptive term "computer" and the regulative term "king." The term "computer" was originally used to denote people, specifically those employed to engage in tedious tasks of rote calculation.[14] The project to mechanize such tasks was originally described as the project of creating a "mechanical computer," and

14 For an interesting history of pre-mechanical calculation, see Grier (2005).

doi: 10.48106/dial.v77.i1

this description was no mere tautology. Once the project succeeded, however, and the human computers disappeared, the meaning of the term "computer" underwent a social shift until it denoted only machines designed to perform such calculations. In fact, the change in the use of the term exhibits a nice symmetry; today, if we call someone a "computer," it is a metaphorical use of the term that suggests extreme proficiency at calculation, a tendency to act without emotion, or some such.

The above change is generally socially recognized as a fundamental change in the *meaning* of the term "computer." In principle, we could have treated the shift differently, for instance, by saying that the term merely expanded its denotation so as to include both human and mechanical, and eventually electronic, computers. There would be no point in doing so, however, for then "computer" would be a disjunctive kind that groups two radically distinct kinds of things together. Our descriptive language is guided by descriptive adequacy, and so simply treating the content of the term as having changed completely individuates the world in a much more useful way.

The same is not true for the institutional pair "king"/"queen" when used in a regulative manner. Kings and queens, historically, are paradigmatic persons who rule a state by decree and who obtain their position by right of birth. Today, however, in the vast majority of countries that still recognize a "king" or "queen," being a king or queen is primarily a symbolic or ceremonial role. While today's kings and queens do have some influence, this influence is so different in kind from the right to rule by decree that the two kinds of "king" or "queen" are beyond any meaningful similarity or comparison. While we may loosely say that "the meaning of being a king or queen has changed," we do not generally consider the term to have changed its semantic content in the same way that the term "computer" has. This is despite the fact that the term categorizes together entities with vastly different social roles. If a historian or sociologist were to uncritically accept the institutional use of the term "king" (or "queen") and try to determine commonalities or differences between kings, the very category of analysis would serve to unnecessarily complicate the inquiry. The term would group together those who ruled by decree as well as those whose social role is effectively a more dignified version of a mascot. If our purpose is descriptive adequacy, then little is to be gained by an individuation scheme that treats those who ruled by decree (old-style kings, the present-day King of Swaziland, etc.) with the current Queen of England or the current Queen of Denmark. Furthermore, it would exclude those whose social role is similar to that of old-style kings and queens, i.e.,

dictators who de facto rule by decree and have their position in virtue of birth, e.g., the North Korean leader Kim Jong-Un.

The point of the above is not to criticize the existence of present-day royalty or to suggest a change in linguistic habit. The point, rather, is that a social scientist who accepts an individuation scheme in which an all-powerful king and the current Queen of Denmark are the same sort of thing[15] is as absurd as a mathematician who treats all lines that bend less than 15 degrees as "straight."

The above considerations concern both cross-institutional identification (i.e., whether the current King of Denmark and the current King of Swaziland are the same sort of thing) and inter-temporal identification (i.e., whether the kings and queens of centuries ago are the same sorts of things as the current Queen of England). It is a fundamental constraint upon inquiry that our criteria of individuation remain *constant*, and here, the regulative use of institutional terms is a poor guide to scientific individuating practices. It is for this reason that the social scientist should feel under no obligation to accept institutional standards of individuation; in fact, she should rearrange the conceptual world as she sees fit.[16]

3 Cases Where the Regulative Use of Institutional Terms Is the Correct Use

The point of the above is not that social scientists should never employ the regulative use of institutional terms as basic terms of inquiry. Institutions do manage to affect the world via declarations, and Searle is correct that this phenomenon is important to understanding our social and institutional world. We can distinguish three reasons for adopting the regulative use as a term of inquiry.

First, our interest may lie precisely in *the objects grouped together* by the regulative use of an institutional term. In this way, as mentioned earlier, we may wish to inquire into the difference between those who are convicted of

15 One could object that kings and queens do form a kind in virtue of their genetic relation to an ancestor. This is so and means that the term, so construed, would be useful for geneticists. Most of the time, however, when considering kings and queens, our interest lies in their social role, and here, the regulative use of the term is a plain obstacle to inquiry.

16 Our account has the additional advantage that it does not overemphasize the role of normativity in the causal processes operative in social reality. See Turner (2010) and Guala (2015) for critical assessments of the (over)use of normativity in the social sciences and social ontology.

doi: 10.48106/dial.v77.i1

committing a crime and those who, while having committed a crime, are acquitted. Or, alternatively, we may be interested in the difference in severity of sentence among those convicted of a crime. Such topics are a staple of criminological and sociological studies that try and determine what effect categories of identity (race, gender, etc.) or socioeconomic attainment have on rates of conviction and severity of sentence. In such cases, our interest lies precisely in a category that exists in virtue of regulative declarations; hence, the regulative category is proper to the study.[17]

Second, the declarations made by institutions have a *causal impact* on the world, and our interest may lie precisely in studying the effect of such an impact. In this way, the criminologist may be interested precisely in the impact of a criminal conviction on one's life prospects. In this case, again, the regulative use is proper to the study in virtue of the causal role of such regulative judgments.

An interesting sub-class of the causal impact that institutional declarations can have is where such declarations have a *symbolic impact* on the objects of such a declaration; we may well be interested in studying the nature and effects of such a symbolic impact. In this way, a historian or sociologist may be interested in changes in self-conception that occur among those people occupying a territory that is widely recognized as being a "country," or changes in self-conception among those recognized as "criminals," and so on.[18]

Third, the social scientist may be interested in a category that need not be governed by regulative use, but the institutional use is close enough, for the purposes of the study, to what they are trying to identify that it is a *useful proxy* for the descriptive use. If a country's rules remain relatively stable over time, the judiciary does a decent job of applying the laws, and the concepts involved happen to individuate reality in a descriptively useful way, the institutional category should be good enough for useful inquiry. Consider, for instance, a scientist who wishes to study whether the color of a motor vehicle has an impact on people's propensity to speed. Strictly speaking, some people who speed will not be among those convicted of speeding, whereas some of those

17 See Wilson (2007) for a related argument that the importance of Searle's work to the social sciences is more limited than one might suppose.

18 In this paper, we mostly speak of classification as a matter of putting objects with similar causal powers together. An anonymous referee points out that the social sciences do more than try to arrive at law-like generalizations. Nothing in our argument prohibits non-causal schemes of individuation that may prove useful in interpretive or normative projects; the point is that the Searlean project does not tie our hands.

convicted will have been innocent. But if all the scientist is looking for is a rough correlation in aggregate and the legal system has been reasonably efficient, then counting all those convicted of speeding as "speeders" should be a good enough sample to do meaningful statistical work.

In endorsing the above regulative uses, we also embrace something close to pluralism about general institutional terms. Good usage will be polysemic; the social scientist will inevitably have to craft the terms of their inquiry to the topic at hand. What we object to, on the grounds discussed, is the idea that regulative bodies should be implicitly granted the power to set the terms of our descriptive agenda.[19]

4 Is Somaliland a Country?

Somaliland is a country. More specifically, we think that, except in the very specific types of cases previously explained, i.e., cases where our epistemic interest is precisely in those objects grouped together by institutional declarations, the social scientist should view Somaliland as a country. Here, we do not base this claim on any specific definition of the term "country." Rather, our judgment reflects the fact that Somaliland, once we ignore regulatory schemes of individuation for the reasons outlined in this paper, seems entirely like paradigm cases of countries, i.e., Kenya, Germany, Chile, Japan, etc.

In this paper, we have explained why we think that the social scientist should, in principle, be very wary of adopting institutional schemes of individuation. This matters, as the currently dominant theory of institutions, i.e., the Searlean theory, effectively adopts and legitimizes institutional schemes of individuation, and, hence, it is worth knowing why the social scientist should feel free to disregard Searle's view. This is so, especially as Searle explicitly recommends that social science adopt his theory of institutions (Searle 2005). Also note that *reflexive* definitions, i.e., definitions on which an entity gains its identity from being *considered* to be the things that it is, long predate Searle.

More important, however, is the question as to the scope of the problem, i.e., the question of how much harm is done by social scientists adopting regulatory schemes of individuation. Our argument is compatible with *quietism* about regulatory individuation, i.e., the view that Somaliland is an edge case, a mere curiosity whose exclusion from the list of countries does no real harm to social analysis. Our view is also compatible with *revisionism* about regulatory

19 We would like to thank an anonymous referee for pressing us to be explicit on this point.

doi: 10.48106/dial.v77.i1

individuation, i.e., the view that Somaliland and cases like it serve to make visible a deep problem that calls for social scientists to abandon regulatory schemes of individuation in favor of a series of successor concepts more suited to descriptive purposes.

The question of which position on the continuum between quietism and revisionism is most justifiable is beyond the ambition of the present work. We can see the appeal of quietism; it would appear ridiculous to expect an economist writing about the correlation between countries in the measured link between inflation and unemployment to worry too much about whether his fundamental categories of analysis are making his job harder than it needs to be.

We can also, however, see the appeal of revisionism. Consider the definition below, intended to capture the regulatory notion of a "country":[20]

> A country is a region that is identified as a distinct entity in political geography. A country may be an independent sovereign state or part of a larger state, as a non-sovereign or formerly sovereign political division, or a geographic region associated with sets of previously independent or differently associated people with distinct political characteristics.

The above definition—in addition to being vague—is disjunctive to an extreme degree. It is akin to a definition of "vegetable" that includes not only ketchup but also all bottles of Worcestershire sauce that are older than three months. The problem with such a disjunction is plain: What possible reason could we have to expect that some underlying causal process could produce similar effects across entities that have been grouped together merely as a matter of a series of historical regulatory contingencies?

Current practice seems to imply at least some deviation from quietism. Social scientists are not naive and have not stayed slavishly faithful to institutional categories of individuation. The *CIA World Factbook*, for example, on its list of countries by gross domestic product, does not list England, Scotland, or Wales among the entries, even though they are generally recognized as countries.[21] It does, however, list the European Union, despite the fact that it is not recognized as being a country. This makes sense, as the interest of the

20 From worlddata.info.

21 *CIA World Factbook* is available at: https://www.cia.gov/library/publications/resources/the-world-factbook/rankorder/2001rank.html.

economist would be in finding a category of individuation that identifies individual units of action, i.e., units with a fair degree of autonomy *qua* matters of economic production and exchange. When it comes to such practices, the present paper serves to justify how such deviations from institutional schemes of individuation are, *contra* Searle, perfectly justified.[22]

The question, however, is whether current practice occupies the appropriate position on the continuum between revisionism and quietism. Note that the *CIA World Factbook* does not list Somaliland, despite there being very little reason not to do so once we abandon the purely regulatory use of "country." In fact, once we take the matter of individuation seriously, we may well have reason to include sub-units of various "countries" under the de facto control of some entity other than the recognized government, i.e., parts of "countries" under the control of rebel groups or drug cartels.[23] This may sound radical, but if our interest lies in discovering the units of political and/or economic action—and hence in groups that have a high degree of autonomy over running their own affairs—then there is little reason to exclude them. We may well learn interesting things by considering such entities *qua* units of economic and political action, for they are effectively no different from "countries" under military or dictatorial control.

5 Conclusion

In this paper, we have argued that social scientists should not be Searleans when it comes to their own categories of analysis, i.e., be wary of employing the regulative use of institutional terms for purposes of individuation. There are two main problems. First, the revisability of institutional judgments is non-epistemically constrained, i.e., mistakes do not get corrected in the same way that we correct them when dealing with descriptive assertions. This means that the social scientist would frequently be better served by employing the descriptive use, and not the regulative use, of institutional terms as a basis of

22 The Searlean could respond by saying that such usage of "country" is a mere loose usage, done for practical purposes. Such a response, however, opens up the line of attack that we have been pressing, for it implicitly admits that the Searlean scheme of individuation is not suited to social science. We would like to thank an anonymous referee for pressing us on this point.

23 The CIA estimates that roughly 20% of Mexico is under the control of the drug cartels. (See "Mexico's government control threatened by criminal groups claiming more territory" in *The Washington Post*, October 29, 2020.) Interestingly, some such drug cartels engage in activities commonly associated with governments, e.g., the provision of social services. See Flanigan (2014) for a discussion of this phenomenon.

doi: 10.48106/dial.v77.i1

individuation. The second problem, and by far the most important one, is due to the fact that institutions individuate in order to regulate, not to describe. Such regulation is holistic, and, hence, the usage of terms will change based on their *syntactic* properties and in response to regulatory pressure, not based on their *semantic* properties and in response to matters of descriptive adequacy.[24] This means that "regulative kinds" will inevitably tend to become disjunctive kinds, and so the law will be prone to the seeming absurdity of classifying ketchup as a vegetable, considering an arm bent 15 degrees to be straight, and so on. This implies that the social scientist will sometimes be better advised to ignore *both* the descriptive and regulative use of institutional terms and to invent institutional categories that have never been subject to regulative declaration at all.*

JP Smit
0000-0003-1524-8845
University of Stellenbosch
jps@sun.ac.za

Filip Buekens
0000-0003-3770-0513
KU Leuven
filip.buekens@kuleuven.be

References

Ásta [Ásta Kristjana Sveinsdóttir]. 2018. *Categories We Live By: The Construction of Sex, Gender, Race, and Other Social Categories*. Oxford: Oxford University Press, doi:10.1093/oso/9780190256791.001.0001.

Eubank, Nicholas. 2015. "Taxation, Political Accountability and Foreign Aid: Lessons from Somaliland." *The Journal of Development Studies* 48(4): 465–480, doi:10.1080/00220388.2011.598510.

Flanigan, Shawn T. 2014. "Motivations and Implications of Community Service Provision by La Familia Michoacána / Knights Templar and other Mexican Drug Cartels." *Journal of Strategic Security* 7(3): 63–83, doi:10.5038/1944-0472.7.3.4.

24 The two problems, i.e., constrained revisability and responsiveness to syntactic properties, are, of course, linked in that both are the result of the fact that our interests in making regulative judgments are practical and normative. I would like to thank an anonymous referee for highlighting this point.

* THANKS

GRIER, David Alan. 2005. *When Computers Were Human*. Princeton, New Jersey: Princeton University Press, doi:10.1515/9781400849369.

GUALA, Francesco. 2015. "The Normativity of Institutions." *Phenomenology and Mind* 9: 118–128, doi:10.13128/Phe_Mi-18157.

—. 2016. *Understanding Institutions: The Science and Philosophy of Living Together*. Princeton, New Jersey: Princeton University Press, doi:10.1515/9781400880911.

GUALA, Francesco and HINDRIKS, Frank. 2015. "A Unified Social Ontology." *The Philosophical Quarterly* 65(259): 177–201, doi:10.1093/pq/pqu072.

HART, H. L. A. 1961. *The Concept of Law*. Oxford: Oxford University Press. Third edition: Hart (2012).

—. 2012. *The Concept of Law*. 3rd ed. Oxford: Oxford University Press. First edition: Hart (1961), third edition with a postscript, edited by Penelope A. Bulloch and Joseph Raz, doi:10.1093/he/9780199644704.001.0001.

HINDRIKS, Frank and GUALA, Francesco. 2015. "Institutions, Rules, and Equilibria: A Unified Theory." *Journal of Institutional Economics* 11(3): 459–480, doi:10.1017/S1744137414000496.

QUINE, Willard van Orman. 1951. "Two Dogmas of Empiricism." *The Philosophical Review* 60(1): 20–43. Reprinted in Quine (1953, 20–46), doi:10.2307/2181906.

—. 1953. *From a Logical Point of View: 9 Logico-Philosophical Essays*. Cambridge, Massachusetts: Harvard University Press. Cited after the revised edition: Quine (1961).

—. 1961. *From a Logical Point of View: 9 Logico-Philosophical Essays*. 2nd ed. Cambridge, Massachusetts: Harvard University Press. Revised edition of Quine (1953), reprinted 1980.

RANSDELL, Joseph. 1971. "Constitutive Rules and Speech-Act Analysis." *The Journal of Philosophy* 68(13): 385–400, doi:10.2307/2025037.

SEARLE, John R. 1995. *The Construction of Social Reality*. London: The Free Press.

—. 2003. "Reply to Barry Smith (2003)." *American Journal of Economics and Sociology* 62(1): 299–309, doi:10.1111/1536-7150.t01-1-00012.

—. 2005. "What Is an Institution?" *Journal of Institutional Economics* 1(1): 1–22, doi:10.1017/S1744137405000020.

—. 2010. *Making the Social World: The Structure of Human Civilization*. Oxford: Oxford University Press, doi:10.1093/acprof:osobl/9780195396171.001.0001.

SMIT, J. P., BUEKENS, Filip and DU PLESSIS, Stan. 2011. "What is Money? An Alternative to Searle's Institutional Facts." *Economics and Philosophy* 27(1): 1–22, doi:10.1017/s0266267110000441.

—. 2014. "Developing the Incentivized Action View of Institutional Reality." *Synthese* 191(8): 1813–1830, doi:10.1007/s11229-013-0370-5.

—. 2016. "Cigarettes, Dollars and Bitcoins—an Essay on the Ontology of Money." *Journal of Institutional Economics* 12(2): 327–347, doi:10.1017/S1744137415000405.

680 Smith, Barry. 2003. "The Ontology of Social Reality." *American Journal of Economics*
681 *and Sociology* 62(1): 283–299, doi:10.1111/1536-7150.t01-1-00012.
682 Turner, Stephen P. 2010. *Explaining the Normative*. Cambridge: Polity Press.
683 Wilson, Robert A. 2007. "Social Reality and Institutional Facts: Sociality Within and
684 Without Intentionality." in *Intentional Acts and Institutional Facts: Essays on John*
685 *Searle's Social Ontology*, edited by Savas L. Tsohatzidis, pp. 139–153. Dordrecht:
686 Springer, doi:10.1007/978-1-4020-6104-2_6.

The Minimalist Theory of Truth and the Generalisation Problem

Li Zhang & Leon Horsten

The Minimalist Theory of Truth must show how it can prove certain truth-involving generalisations. Horwich has proposed two solutions to this challenge over the past decades. The first of these invokes Hilbert's ω-rule and is unacceptable. The second proposal can naturally be viewed in different ways. We show how this second proposal is naturally interpreted as a suggestion to solve the truth generalisation problem using uniform reflection rules. We also argue that this is indeed the right way for Horwich to respond to the truth generalisation problem.

Over the past three decades, Horwich's *minimalism* has been the most discussed deflationary truth theory. Generally speaking, this theory claims that everything about truth can be explained by the collection of underived and unproblematic instances of the equivalence schema

⟨ES⟩ $\langle p \rangle$ is true iff p.

In the literature, the equivalence schema ⟨ES⟩ is also known as the Tarski-schema or T-schema; its instances are known as Tarski-biconditionals or T-sentences. The theory consisting of all underived, unproblematic Tarski-biconditionals, namely, the theory taking all such biconditionals to be *axioms*, is called the "*Minimalist Theory of Truth*" (MT).

Firstly, Horwich believes that truth is *non-substantial*, so we should not define truth with any *substantial* concept. Instead, the meaning of 'is true' is given by the collection of underived, unproblematic instances of the T-schema. Horwich believes that

> our understanding of 'is true'—our knowledge of its meaning—consists in the fact that the explanatorily basic regularity in our use of it is the inclination to accept instantiations of the schema

(E) 'the proposition that p is true if and only if p',

by declarative sentences of English (including any extensions of English). (Horwich 1998, 35, emphasis omitted)

Due to its non-substantiality, truth should remain neutral in debates in other philosophical and non-philosophical areas (Horwich 1998, 52).

Secondly, Horwich argues that MT alone suffices to explain all the truth-involving phenomena (Horwich 1998, 5). Thus, for instance, without equating truth with utility or any other substantial concept, MT suffices to explain that "true beliefs help us to achieve practical goals" (Horwich 1998, 44). In what follows, we denote the first point Horwich has made as *the neutrality thesis* and the second as the *adequacy thesis* of minimalism (Gupta 1993, 361).

Despite Horwich's clever arguments for the two minimalistic theses, many logicians and philosophers insist that Horwich's minimalism is far from correct since there are many truth-involving facts that cannot be explained by it. In particular, it cannot prove truth generalisations that we regard as acceptable. For instance, it is unclear how MT proves

(1) Every proposition of the form $p \rightarrow p$ is true,

or

(2) Every proposition is such that either it or its negation is true.

In fact, many believe that it is impossible for MT to prove sentences such as (1) and (2). In the literature, this problem is known as *the truth generalisation problem* (Halbach 2014, 57; Raatikainen 2005, 177). Horwich has formulated two proposals in response to this challenge, but, as they stand, neither of them decisively answers the problem. We will defend an amplification and extension of Horwich's second proposal and argue that this successfully tackles the truth generalisation problem within the framework of truth-theoretic minimalism.

The structure of this paper is as follows: in section 1, we reformulate Horwich's minimalist truth theory in such a way that some unclarities of his original formulation are removed. In section 2, we show why MT and its modifications cannot prove intuitively acceptable truth generalisations. In section 3, we evaluate Horwich's two proposals in the light of critiques of them that have appeared in the literature. In section 4, we show how MT proves an ample collection of truth generalisations when strengthened with

uniform reflection rules, and we argue this to be in line with Horwich's second proposal. In section 5, we conclude this paper by suggesting that Horwich should accept our formulation of the reflection rules proposal since it coheres best with his other truth-theoretic theses.

1 Reformulating MT

It has been recognized that several aspects of the formulation of MT are unclear. In particular, it is not clear which Tarski-biconditionals belong to MT's axioms and which do not. Moreover, it is not clear how taking propositions as primary truth-bearers increases MT's proof-theoretic strength. Thus, we suggest two modifications of MT in this paper. First, by applying the T-schema to sentences *that themselves do not contain the truth predicate*, we obtain a precise description of MT's axioms. Second, we take sentences to be primary truth-bearers. Given these two modifications, MT is equivalent to the axiomatic truth theory TB (for "Tarski-biconditionals") when we take the Peano Arithmetic to be its base theory.[1]

One reason for MT's vagueness is Horwich's approach to truth-theoretic paradoxes. Horwich concedes that if some instances of the T-schema are included in MT's axioms, MT proves contradictions. He demonstrates this in a familiar way by applying the T-schema to the sentence

> THE PROPOSITION FORMULATED IN CAPITAL LETTERS IS NOT TRUE.

He argues that the only acceptable strategy for this problem is to exclude some instances of the T-schema from the axioms of MT (Horwich 1998, 40–41). The spirit of his approach to paradoxes has been shared by prominent logicians, including Tarski: by putting different constraints on the scope of the T-schema, we obtain different formal truth theories. These theories capture central uses of the truth predicate while, in the meantime, being adroit at avoiding truth-theoretic paradoxes. What renders Horwich's strategy different is that he does not give a specification of either the permitted or prohibited instances of the T-schema; he only requires that the collection of MT's axioms should be a maximally consistent set of sentences (Horwich 1998, 42). Unfortunately, McGee has shown there are uncountably many mutually incompatible sets

1 TB is also sometimes denoted as DT (for "disquotational theory") in the literature (Halbach 2014, 53).

doi: 10.48106/dial.v77.i1

that satisfy this requirement; none of them are recursively axiomatisable. Therefore, Horwich must impose more constraints on the instances of the T-schema (McGee 1992, 236–237).

TB is axiomatisable and consists of unproblematic Tarski-biconditionals, which renders it a suitable substitute for MT. However, far be it from us to claim that TB is the *only* suitable substitute for MT. Many natural axiomatic disquotational theories of truth would do just as well. For instance, if one would substitute variants of Halbach's theory of *Positive Tarski-biconditionals* (2009) for MT instead, then the arguments of the present article would still go through.

Most logicians who are interested in formal truth theories, such as Tarski (1944, 342), McGee (1992, 235), Halbach (2014, 12), and Cieśliński (2018, 1083, n.8), take sentences to be primary truth-bearers. The reasons for their choice are quite straightforward: propositions are ill-understood and controversial, whereas we have rigorous and widely accepted syntactical theories of sentences.

Horwich nonetheless insists on formulating minimalism in terms of propositions because he believes that there exist propositions that cannot be expressed by current human languages:

> Patrick Grim pointed out to me that the minimal theory cannot be regarded as *the set* of propositions of the form $\langle \langle p \rangle$ is true iff $p \rangle$; for there is no such set. The argument for this conclusion is that if there were such a set, then there would be distinct propositions regarding *each* of its subsets, and then there would have to be distinct axioms of the theory corresponding to those propositions. Therefore there would be a $1-1$ function correlating the subsets of MT with some of its members. But Cantor's diagonal argument shows that there can be no such function. Therefore, MT is not a set. In light of this *result* [our emphasis], when we say things like "$\langle A \rangle$ follows from the minimal theory," we must take that to mean, not that the relation of *following from* holds between $\langle A \rangle$ and a certain entity, the minimal theory; but rather that it holds between $\langle A \rangle$ and *some part* of the minimal theory—i.e., between $\langle A \rangle$ and some set of propositions of the form $\langle \langle p \rangle$ is true iff $p \rangle$. (Horwich 1998, 20–21, n.4)

The particular argument of Grim that is alluded to here goes as follows (see Grim 1988): Suppose there were a set S of all truths, and consider all subsets

of S, i.e., all members of the power set $\mathcal{P}(S)$. To each element of this power set will correspond a truth. To each element of the power set, for example, a particular truth p either will or will not belong as a member. In either case, we will have a truth: that p is a member of that element or that it is not. There will then be at least as many truths as there are elements of the power set $\mathcal{P}(S)$. But by Cantor's theorem, we know that the power set of any set will be larger than the original. There will then be more truths than there are members of S, and for any set of truths S, there will be some truth left out. There can therefore be no set of all truths.

The quotation by Horwich shows that he regards Grim's argument as definitive: he takes the conclusion of the argument as a philosophical *result*. But it is far from clear whether, in the absence of a detailed, widely accepted theory of propositions and their constituents, Grim's argument is persuasive. To give but one example of a worry that one might have here,[2] observe that Grim's succinct argument presupposes that for each subset B of S, there exists a *proposition* of the form $p \in S\,(\neg p \in S)$. For this to be the case, for each such subset B, there has to be an individual concept of B as a part of this proposition. But whether all such individual concepts exist is a substantial and unsettled philosophical question.

In view of this, there seems to be no pressing need for Horwich to take truth to be a property of propositions. Nonetheless, we do not ask of Horwich that he abandon his views of the kinds of entities that are the bearers of truth. A sub-class of the totality of all propositions is the *set* of all propositions that can be expressed by sentences belonging to some fixed language. The theory of true *sentences* (of some language) can then be seen as a special case of Horwich's more general theory of truth of propositions. So the argument that is developed in the subsequent sections intends to support the thesis that, *as far as truth of (propositions expressed by) sentences goes*, Horwich's arguments of the early 2000s concerning truth generalisations were at least a decade ahead of their time, albeit not fully fleshed out. That Horwich might well be sympathetic to such an interpretation of his views concerning truth generalisations is indicated by the passage in his *Truth* book, where he says that

> ordinary language suggests that truth is a property of propositions, and that utterances, beliefs, assertions, etc., inherit their truth-

2 We do not have the space to go deeply into the literature that has been generated by Grim's argument.

doi: 10.48106/dial.v77.i1

> like character from their relationship to propositions. However, [previous considerations] show that this way of seeing things has no particular explanatory merit. The truth-like conception for each type of entity is equally minimalistic. And by assuming any one of them we can easily derive the others. (Horwich 1998, 102)

2 The Truth Generalisation Problem

A non-trivial general claim of the form "*Every* x is φ" cannot be proved by a finite collection of premises, each of which asserts that a_i is φ, for some i, except if there is an additional premise that says that every object is one of this finite number of a_i's. This also applies to MT in the sense that a general claim of the form "For *every* sentence x of the form $p \rightarrow p$, x is true," for example, cannot be proved in MT. Indeed, it has been *proved* that such a truth generalisation cannot be proved in TB (= MT) (Halbach 2014, 56–57). Many such truth generalisations appear to be conceptual truths about the concept of truth. In particular, this is so for the classical compositional axioms of truth that state that truth commutes with the logical connectives. Moreover, there are valid philosophical and natural language arguments whose *validity* depends not just on Tarski-biconditionals but also on compositional truth axioms (Fujimoto 2022; Horsten 2010).

This poses a challenge to the minimalist theory of truth: recall that MT's adequacy thesis claims that *all* facts whose expression involves the truth predicate can be explained by assuming no more about truth than instances of the equivalence schema (Horwich 1998, 23). A number of philosophers and logicians, including Armour-Garb (2010, 698), Gupta (1993, 363–364), Halbach (2001, 1959–1960), and Soames (1997, 30–31), regard its inability to prove truth generalisations as a serious defect of MT.

One may be tempted to appeal to "McGee's trick" (McGee 1992, 238) and contend that since it is always possible to find a T-sentence that is equivalent to a given truth generalisation, when MT is not identified with TB but instead with TB plus such additional Tarski-biconditionals, MT is capable of proving all acceptable truth generalisations. Indeed, by the diagonal lemma, for every truth generalisation A, there is a sentence κ such that

$$\vdash \kappa \leftrightarrow (T(\kappa) \leftrightarrow A).$$

By associativity of $\leftrightarrow$, the Tarski-equivalence $\kappa \leftrightarrow T(\kappa)$ is provably equivalent to A, where A is an acceptable truth generalisation. However, it is widely ac-

cepted that sentences such as that expressed by κ should not be allowed in ⟨ES⟩ since, by exactly the same procedure, it is possible to find a T-sentence equivalent to "Santa Claus exists," which should not follow from any acceptable truth theory.

In sum, the generalisation problem poses a serious challenge for Horwich's truth theory.

3 Horwich's Responses

It is not clear exactly when Horwich came to the conclusion that MT cannot prove acceptable truth generalisations. But it is clear he wants to resolve this problem by strengthening MT with further theoretical resources. Moreover, it is possible to group his many responses into two categories: the *ω-rule proposal* and a *reflection-based proposal*. We review these proposals in turn.

3.1 *Horwich's First Attempt: The ω-Rule*

In the postscript of the revised *Truth*, Horwich formulates his first attempt at solving the truth generalisation problem. There, he writes:

> However, it seems to me that in the present case, where the topic is *propositions*, we can find a solution to this problem. For it is plausible to suppose that there is a truth-preserving rule of inference that will take us from a set of premises attributing to each proposition some property, *F*, to the conclusion that all propositions have *F*. No doubt this rule is not *logically* valid, for its reliability hinges not merely on the meanings of the logical constants, but also on the nature of propositions. But it is a principle we do find plausible. We commit ourselves to it, implicitly, in moving from the disposition to accept any proposition of the form "*x* is *F*" (where *x* is a proposition) to the conclusion "All propositions are *F*." So we can suppose that this rule is what sustains the explanations of the generalizations about truth with which we are concerned. Thus we can, after all, defend the thesis that the basic theory of truth consists in some subset of the instances of the equivalence schema. (Horwich 1998, 137–138)

doi: 10.48106/dial.v77.i1

It has been acknowledged that the above-mentioned truth-preserving rule amounts to a form of the ω-rule (Raatikainen 2005, 175). Hilbert introduces this principle in the following manner:

> If it has been proved, for any given numeral δ, that the formula
>
> $$\mathfrak{A}(\delta)$$
>
> is always a correct numerical formula, then the formula center
>
> $$(x)\mathfrak{A}(x)$$
>
> can be laid down as a starting formula [Ausgangsformel]. (Hilbert 1931, 1154)

Feferman rightly observed that Hilbert's own formulation of the ω-rule is somewhat vague (Feferman 1986, 212). The ω-rule is perhaps more clearly expressed as: "From infinitely many premises $\varphi(0)$, $\varphi(1)$, ... that result from replacing the numerical variable n in $\varphi(n)$ with the numeral for each natural number, conclude $\forall x\varphi(x)$" (Hazen 1998).

The ω-rule is a strong rule: When enriched with this rule, PA proves true arithmetic (Hazen 1998). With regard to the generalisation problem, when augmented with the ω-rule, MT is able to prove all acceptable truth generalisations. Take a finite first-order language as an example; every sentence of the form $p \rightarrow p$ is a theorem of this language. Enumerate all sentences of the form $p \rightarrow p$, so each of them is represented by a numeral. Apply T-sentences of MT to them, so for each n, $T(n)$. By the ω-rule, we obtain the general claim $\forall x T(x)$.

However, certain features of the ω-rule render this proposal problematic and, in particular, unacceptable to the minimalist truth theory. Raatikainen has argued that we, as finite human beings, cannot take infinitely many premises into consideration simultaneously. Therefore, even if the theory MT + the ω-rule is capable of proving acceptable truth generalisations, those generalisations are beyond the reach of ordinary human beings (Raatikainen 2005, 176). This problem with the ω-rule cannot be overcome: it simply has no effective (read: recursively enumerable) equivalent. Moreover, the proof-theoretic strength of the ω-rule makes it specifically unacceptable to the minimalist truth theory. When enriched with this rule, Peano Arithmetic proves all true arithmetic sentences. True arithmetic is not axiomatisable, while MT is intended to be an axiomatised truth theory.

It is not clear whether or not Horwich has accepted critiques of his first proposal. In a recent publication, Horwich still seems to propose using the ω-rule as a solution to the truth generalisation problem:

> For it is plausible to suppose that there is a truth-preserving rule of inference that will take us from a set of premises attributing to each proposition of a certain form some property, G, to the conclusion that the *all* proposition have property G. And this rule—not *logically* valid, but nonetheless necessarily truth-preserving given the nature of proposition—enables the general facts about truth to be explained by their instances. (Horwich 2003, 84, n.14, cited after reprint)

Yet, in most of his recent writings, Horwich advocates an alternative resolution based on an introspective process. To this proposal, we now turn.

3.2 *Horwich's Second Attempt: Reflection*

Over the years, Horwich's formulation of his second proposal has varied, and it is not easy to select a preferred formulation from these variants. Extant critiques of his various formulations are indecisive. Nonetheless, we will argue that all variants of Horwich's second proposal need emendation in order to solve the truth generalisation problem.

A first formulation of Horwich's second attempt emerges in (2001):

> Whenever someone can establish, for any *F*, that it is *G*, *and recognizes that he can do this*, then he will conclude that every *F* is *G*. (Horwich 2001, 157)

Call this *Solution 2.0*. This solution also consists in adding an additional rule of inference to MT. But the additional rule of inference of Solution 2.0 is different from the ω-rule.

In a revised version (2010) of the same paper, Horwich formulates a variant of this new proposal, which in effect amounts to a further, *substantially different* proposal:

> Whenever someone is disposed to accept, for any proposition of structural type F, that it is G (and to do so for uniform reasons) then he will be disposed to accept that every F-proposition is G. (Horwich 2010, 45)

doi: 10.48106/dial.v77.i1

To the above statement, he adds the following *proviso*:

> We cannot conceive of there being additional Fs—beyond those Fs we are disposed to believe are G—which we would not have the same sort of reason to believe are Gs. (Horwich 2010, 44)

Call the proposal that is encapsulated in the previous two quotations *Solution 2.1*. (Horwich endorses this same solution in 2003, 84.)

Armour-Garb argues that Solution 2.1 is unsatisfactory because

> one will not be disposed to accept (the proposition) that all F-propositions are G, from the fact that, for any F-proposition, she is disposed to accept that it is G (NB, even for uniform reasons), unless she is *aware* of the fact that, for any F-proposition, she is disposed to accept that it is G. (Armour-Garb 2010, 699)

The *proviso* Horwich added to Solution 2.1 does not provide such an awareness component. It merely adds a *negative* condition ("not being able to conceive of there being F's that are not G"), while Armour-Garb's awareness requirement is a positive condition. Nonetheless, Solution 2.0 incorporates exactly the awareness condition that Armour-Garb insists on ("and recognises that he can do this").

Armour-Garb is making a psychological observation here, but there is an accompanying *rational* point to be made also. If one does not *recognise* that for any *F*-proposition, she is disposed to accept that it is *G*, then she is not, without further ado, *rationally required* to believe that every *F*-proposition is *G*. *Ought* implies *can*, and, in this situation, she simply lacks the ground for accepting that every *F*-proposition is *G*.[3] For this reason, Horwich's Solution 2.0 must be regarded as superior to his Solution 2.1.

Nonetheless, Armour-Garb would not be satisfied with Solution 2.0 either. He argues that the switch, in the move from the premise to the conclusion of the rule of inference in Solution 2.1, of "for any F-proposition" from outside the "disposed to accept"-context to inside the "disposed to accept"-context, is "viciously circular." He is certainly right that this quantifier shift, which is also present in Solution 2.0, is not derivable in classical logic. Nonetheless, we take issue with this aspect of Armour-Garb's critique of Horwich's second proposal. Indeed, we agree with Cieśliński that Armour-Garb's dismissal of

3 Further discussion of these important matters can be found in Horsten (2021).

Horwich's second solution on the grounds of its being viciously circular is "hasty" (Cieśliński 2018, 1082): we will come back to this later.

It is time to spell out the content of Horwich's Solution 2.0, i.e., the first quotation in this section, in more precise terms. We do this by formalising Horwich's informally expressed—and somewhat vague—rule in first-order logic. In our formalisation of the first quotation in this section, we want to be charitable to Horwich. We do not claim that Horwich would agree with our formalisation (Horwich can speak for himself), but we will argue that there are good reasons for him to do so. Firstly, Solution 2.0 contains the phrase "*will* conclude," making it seem like a psychological prediction.[4] If it is taken in this way, then whether it is true or not is an empirical matter. But this is presumably not what Horwich intends. Rather, what he means is that the agent will be disposed to draw this conclusion *if she is rational*. In other words, Horwich purports to propose a rational *rule of inference* here. So it might be better to replace, in Solution 2.0, "will conclude" with "may (rationally) conclude," or perhaps even "should (rationally) conclude." Secondly, since we are concerned with *establishing* truth generalisations, we identify the concepts "being disposed to accept" and "recognising" with being *provable*. In particular, we interpret the clause "*and recognizes that he can do this*" as *de re* provability of an arbitrary *F that* it is *G*. Thirdly, we identify provability with provability in the background theory, which is *MT*. If we were to identify provability with provability in the system *including the rule*, then the proposed rule would indeed be viciously circular, confirming Armour-Garb's (unfounded) suspicions. But if we identify provability with provability in *MT*, then there is no circularity. Fourthly, we *omit* the concept of provability ("being disposed to accept") from the conclusion of the rule. With these precisifications in place—which we take to be reasonable, but we leave it open whether they are *exactly* in accordance with what Horwich intended—we obtain the following schematic rule:[5]

$$\frac{\vdash \forall x : F(x) \to Bew_{\mathrm{MT}}(G(x))}{\vdash \forall x : F(x) \to G(x)}.$$

We will call this rule H (for: "Horwich"). Observe that, unlike the ω-rule, H is an *effective* rule: adding it to MT yields an axiomatic system.

Worries based on the lottery paradox might cause one to doubt the rationality of rule H. For any ticket (in a large, fair lottery), I believe that it is

4 Cieśliński sees this as the main weakness of Horwich's recent views; see Cieśliński (2017, 80).

5 In the interest of readability, we are sloppy with Gödel coding here as well as later on in this article.

doi: 10.48106/dial.v77.i1

not the winning ticket (and I believe this for "uniform reasons"). But from this, I am not prepared to infer that every ticket is a losing ticket (Kyburg 1970, 56). Nonetheless, such a worry would be ill-founded, for the situation under consideration is different in one key respect. The irrationality of the lottery paradox inference stems from the fact that many small but non-zero probabilities (of being the winning ticket) can add up to a large probability (of one of a large collection of tickets being the winning one). But what is provable has probability 1 rather than $1-\epsilon$ (for some small ϵ) since provability in a sound system from necessary premises is itself necessary, and necessary truths by a Kolmogorov axiom for probability receive probability 1. So, the fair lottery phenomenon is irrelevant to the evaluation of rule H.[6]

4 Uniform Reflection and Truth Generalisations

We have seen that Horwich recognises that H is not an admissible inference rule of first-order logic. The main questions that we want to answer in this section about H are the following: *To what extent and in which way does adding H to MT allow us to prove truth generalisations?* Moreover: *Is H a rational rule of inference?*

4.1 *H and Uniform Reflection*

It is clear that given a *sound* theory S, adding H (with Bew_{MT} replaced by Bew_{S}) to S results in a sound system. So, in particular, MT + H is a sound system.

Next, we make the crucial observation that H is equivalent to a reflection rule that has been intensively investigated in proof theory. To this end, we first recall the notion of *uniform reflection principle* for a theory S (denoted as $RFN(S)$),

$$\forall x : Bew_{\mathrm{S}}(\varphi(x)) \rightarrow \varphi(x),$$

and the notion of *uniform reflection rule* for a theory S (denoted as UR_S),

$$\frac{\vdash \forall x : Bew_{\mathrm{S}}(\varphi(x))}{\vdash \forall x : \varphi(x)}.$$

6 An extended discussion of the relevance or irrelevance of the lottery paradox in this context can be found in Cieśliński (2017, sec. 13.5).

Feferman has proved the remarkable little fact that RFN(S) is equivalent to UR_S (Feferman 1962, Theorem 2.19). In light of this, it is easy to see that H is equivalent to UR_{MT} (and therefore also to RFN(MT)): the $\Rightarrow$-direction is obvious, and the $\Leftarrow$-direction follows immediately from Feferman's theorem.

At this point, a connection with Horwich's *first* solution also becomes apparent. Indeed, the uniform reflection rule is widely seen as an effective version (a "tamed" version) of the ω-rule. Horwich's appeal to the ω-rule was (rightly) rejected by Raatikainen on account of its non-effectiveness. Uniform reflection rules cannot be rejected on the same grounds.

We will now see how the main observation of this subsection allows us to answer the question to what extent H enables us to prove truth generalisations.

4.2 *Deriving Truth Generalisations*

Let us denote MT + H as MT_1. Now that we have made Horwich's Solution 2.0 precise, we address the question whether MT_1 can prove all intuitively acceptable truth generalisations. An apparent counterexample is a proposition such as "There are as many truths as there are untruths" (Gupta 1993, 363). But this is a second-order statement involving not just sentences but also *sets* of sentences. So, it falls outside the scope of MT (= TB), which cannot even express *claims* involving sets of sentences.

The truth theory that takes the axioms that state that truth commutes with the logical connectives for sentences that do not themselves contain the notion of truth is called *CT*. It is fairly generally accepted that in CT, a vast amount of intuitively acceptable truth generalisations logically follow (Horsten 2011, chap. 6). So if Horwich can derive the truth axioms of CT, then he has made significant progress towards solving the truth generalisation problem. Nonetheless, it would be an exaggeration to say that *all* intuitively acceptable truth generalisations are provable in CT:[7] the truth generalisation "All arithmetical theorems of CT are true," for instance, is not provable even in CT.

With only one exception, the compositional truth axioms of CT can indeed be derived in MT_1 (Horsten and Leigh 2017). As an example, let us consider the compositional axiom for negation:

$$\forall x \in \mathcal{L}_{PA} : T(\neg x) \leftrightarrow \neg Tx.$$

7 Thanks to an anonymous referee for making this point.

doi: 10.48106/dial.v77.i1

Every *instance* of this axiom can be proved in TB (using Tarski-biconditionals). Moreover, *that* every instance can be proved in TB can be uniformly recognised (i.e., proved) as a combinatorial fact even in the background theory PA. So we have:

$$PA \vdash \forall x \in \mathcal{L}_{PA} : Bew_{\mathrm{MT}}(T(\neg x) \leftrightarrow \neg Tx).$$

Then, by $\mathrm{UR}_{\mathrm{MT}}$, we indeed obtain $\forall x \in \mathcal{L}_{PA} : T(\neg x) \leftrightarrow \neg Tx$.

The other compositional axioms can be derived in a similar way in MT_1, with the sole exception of the quantifier axiom:

$$\forall \varphi(x) \in \mathcal{L}_{PA} : T(\forall x \varphi(x)) \leftrightarrow \forall x T\varphi(x).$$

We cannot prove in MT, for every $\varphi(x) \in \mathcal{L}_{PA}$, that $T(\forall x \varphi(x)) \leftrightarrow \forall x T\varphi(x)$. The reason is that TB (= MT) only contains Tarski-biconditionals for *sentences*, i.e., for *closed* formulas. In order to prove, for each $\varphi(x) \in \mathcal{L}_{PA}$, that $T(\forall x \varphi(x)) \leftrightarrow \forall x T\varphi(x)$, we need a slight strengthening of the Tarski-biconditionals of TB, namely, the *uniform* arithmetical Tarski-biconditionals, which are the sentences of the form $\forall x(T\varphi(x) \leftrightarrow \varphi(x))$ for formulas $\varphi(x) \in \mathcal{L}_{PA}$. The resulting slight strengthening of TB is called *UTB*.

How do we derive these uniform Tarski-biconditionals? We can prove them in MT_1 as follows (Horsten and Leigh 2017, Theorem 9).[8] Every instance of a given uniform (arithmetical) Tarski-biconditional can be proved in TB. This combinatorial fact can again be proved even in PA:

$$PA \vdash \forall x \in \mathcal{L}_{PA} : Bew_{\mathrm{MT}} : T\varphi(x) \leftrightarrow \varphi(x).$$

So, by applying $\mathrm{UR}_{\mathrm{MT}}$ in MT_1 to this fact, we obtain the result. Now, in a second stage, we can proceed as we did with the negation axiom. But to carry out this proof, we need to appeal to $\mathrm{UR}_{\mathrm{MT}_1}$, which is the uniform reflection rule for MT_1:

$$\frac{\vdash \forall x : Bew_{\mathrm{MT}_1}(\varphi(x))}{\vdash \forall x : \varphi(x)},$$

where φ can be any arithmetical formula, and Bew_{MT_1} formally expresses provability in MT_1. For the same reasons as why $\mathrm{UR}_{\mathrm{MT}}$ exceeds MT, the rule

8 Theorem 9, obtained in Horsten and Leigh (2017), is based on uniform reflection principles rather than rules, but we have seen above that by an argument due to Feferman, the two are provably equivalent.

$\mathrm{UR}_{\mathrm{MT}_1}$ exceeds MT_1. If we apply this inference rule to the earlier obtained fact that PA proves

$$\forall x \in \mathcal{L}_{PA} : Bew_{\mathrm{MT}_1}(T(\forall x \varphi(x)) \leftrightarrow \forall x T\varphi(x)),$$

then we obtain the desired result that $\forall x \in \mathcal{L}_{PA} : T(\forall x \varphi(x)) \leftrightarrow \forall x T\varphi(x)$.

In sum, we can prove all the compositional truth axioms of CT and, therefore, many intuitively acceptable truth generalisations in $\mathrm{MT}_2 = \mathrm{MT}_1 + \mathrm{UR}_{\mathrm{MT}_1} = \mathrm{MT} + \mathrm{UR}_{\mathrm{MT}} + \mathrm{UR}_{\mathrm{MT}+\mathrm{UR}_{\mathrm{MT}}}$ (Horsten and Leigh 2017). In other words, many truth generalisations follow from two iterations of uniform reflection on MT. Even more truth generalisations can be proved when this strategy is iterated further. By adding further uniform reflection principles to MT_2, for instance, also the truth generalisation "All arithmetical theorems of CT are true" becomes provable.

At this point, we see that we have to go slightly beyond our charitable interpretation of Horwich's Solution 2.0. Horwich claims that *one* level of reflection on MT suffices to prove all acceptable truth generalisations. We now see that *two* levels of reflection on MT are required. Given the equivalence between Horwich's rule H and Feferman's uniform reflection rule, all acceptable truth generalisations can be derived in the theory MT + H + H′, where H′ is just like H, except that its background theory is MT + H instead of MT:

$$\frac{\vdash \forall x : F(x) \rightarrow Bew_{\mathrm{MT}+\mathrm{H}}(G(x))}{\vdash \forall x : F(x) \rightarrow G(x)}.$$

In sum, if H and H′ are *rational* rules of inference, then Horwich was very much on the right track.

4.3 *Rationality*

Uniform reflection rules are rules that contain the required "awareness" component in the antecedent (the agent has to have a proof) and that are also, *pace* Armour-Garb, not circular in any way. In addition, in the premise of uniform reflection rules, the awareness/recognition component that is required is *proof* from the Tarski-biconditionals.

On our interpretation and emendation of his view, Horwich contends that it is *rational* to add $\mathrm{UR}_{\mathrm{MT}}$ and $\mathrm{UR}_{\mathrm{MT}_1}$ to MT. With this, he would not be alone. In his work on *implicit commitment*, Feferman claimed that if an agent explicitly accepts a theory S, then she also ought to accept uniform reflection principles

doi: 10.48106/dial.v77.i1

and rules for S, such as $\mathrm{UR_S}$ and $\mathrm{UR_{S+UR_S}}$ (Feferman 1991, 2, 44). Acceptance of $\mathrm{UR_S}$ is, in his view, *implicit* in the acceptance of S, and acceptance of $\mathrm{UR_{S+UR_S}}$ is "implicitly implicit" in the acceptance of S.

Feferman did not give an epistemological argument for *why*, if one accepts a theory S, one should also accept $\mathrm{UR_S}$ (and $\mathrm{UR_{S+UR_S}}$). A recent attempt to provide such an argument is given by Fischer in (2023), which can, in retrospect, be seen as one attempt to develop Horwich's Solution 2.0 in detail. A discussion of Fischer's argument is outside the scope of this article. Here, we restrict ourselves to a few remarks on the issue. The uniform reflection rule for the theory that one is currently working in expresses a form of trust or confidence in this theory. If the theory one is working in is justified, then this trust is also justified, and therefore, accepting the uniform reflection rule is justified. The theory that is relevant in the present context is the truth theory MT. Horwich argues that this theory is indeed justified because Tarski-biconditionals express the content or meaning of the concept of truth (Horwich 2010, 17). Therefore, making one's trust in MT explicit by accepting $\mathrm{UR_{MT}}$ and $\mathrm{UR_{MT_1}}$ is rational.[9] Since, by Feferman's theorem, H is equivalent to $\mathrm{UR_{MT}}$, and H′ is equivalent to $\mathrm{UR_{MT_1}}$, H and H′ are, therefore, also rational inference rules.

5 Horwich Vindicated?

There have been two phases in the history of truth-theoretic deflationism. In the first phase, disquotational axioms were taken to express the full content of the concept of truth. This phase comprises, a.o., Quine's views on truth as a tool for semantic ascent and descent (Quine 1970, 10–13) and the prosentential theory of truth (Grover, Camp and Belnap 1975). Horwich's minimalism is often viewed as a late and particularly bright exponent of this phase of deflationism. In the second phase, compositional axioms were taken to express basic properties of the concept of truth. This phase started sometime in the 1980s, partly under the influence of Davidson's truth-conditional compositional approach to natural language semantics (Davidson 1967). During much of this second phase, Horwich's views on the concept of truth came to be increasingly seen as dated and untenable. As a result of this, his writings about the generalisation problem after the first edition of his book *Truth* were mostly ignored by the truth-theoretic community.

9 Considerations such as these may provide at least the beginnings of a response to Cieśliński's complaint above (*confer supra*, page XX) that Horwich's theory is too psychological.

Perhaps we now experience the dawn of a third phase in the history of truth-theoretic deflationism, in which the relation between the concept of truth on the one hand and reflection principles on the other hand plays a major role. In particular, it is currently a hotly debated question whether by making use of reflection principles or rules, disquotationalism can solve the generalisation problem. We make no attempt to adjudicate this discussion here. But we have seen that Horwich anticipated the current philosophical debate already in the early 2000s. So rather than being a truth-theoretic dinosaur, at the time Horwich's views were ahead of their time—which, of course, does not mean that they are in any way definitive.

The main reason why Horwich's thoughts about the relation between reflection principles and truth generalisations were ignored is that Horwich's view about this problem was not completely precise and was connected to other views of his that can be separated from the problem at issue. Horwich was committed to propositions as the bearers of truth but did not give a precise theory of propositions. At the same time, he was also committed to the background disquotational theory as a maximal consistent collection of propositions, which prevents it from being recursively axiomatisable and, therefore, prevents it from being learnable. But we have seen that a *derived* notion of true proposition *expressible in a given language* makes perfect sense in Horwich's framework. Moreover, Horwich's requirement of MT being a maximal consistent collection of propositions is unrelated to his solution proposal to the generalisation problem and can, therefore, simply be rejected—which is exactly what the truth-theoretic community has largely done. In sum, Horwich's views from the early 2000s on the truth generalisation problem can be disentangled from the further commitments and unclarities with which he connected them.

The imprecision of his treatment of the generalisation problem prevented Horwich from working out the technical details with full precision. For instance, he did not see that *two* rounds of uniform reflection are needed in order to derive the compositional truth principles from the disquotational axioms. Nonetheless, Horwich did see that his strategy for dealing with the generalisation problem is in line with his two main minimalistic theses: the neutrality thesis and the adequacy thesis. Reflection rules are not truth-theoretic (or *philosophical*) but *mathematical* rules. Uniform reflection rules are universally seen as mathematical rules because they have substantial mathematical consequences: they are canonical ways of extending the mathematical strength of a theory. Therefore, strengthening MT with uniform reflection rules does

doi: 10.48106/dial.v77.i1

not affect the neutrality of the theory of *truth*. (Indeed, as mentioned earlier, MT can be taken to be proof-theoretically conservative over its background theory PA.) Moreover, since CT is derivable from MT by means of two rounds of uniform reflection, and CT proves the needed truth generalisations, a solution to the generalisation problem is reached, whereby the challenge to the adequacy thesis is answered.

The more recent debate about the connection between reflection principles and the truth generalisation problem developed only after 2015, and it developed largely independently from Horwich's views on the generalisation problem. Moreover, we now see further and more clearly in these matters than Horwich did around 2002. Yet it would be a mistake to take Horwich's early thoughts on this issue to be merely of historical relevance ("give credit where credit is due"). The appeal to proof-theoretic reflection principles and rules as a means to derive compositional truth axioms is sometimes seen as a mere "technical" manoeuvre. But Horwich, at the time, did not know any of the proof-theoretic literature concerning reflection principles and hit on the basic idea *in tempore non suspecto*. Purely by philosophically thinking about how to solve the generalisation problem in a disquotational framework, he, in one of his proposals (Solution 2.0), arrived at the view that the compositionality or truth follows by the uniform reflection rule from disquotational principles. This is simply amazing, and it shows that rather than being merely a technical trick, it is a very natural theoretical view to take.*

Li Zhang
0000-0002-7766-7263
Tsinghua University
l-zhang17@mails.tsinghua.edu.cn

Leon Horsten
0000-0003-3610-9318
University of Konstanz
Leon.Horsten@uni-konstanz.de

* The authors are indebted to the members of the Antos-Horsten doctoral seminar at the University of Konstanz, where an early version of this article was presented, for helpful and thoughtful comments. They also thank two anonymous referees, whose comments have led to substantial improvements of this article. The first author's research for this article was supported by a Chinese Scholarship Council Fellowship (CSC No. 201906210194) for carrying out doctoral research at the University of Bristol and at the University of Konstanz. Without the support of this scholarship, the present article could not have been written.

References

ARMOUR-GARB, Bradley. 2010. "Horwichian Minimalism and the Generalization Problem." *Analysis* 70(4): 693–703, doi:10.1093/analys/anq073.

BEALL, J. C. and ARMOUR-GARB, Bradley, eds. 2005. *Deflationism and Paradox*. Oxford: Oxford University Press, doi:10.1093/oso/9780199287116.001.0001.

CIEŚLIŃSKI, Cezary. 2017. *The Epistemic Lightness of Truth: Deflationism and Its Logic*. Cambridge: Cambridge University Press, doi:10.1017/9781108178600.

—. 2018. "Minimalism and the Generalisation Problem: On Horwich's Second Solution." *Synthese* 195(3): 1077–1101, doi:10.1007/s11229-016-1227-5.

DAVIDSON, Donald. 1967. "Truth and Meaning." *Synthese* 17(1): 304–323. Reprinted in Davidson (1984, 17–36), doi:10.1007/BF00485035.

—. 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press, doi:10.1093/0199246297.001.0001.

EWALD, William Bragg, ed. 1996. *From Kant to Hilbert: A Source Book in the Foundations of Mathematics. Volume II*. Oxford: Oxford University Press, doi:10.1093/oso/9780198505365.001.0001.

FEFERMAN, Solomon. 1962. "Transfinite Recursive Progressions of Axiomatic Theories." *The Journal of Symbolic Logic* 27(3): 259–316, doi:10.2307/2964649.

—. 1986. "Introductory Note to 1931c [Gödel (1931)]." in *Collected Works, Volume I: Publications 1929–1936*, pp. 208–212. Oxford: Oxford University Press. Edited by Solomon Feferman, John W. Dawson, Jr., Stephen C. Kleene, Gregory H. Moore, Robert M. Solovay and Jean van Heijenoort, doi:10.1093/oso/9780195147209.003.0019.

—. 1991. "Reflecting on Incompleteness." *The Journal of Symbolic Logic* 56(1): 1–49, doi:10.2307/2274902.

FISCHER, Martin H. 2023. "Another Look at Reflection." *Erkenntnis* 88(2): 479–509, doi:10.1007/s10670-020-00363-9.

FUJIMOTO, Kentaro. 2022. "The Function of Truth and the Conservativeness Argument." *Mind* 131(521): 129–157, doi:10.1093/mind/fzaa083.

GÖDEL, Kurt. 1931. "Besprechung von Hilbert (1931)." *Zentralblatt für Mathematik und ihre Grenzgebiete* 1: 260. English translation in van Heijenoort (1967, 596–616); reprinted in Gödel (1986, 213/214).

—. 1986. *Collected Works, Volume I: Publications 1929–1936*. Oxford: Oxford University Press. Edited by Solomon Feferman, John W. Dawson, Jr., Stephen C. Kleene, Gregory H. Moore, Robert M. Solovay and Jean van Heijenoort, doi:10.1093/oso/9780195147209.001.0001.

GRIM, Patrick. 1988. "Logic and Limits of Knowledge and Truth." *Noûs* 22(3): 341–367, doi:10.2307/2215708.

GROVER, Dorothy L. 1992. *A Prosentential Theory of Truth*. Princeton, New Jersey: Princeton University Press, doi:10.1515/9781400862689.

GROVER, Dorothy L., CAMP, Joseph L., Jr. and BELNAP, Nuel D., Jr. 1975. "A Prosentential Theory of Truth." *Philosophical Studies* 27(2): 73–125. Reprinted as Grover (1992, ch. 3), doi:10.1007/bf01209340.

GUPTA, Anil. 1993. "Minimalism." in *Philosophical Perspectives 7: Language and Logic*, edited by James E. TOMBERLIN, pp. 359–369. Atascadero, California: Ridgeview Publishing Co., doi:10.2307/2214129.

HALBACH, Volker. 2001. "Disquotational Truth and Analyticity." *The Journal of Symbolic Logic* 66(4): 1959–1973, doi:10.2307/2694987.

—. 2009. "Reducing Compositional to Disquotational Truth." *The Review of Symbolic Logic* 2(4): 786–798, doi:10.1017/s1755020309990220.

—. 2011. *Axiomatic Theories of Truth*. 1st ed. Cambridge: Cambridge University Press, doi:10.1017/cbo9780511921049.

—. 2014. *Axiomatic Theories of Truth*. 2nd ed. Cambridge: Cambridge University Press. First edition: Halbach (2011), doi:10.1017/cbo9781139696586.

HAZEN, Allen Patterson. 1998. "Non-Constructive Rules of Inference." in *The Routledge Encyclopedia of Philosophy*, edited by Edward J. CRAIG. London: Routledge. The Routledge Encyclopedia was made available online in 2002 and is now regularly updated, doi:10.4324/9780415249126-Y014-1.

VAN HEIJENOORT, Jan, ed. 1967. *From Frege to Gödel: A Source Book in Mathematical Logic 1879–1931*. Cambridge, Massachusetts: Harvard University Press.

HILBERT, David. 1931. "Die Grundlegung der elementaren Zahlenlehre." *Mathematische Annalen* 104(1): 485–494. Translated as "The Grounding of Elementary Number Theory" in Ewald (1996, 1148–1156), doi:10.1007/BF01457953.

HORSTEN, Leon. 2010. "On a Necessary Use of Truth in Epistemology." in *The Analytical Way: Proceedings of the 6th European Congress of Analytic Philosophy*, edited by Tadeusz CZARNECKI, Katarzyna KIJANIA-PLACEK, Olga POLLER, and Jan WOLEŃSKI, pp. 371–376. Studies in Logic n. 25. London: College Publications.

—. 2011. *The Tarskian Turn: Deflationism and Axiomatic Truth*. Cambridge, Massachusetts: The MIT Press, doi:10.7551/mitpress/9780262015868.001.0001.

—. 2021. "On Reflection." *The Philosophical Quarterly* 71(4): 738–757, doi:10.1093/pq/pqaa083.

HORSTEN, Leon and LEIGH, Graham E. 2017. "Truth is Simple." *Mind* 126(501): 195–232, doi:10.1093/mind/fzv184.

HORWICH, Paul. 1990. *Truth*. Oxford: Blackwell Publishers. Second edition: Horwich (1998).

—. 1998. *Truth*. 2nd ed. Oxford: Oxford University Press. First edition: Horwich (1990), doi:10.1093/0198752237.001.0001.

—. 2001. "A Defense of Minimalism." *Synthese* 126(1/2): 149–165. Substantially revised version in Horwich (2010, 35–56), doi:10.1023/a:1005279406402.

—. 2003. "A Minimalist Critique of Tarski on Truth." in *Philosophy and Logic in Search of the Polish Tradition: Essays in Honour of Jan Woleński on the Occasion of*

His 60th Birthday, edited by Jaakko HINTIKKA, Tadeusz CZARNECKI, Katarzyna KIJANIA-PLACEK, Tomasz PLACEK, and Artur ROJSZCZAK, pp. 3–12. Synthese Library n. 323. Dordrecht: Kluwer Academic Publishers. Reprinted, in substantially revised form, in Beall and Armour-Garb (2005, 75–84) and in Horwich (2010, 79–97), doi:10.1007/978-94-017-0249-2_1.

—. 2010. *Truth – Meaning – Reality*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199268900.001.0001.

KYBURG, Henry E., Jr. 1970. "Conjunctivitis." in *Induction, Acceptance, and Rational Belief*, edited by Marshall SWAIN, pp. 55–82. Synthese Library n. 26. Dordrecht: D. Reidel Publishing Co., doi:10.1007/978-94-010-3390-9_4.

MCGEE, Vann. 1992. "Maximal Consistent Sets of Instances of Tarski's Schema (T)." *Journal of Philosophical Logic* 21(3): 235–241, doi:10.1007/bf00260929.

QUINE, Willard van Orman. 1970. *Philosophy of Logic*. Foundations of Philosophy Series. Englewood Cliffs, New Jersey: Prentice-Hall, Inc. Second edition: Quine (1986).

—. 1986. *Philosophy of Logic*. 2nd ed. Cambridge, Massachusetts: Harvard University Press. First edition: Quine (1970).

RAATIKAINEN, Panu. 2005. "On Horwich's Way Out." *Analysis* 65(3): 175–177, doi:10.1111/j.1467-8284.2005.00546.x.

SOAMES, Scott. 1997. "The Truth about Deflationism." in *Philosophical Issues 8: Truth*, edited by Enrique VILLANUEVA, pp. 1–44. Atascadero, California: Ridgeview Publishing Co., doi:10.2307/1522992.

TARSKI, Alfred. 1944. "The Semantic Conception of Truth and the Foundations of Semantics." *Philosophy and Phenomenological Research* 4(3): 341–375, doi:10.2307/2102968.

The Problem of Thomistic Parts

Fr. James Dominic Rooney, OP

Thomas Aquinas embraces a controversial claim about the way in which parts of a substance depend on the substance's substantial form. The substantial form is responsible for the identity/nature of the parts of the substance such a form constitutes. Aquinas's controversial claim can be roughly put as the view that things are members of their kind in virtue of their substantial form.

The aim of this paper will be to defend Aquinas's claim that every time the *x*s come to compose a *y*, those *x*s have to undergo a change in kind membership. After defending the Thomistic account, I propose that approaching problems of material composition as a Thomist has a significant, oft-overlooked advantage of involving a thoroughgoing naturalistic methodology that resolves such problems by appeal to empirical considerations.

Thomas Aquinas embraces a controversial claim about the way in which parts of a substance depend on the substance's substantial form. On his metaphysics, a "substantial form" is not merely a relation among already existing things, in virtue of which (for example) the arrangement or configuration of those things would count as a substance. The substantial form is rather responsible for the identity or nature of the parts of the substance such a form constitutes (Marmodoro and Page 2016, 17–18). Substantial forms thus do not have substance-parts as that which they characterize, i.e., their matter. However, the implication is that if some substances come to compose another substance as proper parts, those things that become parts must ipso facto cease to be substances. Conversely, if a material part ceases to compose a substance as a part, that thing will become a substance or a heap of substances. Aquinas's controversial claim can be roughly put as the view that things are members of their kind in virtue of their substantial form. When a part ceases to compose a substance, it ceases to be that kind of thing that it was when it composed its parent substance and, so, loses all of the properties or powers that are associated with being a part of that substance.

As an illustration of the implications of this claim, consider the death of Socrates. Aquinas holds that "the soul […] is the form of the whole body and each of its parts. […] Thus, it is necessary that each part of a man and that of an animal receive its existence and species from the soul as from its proper form"[1] (*QDA*,[2] a. 10 resp.). When Socrates dies, all of his parts—his body and his eyes and his skin—cease to have their act of existence and cease to be the things that they were when they composed Socrates. Socrates's corpse does not have eyes or hands or skin, at least properly speaking, because, after the soul leaves the body, "neither eye nor flesh nor any part remains except equivocally"[3] (*QDA*, a. 10 resp.). To put it simply, Aquinas's claim results in the implication that every time the *x*s come to compose a *y*, those *x*s have to undergo a change in kind membership (Koslicki 2008, 147).

This has been called the "homonymy principle," and it follows from Aquinas's view of substantial forms and, specifically, from the position that substantial forms inform prime matter rather than substance-parts. Consequently, a substantial form must account for the determinate actuality of every part of the substance. Yet the homonymy principle has appeared to many to be so counterintuitive as to practically require a belief in the existence of substance-parts of substances. Kathrin Koslicki argues that if the homonymy principle were true, it would be impossible to explain continuity in change.[4] The aim of this paper will be to defend that the Thomistic claim that substantial forms account for the determinate actuality of every part of a substance is plausible and coherent. After defending the Thomistic account, I propose that approaching problems of material composition as a Thomist has a significant, oft-overlooked advantage of involving a thoroughgoing naturalistic methodology that resolves such problems by appeal to empirical considerations.

1 Being a Part of a Thomistic Substance

The Thomistic claim about substantial forms would not be controversial merely if it held that parts ceased to be parts when they ceased to compose a substance or that something became a part when it composed something else.

1 "anima […] [e]st enim forma et totius corporis, et cuiuslibet partis eius. […] Unde oportet quod quaelibet pars hominis et animalis recipiat esse et speciem ab anima sicut a propria forma."

2 Thomas Aquinas, *Quaestiones Disputatae de Anima*.

3 "neque oculus neque caro neque aliqua pars remanet nisi aequivoce."

4 This remains true in Koslicki (2018, 217–220).

One could think, for this reason, there was a problem if we were to characterize "substances" merely as those material objects that are not parts of any other. What seems to be missing from my characterization of a substance is the way in which a substance is a properly unified thing, as would result from having a substantial form that actualizes all of that substance's parts, making it one *kind* of thing. Stump points out, for example, that on Aquinas's view of what it is to be a substance, "the ability to exist on its own is a necessary but not a sufficient condition for something's being a substance" (Stump 2003, 42).

Stump attempts to appeal to the contemporary metaphysical concept of emergence as that which sets apart Aquinas's view of substance. But she contrasts her view of emergence with other contemporary views because "on Aquinas's way of thinking about material objects what can emerge when form is imposed on matter is not just properties but substances" (Stump 2003, 196). She defines what it is to be an emergent whole, i.e., a substance: "W is an emergent thing if and only if the properties and causal powers of W are not simply the sum of the properties and causal powers of the constituents of W when those constituents are taken singillatim, outside the configuration of W" (Stump 2003, 43). Non-substances, like artifacts, are nothing more than "the sum of [their] parts" because the properties and powers of such wholes are nothing over and above the properties and powers of their parts (Stump 2003, 44).

Obviously, Stump's claim can be misleading without further qualification. As Marmodoro and Page point out, emergence is overly permissive as a criterion of substancehood: "There are plenty of examples of material objects having—on account of their structure or external relations—emergent properties or functions that the parts individually do not have, without such objects *ipso facto* being substances" (Marmodoro and Page 2016, 4). Yet Stump's claim is not that it is sufficient for something to be a substance if the parts actually composing some substance lack properties or powers individually that the whole substance possesses. Rather, Stump's claim is that a substance has powers and properties that are not a sum of the powers and properties of the parts that could *potentially* come to compose it. That is the sense of the qualification that these parts must be considered apart from the actual configuration of the whole.

Similarly, Stump's definition would be an insufficient characterization of Aquinas's views if it was understood as presuming this claim: that one and the same thing can come to have properties or powers in virtue of composing another substance as a proper part; or, that one and the same thing can lose

doi: 10.48106/dial.v77.i1

powers or properties when ceasing to compose a whole and becoming a substance. Such a view would violate Aquinas's homonymy principle. Aquinas's case of the death of Socrates showed that his body could not be identical when it was actually alive and when it was a corpse; "neither eye nor flesh nor any part remains except equivocally"[5] (*QDA*, a. 10 resp.). At the moment that Socrates dies, his body ceases to exist, and a corpse (or, more accurately, a heap of substances) comes into existence. As all his parts go out of existence when his soul ceases to compose his body, no parts of Socrates are found in his corpse. Socrates's substantial form informs prime matter directly, and the *only* matter that persists over a change of substances is prime matter.

Stump's characterization of a substance as an "emergent whole" aims to capture this relation by noting that the parts actually composing a substance are not identical with the things that potentially compose it, and similarly for the properties and powers of those parts. A substantial form is precisely that form which accounts for the existence of material substances in general, including those that might be part-less simples, and therefore accounts not only for the composition of some parts into a whole but for the *matter* of the whole. Aquinas draws a distinction between two senses a material composite can have matter. One is the familiar sense in which a material substance has its *integral* parts, such as my hands or fingers or toes, which are the material parts composing me. These are the "proximate matter" of a material substance. Another sense is much less familiar. This is the way of considering matter in a general sense as a part of any material composite substance. And Aquinas, indeed, is known for characterizing this sense of matter as a *potentiality*. This potentiality is what Aquinas calls "prime matter" (*materia prima*):

> Prime matter is [...] matter without any form at all, "materiality" (as it were) apart from configuration. When it is a component in a matter-form composite, prime matter is the component of the configured composite which makes it the case that the configured thing can be extended in three dimensions and can occupy a particular place at a particular time. But by itself, apart from form, prime matter exists just potentially; it exists in actuality only as an ingredient in something configured. So we can remove form from prime matter only in thought; everything which exists in reality is configured in some way. For this reason, Aquinas sometimes says that form is the actuality of anything. Configuration

5 "recedente anima, neque oculus neque caro neque aliqua pars remanet nisi aequivoce."

> or organization is necessary for the existence of anything at all; without form, nothing is actual. (Stump 2003, 37)

Aquinas's prime matter is not one obscure material stuff that is a part of every object, an ultimate material substrate of which everything else is a modification (*pace* Jaworski 2016, 332). Aquinas is strongly against such a theory, in fact, as he argues that prime matter must not be a special kind of matter, in the sense in which my flesh or calcium are kinds of matter, but needs to be devoid of all form. If all material objects had one substratum with its own form, and this substratum was part of every material object, he argues that the substratum would be the only true substance and every other object would be a modification of it (see Thomas Aquinas, *De Substantiis Separatis*, c. 6). It would be the inverse of an atomistic universe, with all everything forming one "blobject."

Prime matter is thus not an integral part but the potential to be a material object, considered apart from any particular actual way something could be a material object by being a member of a determinate kind of thing. The characterization of prime matter as the potentiality for a thing to have a location in space-time and extension in three dimensions follows from the fact that Aquinas holds these features as proper to all material objects in general, of any kind. These features of matter in general are not merely a relation or a feature of our concepts or definitions of matter but are essential to matter in general, i.e., "the potentiality of matter is nothing other than its essence" (*ST*[6] I, q. 77, a. 1, ad. 2; see also Wippel 2000, 319). Aquinas's claims about prime matter are therefore claims that what is essential in being a "material object" is that something has dimensions and spatiotemporal configuration, but not that one has those features in any determinate way. For something to be "material" is only to have indefinite dimensions and space-time location. Prime matter plays a role aside from being the principle in virtue of which things have dimensions because, as we will see, prime matter plays a theoretical role in how we should understand certain kinds of material changes. Prime matter is the matter *from* which some set of integral material parts are constituted (*pace* Dumsday 2021, 2721; see further *ST* I, q. 66, resp.).

Inasmuch as prime matter is only the potential to be a material substance, Aquinas therefore holds that prime matter could not exist without having a substantial form to give it particular properties and to determine what *actual*

6 Thomas Aquinas, *Summa Theologiae*.

doi: 10.48106/dial.v77.i1

dimensions or location would have. Moreover, Aquinas argues that prime matter alone *necessarily* cannot constitute any particular kind of object, as it is strictly *contradictory* to claim that prime matter exists without being informed by any substantial form. For prime matter to exist by itself, without the actuality of any form, would be to say that something purely potential could be actually existent yet without being actual in any way. That would be nonsense (*ST* I, q. 66, a. 1; cf. Quodlibet III, q. 1, a. 1, resp.). All this is to say, in short, that prime matter is a *metaphysical part* of material composites, having a similar kind of relation to a substance as that which a substantial form does: as Aquinas puts it, prime matter is "*incomplete* being without the substantial form" (Marmodoro and Page 2016, 12).

Aquinas's idea can then be put much more simply: substances are the place where the buck stops for existing, being actual, composing, or having properties or powers. Substances are what really exist, whereas parts only have existence insofar as they are parts of substances: "Only the composite whole [viz., a substance] has existence (*esse*), properly speaking. All the other parts of the substance[, …] though things of a certain kind, nevertheless exist only in an improper sense, in virtue of the whole's existence" (Pasnau 2011, 624). Aquinas therefore distinguishes two ways of attributing existence to things either as substances or as modifications "in" another:

> Existence (*esse*) is attributed to something in two ways. In one way, as to that which properly and truly has existence or exists, and in this way it is attributed only to a substance that subsists *per se*. Thus *Physics* I [186b4–8] says that a substance is what truly is. All those things, on the other hand, that do not subsist *per se*, but are in another and with another—whether they are accidents or substantial forms or any sort of parts—do not have existence in such a way that they truly exist, but existence is attributed to them in another way—that is, as that by which something is—just as whiteness is said to be not because it subsists in itself, but because by it something has existence-as-white (*esse album*).[7] (Thomas

7 "Uno modo ut sicut ei quod proprie et vere habet esse vel est. Et sic attribuitur soli substantiae per se subsistenti: unde quod vere est, dicitur substantia in I Physic. Omnia vero quae non per se subsistunt, sed in alio et cum alio, sive sint accidentia sive formae substantiales aut quaelibet partes, non habent esse ita ut ipsa vere sint, sed attribuitur eis esse alio modo, idest ut quo aliquid est; sicut albedo dicitur esse, non quia ipsa in se subsistat, sed quia ea aliquid habet esse album."

> Aquinas, Quodlibet IX, q. 2, a. 2, resp.; translated by Pasnau 2011, 624)

Aquinas, therefore, defines forms in terms of their role, forms being that in virtue of which something has actual existence in some way: "all that from which something has existence [*esse*], whether that existence is substantial or accidental, is able to be called a 'form' [...] and because form makes [something] to be in actuality, therefore form is said to be an actuality [*actus*]" (Thomas Aquinas, *De Principiis Naturae*, caput 1, 5). As prime matter lacks any actuality and is essentially a potentiality for being configured by a form, the substantial form of a given substance accounts for everything in terms of which that material thing is a determinate member of its kind, e.g., having essential properties or powers (Stump 2003, 38). According to Aquinas's way of thinking, "there is no such thing as existence beyond the specific ways of functioning manifested by specific kinds of things" (Pasnau 2012, 492). This is why Aquinas claims that a "substantial form gives being [*esse*] to matter *simpliciter*" (*SLM*,[8] 775).

Further, corresponding to the way in which there are two senses as to what the "matter" is in a material substance, Aquinas distinguishes two senses of what potentiality in a substance that the substantial form actualizes. Even though the matter *from which* every material composite is constituted is prime matter, and "no other substantial form intervenes between [a substantial form such as] the soul and prime matter" (*QDA*, a. 9, resp.), no object is merely actualized prime matter. Instead, the matter *of which* some material substance is composed is its proximate matter, e.g., its integral parts. Hence, the immediate potentiality that the human soul makes actual is a living human body (and all its parts); "the human body is the matter proportionate to the human soul, and it is related to the soul as potency to actuality" (*QDA*, a. 1, ad 5). Aquinas considers such proximate matter brought into existence by the substantial form as a *particular way* in which the potentiality of prime matter is actualized. Since no one substantial form actualizes all of the potential of prime matter, this is only some specific potentiality of matter that corresponds to the specific actuality that a substantial form brings about: whatever matter, under whatever determinate conditions, that is essential to the kind of substance the form constitutes.

Consequently, integral parts are *actually* what they are only in virtue of composing their substance. The actuality of the parts "is in some sense derived

8 Thomas Aquinas, *Sententia libri Metaphysicae*.

doi: 10.48106/dial.v77.i1

from the actuality of the whole, inasmuch as the whole substance, including all of its parts, shares in just a single existence" (*QDA*, a. 1, ad 5). Conversely, every part of a substance, merely by being a part, is something "in potentiality" (*in potentia*) to the substantial form of that substance. And Aquinas draws this conclusion quite clearly:

> That parts [of a substance] are in potentiality alone is apparent because none among them is separate, inasmuch as, given that all the parts, insofar as they are parts, are united in a whole. For everything which exists actually ought to be distinct from other things, because one thing is distinguished from another by its own actuality and form [...]. But those things, which are taken to be parts, are separated from each other when the whole dissolves, then indeed they are beings in actuality, surely not as parts, but as matter existing in privation from the form of the whole. Just as, clearly, in the case of earth, fire, and air, which, when those are parts of a mixed body, are not actually in existence, but only potentially [existing] in a mixture. When they become truly separated, then they are in actual existence and are not parts. For, none of the elements, before they are arranged (that is, before they are altered in the mixture and become one mixed thing [composed] from those elements), is one [element] with another, except in the sense that a heap of stones is one thing *secundum quid* [i.e., in some qualified sense] and not simply.[9] (*SLM*, 1632–1633)

Obviously, on this way of understanding forms as being that in virtue of which not only some parts are configured into a composite material substance, but that all of that substance's matter *exists,* it is not easy to see what pluralism about substantial forms could mean. If a substance had two substantial forms, this would be for one and the same substance to exist "twice over," and that

9 "Et quod partes sint in potentia tantum, patet, quia nihil de numero earum est separatum; immo omnes partes inquantum sunt partes, sunt unitae in toto. Omne enim quod est in actu, oportet esse ab aliis distinctum, quia res una dividitur ab alia per suum actum et per formam, sicut supra dictum est. Quando autem ea, quae ponuntur partes, fuerint separata abinvicem dissoluto toto, tunc quidem sunt entia in actu, non quidem ut partes, sed ut materia existens sub privatione formae totius. Sicut patet de terra et igne et aere, quae quando sunt partes corporis mixti, non sunt actu existentia, sed potentia in mixto; cum vero separantur, tunc sunt in actu existentia, et non partes. Nullum enim elementorum antequam digeratur, idest antequam per alterationem debitam veniat ad mixtionem, et fiat unum mixtum ex eis, est unum cum alio, nisi sicut cumulus lapidum est unum secundum quid, et non simpliciter."

just seems nonsense. Aquinas therefore treats pluralism about substantial forms as a conceptual confusion: "Since every form gives a certain *esse*, and it is impossible for one thing to have two substantial existences (*esse*), it is necessary that if the first substantial form coming to matter gives substantial *esse* to it, a second superadded form must give an accidental existence (*esse*)" (*In II Sent.*, dist. 18, q. 1, art. 2, corp.; translated by Wippel 2011, 122). Forms either make a substance exist, simply speaking, or they otherwise configure that substance to exist in some way (e.g., as having a property). And if the substance already exists, any further forms in the substance can only bring about modifications within that already-existing substance; i.e., further forms would not be *substantial* forms.

It is easy to misunderstand Aquinas's claim about parts as in potential to the substantial forms of the substances they compose. Pasnau sees in Aquinas's claim that parts of substances only exist potentially as an attempt "to distinguish a thing from its existence, as if it is one kind of question to ask whether a thing is real, and another kind of question to ask whether it exists" (Pasnau 2011, 627). Pasnau is assuming that Aquinas's view is that one and the same thing becomes a potential thing when it is a part and then is an (actually) existent thing when it becomes a substance, where both potential and actual things are real.[10] But Aquinas is more radical: he is not going to countenance material parts or wholes "surviving" substantial changes of these sorts. As Koslicki notes of Aquinas's views, "no object that is *not* already part of a whole that is unified under a single form can survive *becoming* part of such a whole; and no object that *is* already part of such a whole can survive *ceasing* to be part of it" (Koslicki 2008, 147). So it would be strictly false, on Aquinas's view, that one and the same thing could be characterized at one time as a part and at another as a substance. Potential parts of substances are not the same things that are the actual substances they can become.

Aquinas, therefore, also treats the view that a substance can have other substances as parts as a conceptual confusion. Having other substances as proper parts is just what it is to be an aggregate and not a substance. Substantial forms, on Aquinas's view, account for the existence of a substance precisely because they account for the existence of every part of that substance:

> The soul as the form of the body [...] is united directly to the whole body, because it is the form of the body as a whole and of each

10 Pasnau thinks Aquinas's response requires an appeal to the doctrine of a "real distinction" between "essence and existence" (2011, 626–627).

doi: 10.48106/dial.v77.i1

> of its parts. And this must be maintained, for, since the body of a man or that of any other animal is a certain natural whole, it will be said to be one because it has one form whereby it is perfected, and not simply because it is an aggregate or a composition, as occurs in the case of a house and other things of this kind. Hence each part of a man and that of an animal must receive its act of existing and species from the soul as its proper form. (*QDA*, a. 10, resp.; translated by John Patrick Rowan in 1949)

Notice, however, that this way of thinking entails that substantial form is *intrinsic* to the substance and all of its parts. An arrangement, for example, is not something intrinsic to the things arranged, and this is what makes an arrangement an accidental rather than a substantial form—it is only a relation among substances. Aquinas uses the illustration that a mass of bronze coming to be a statue only involves an accidental change or alteration because "the bronze, before the advent of the form or figure, has actual existence, and its existence does not depend on that figure" (*De Principiis Naturae*, caput 1, 8). If the statute's shape were a substantial form, that shape would not only result in the existence of bronze shaped-as-a-statue but the existence of its matter as well. "A form must be something *of* that to which it gives existence, for form and matter are intrinsic principles constituting the essence of a [corporeal] thing" (*De Principiis Naturae*, caput 1, 8).

Similarly, a substantial form is not like a causal *agent* internal to some parts, e.g., gathering them together or pushing them through space. To say then that a substantial form is that in virtue of which a substance exists or is actual is not to say that the substantial form creates or generates its own material parts. A chemist who makes a new chemical compound by combining the constituents in the right way is bringing into existence that compound, certainly, but in a different sense. Aquinas thinks of a causal agent as making some matter *to have a form*: "corporeal forms are caused […] by matter being brought from potentiality into act by some composite agent" (*ST* I, q. 65, a. 4, resp.; translated by the English Dominican Fathers in 1912–1936). This account of causal agency, even though utilizing an act-potency distinction, presumes that forms play a distinct role. The forms are that *in* what is actual, whereas the agent remains outside of what she actualizes. Whereas the chemist does not become the chemical compound she mixes up, the substantial form and the matter actualized by it "must have one and the same act of existing (*esse*),

something which is not true of an efficient cause and an effect to which it gives *esse*" (Wippel 2011, 124).

Aquinas's controversial claim can then be stated more fully as follows: the substantial form not only accounts for the existence of the substance and the composition of other material parts in a whole substance but for *everything that is essential to the parts*, whether existence or actuality or powers or properties. As we saw, this claim entails that a substance ceases to exist when it begins to compose a part of something else. Then, given that the substance no longer exists when it becomes a part, all of its properties or powers also cease to exist. Similarly, composing a whole with certain properties, like being human, entails that the parts also have certain properties in virtue of being parts. Thus, my hand is a human hand merely in virtue of composing me, but it ceases to be a hand when it ceases to compose me.

2 The Puzzle of Parts

Now, a puzzle looms. Aquinas's claims about Socrates's hand ceasing to be a hand when he dies or his body ceasing to be a body both seem empirically false. Consider a case presented by William Jaworski as a counter-example to Aquinas's theory of composition:

> Oxygen. In the process of respiration, oxygen atoms, as molecular oxygen (O_2), enter a human bloodstream. Those atoms oxidize red blood cells, becoming parts of those cells and, by extension, a human being. After circulating in human blood, those same oxygen atoms are eventually expelled, albeit in a different molecular configuration (CO_2).[11]

If we assume that oxygen atoms, molecules, and human beings are all substances, Aquinas is apparently committed to saying that these oxygen atoms were not the same atoms at every point in this process. As Jaworski puts it, "That atom does not survive being incorporated into me. It is instead replaced by something else—something that perhaps has many of the same characteristics as the original atom, but that is nevertheless numerically different from it" (2016, 118). Aquinas appears, therefore, to claim that when those atoms begin to compose a human being, those atoms ipso facto cease to exist.

11 I have made the case more specific. See below.

doi: 10.48106/dial.v77.i1

In fact, Aquinas would be committed to the stronger claim that those atoms *never* existed because they always composed some other substance at every step of the case.

Yet we see no such replacements happening when substances come to acquire new parts; the oxygen atom does not appear to be replaced by a "token" oxygen look-alike. We could have used radioactive isotopes to "tag" the atoms within the molecular oxygen and then identify the same two atoms at every point in the process. If these atoms ceased to exist or never existed, how could we track each particular atom, their properties, and their causal powers? Oxygen atoms do not disappear when they compose other molecules, nor do their properties or powers just cease to exist. Cases like Oxygen are not exceptional or infrequent. When we break up a composite substance, the ingredient substances can come back into full existence with entirely the same properties they had before they composed anything. Oxygen atoms do not just "pop" into existence when we break up, e.g., H_2O molecules with hydrolysis; the atoms were parts of the molecular structure itself! The Thomist view thus appears straightforwardly empirically false.

This puzzle should not be as puzzling as it might seem. Aquinas holds that substances cannot be parts of other substances, and if an atom becomes a part of a molecule, ipso facto, that atom ceases to be a substance. Nevertheless, Aquinas does not hold merely that the atom no longer exists. Rather, his claim is simply that the thing that was an atom substance *became* an atomic part of a molecule. For Aquinas, if the oxygen atom was a substance and remained a substance over the event described in Oxygen, the atom would not compose that molecule but only become, at best, spatially co-located with the molecule. More accurately, as molecules are not separable things from the atoms that compose them, it would be that molecules are nothing more than spatial arrangements of atoms; i.e., molecules are not genuine material objects but pseudo-objects.

Further, every substance exists in a determinate way, that is, as a member of a kind. If *oxygen atom* and *hemoglobin protein* are each a distinct kind, it is easy to see that the implicit assumption is likely empirically false. When incorporated into an oxyhemoglobin molecule, an oxygen atom is configured differently from when it is not incorporated into a protein. When, in respiration, oxygen atoms are incorporated into the hemoglobin in red blood cells, those atoms bond with the hemoglobin, and their structure changes. A free-floating oxygen atom undergoes a series of changes when it bonds with hemoglobin, such that it comes to have different properties and structural

relations to other things (e.g., the hemoglobin—see, for example, Jenkins and Van Kessel 2002, 122). In becoming a part of a protein, then, the Thomistic account holds that the oxygen atom ceases to be a substance when it becomes a structural component of the substance that is the protein. It is no part of the account, *pace* Jaworski's objection, that the oxygen atom is replaced with a completely identical look-alike when it bonds with hemoglobin; instead, the oxygen atom becomes quite different structurally and in its other properties at the moment it becomes a structural component of a protein.

Consider for a moment the simpler case of H_2O molecules. H_2O and O_2 are distinct molecules with distinct properties and powers. These two molecules have distinct properties and powers because they are distinct structurally. Further, their oxygen parts are distinct structurally as well: the oxygen in H_2O has two distinct covalent bonds with hydrogen, and dioxygen's atom parts have a double covalent bond. These kinds of bonds modify the distinct properties of the whole in a way that the whole has properties and powers distinct from other possible configurations, but it is also true that the oxygen atoms being so bonded are distinct in powers and properties from a single free-floating oxygen atom. Thus, an individual atom might react under certain conditions (hydrogen gas will react with O_2 in combustion), whereas in the molecule it does not react so (H_2O does not combust, even as a gas). If each constitutes a distinct kind of substance, it is not clear how molecular oxygen coming to compose a thing of a distinct kind has *not* ceased to be an instance of "molecular oxygen."

The only way in which the case of Oxygen could be a counter-example to Aquinas's theory of composition is if it described a case where something came to compose another without any change of essential properties. First, the case does not plausibly show this, since it is an empirical matter whether there *is* such a natural kind as an "oxygen atom,"[12] and the atom in the case underwent a great deal of changes that contrast with the way that we ordinarily take an atom to be determinate when it is *not* part of those compounds. Second, it is not clear how such a case *could* disprove Aquinas's views without assuming what it intends to refute. Aquinas's views are that something only counts as an instance of material composition when the parts depend on the whole in a certain way. If the case was taken to describe merely extrinsic changes of spatial location among the atoms, Aquinas would just flatly deny that the

12 Kerry McKenzie has written extensively in criticism of the view that particles are fundamental entities; see McKenzie (2014, 2011) and McKenzie and Muller (2017).

doi: 10.48106/dial.v77.i1

atoms composed anything throughout the process. Instead, Aquinas's claim only entails that the atom, when it composes a molecular substance, is at best a derivative property-bearer in virtue of that substance (if it is the right integral part of the molecule to do so) and that any properties it has that result from being a part of a molecule would cease when it ceases to compose that molecule.

In fact, the puzzle of parts is actually not a puzzle about material composition but a puzzle about what persists over substantial change or the change of one substance into another substance (there would be *no* puzzle if there were no changes of parts). And Aquinas's perspective, bluntly, is that the wrong place to look for continuity in change is in what it is to be a substance. The right place to account for continuity in substantial change is in the relation among the substances that go into or come out of existence while carefully distinguishing the material parts involved in the changes. The reason that H_2O can be split into oxygen and hydrogen by electrolysis is not that there was both molecular oxygen and a water molecule spatially co-located at the beginning of the process. Rather, the reason one can split out these two components is because water molecules are such that they can be decomposed into hydrogen and oxygen atoms.

What we should appeal to in order to account for continuity in substantial changes is the actuality or potentiality corresponding to the substance (and its integral parts') ability to undergo the relevant changes. Hydrogen and oxygen atoms are "potential parts" of water molecules because they are the proximate matter essential to being a thing of the kind "water molecule." H_2O does not exist without them. But it is not essential to hydrogen or oxygen atoms to constitute H_2O. Further, we can assume H_2O molecules have distinguishable integral parts, such that we can identify the hydrogen and oxygen atomic parts; that is, the parts can have their own properties that they bear in virtue of being parts of that substance (the parts bear properties derivatively). Then, we can say that when the water molecule is decomposed in hydrolysis, there are two senses in which *the same* integral parts are what became one hydrogen and two oxygen atom substances.

On the one hand, the matter *from which* they were constituted is just the same matter that came from the molecule because their coming into existence consists in an actualization of that potential—prime matter—that was formerly "in" the water molecule. They did not "pop" into existence from nowhere. On the other hand, the material integral parts of the water molecule were characterized as (derivative) property bearers that had their own inter-

nal structures and properties in virtue of being essential parts of the water molecule. When the water molecule decomposes into the atoms, the new substances only need to lose those properties that were essential to the whole they composed. We can imagine, for example, that the hydrogen atomic part being tagged with an isotope is a property that is accidental to both the molecule and the hydrogen. If they bore any properties that were unique to themselves as the parts, these would be accidental to the water molecule and could come to characterize the new substances as well. There is nothing preventing the hydrogen atom substance, resulting from the decomposition of the molecule, from likewise being characterized by the isotope tagging.

It might be alleged that, in these cases, "scientists do not claim to be tracing powers, but things that bear the powers" (Pawl and Spencer 2016, 138). But Aquinas's view does not require thinking that all we track are only powers of oxyhemoglobin. Instead, some parts can bear properties insofar as a substance can have a property *in*, or in virtue of, one of its parts,[13] where the whole will be the subject even of accidental properties of the parts, as when, for example, I have the property of "being wounded" in virtue of my foot being wounded. And one of these integral parts can be such that it can become a numerically distinct substance, bearing that property in its own right, when it ceases to compose the whole. Atomic parts are just like this. There does not seem to be any empirical reason to think that, in tracking an oxygen atom through my body with a radioactive isotope, we need more than Aquinas's account can give: a certain isotope was introduced into my body, and, in virtue of a chemical change, became composed as a part of one of my atomic parts; that atomic integral part of me was tracked, in virtue of the radioactive properties now associated with that atomic part, and then the isotope part or the atomic part ceased to compose me, eventually (see further Toner 2008).

Yet, Pasnau alleges there is the *inverse* problem of that posed by Jaworski: how to explain the fact that *exactly similar* properties persist over substantial changes. His example is that the skin color of Socrates can be identical to the skin color of Socrates's corpse a moment after death, and "it seems nothing short of miraculous that, without that form, the corpse retains so many exactly similar accidents" (Pasnau 2011, 585). Here again, it is important to note that while it is true that Socrates's substantial form is that in virtue of which Socrates and his parts are characterized by essentially human proper-

13 Aquinas draws this very contrast between properties and parts. Properties, accidents, are not "particular things." But parts, even though they too are dependent entities like properties, can be considered particular things in ways that properties cannot (*QDA*, a. 1, ad. 9).

doi: 10.48106/dial.v77.i1

ties, Aquinas's claim is *not* that no qualitatively similar kind of property or part, even an *exactly* similar property, could ever characterize anything else (Quodlibet I, q. 4, a. 1; translated by Sandra Edwards in 1983). There is no reason that Socrates and Socrates's corpse could not be qualitatively identical in regard to skin color (and Aquinas does say that they could be; *De Ente et Essentia*, c. 5). Aquinas's view only requires holding that Socrates's corpse is not the same substance as Socrates's body and that no essentially human properties "survive" Socrates's demise.

Some kinds of substances, given their proximate matter, have the potential to become other substances, whereas others have integral parts such that those parts can become substances in various ways. For that reason, organ transplants are not a metaphysical mystery. A heart, when detached and "on ice," is no longer a part of any particular human, although it is suited to become the heart of another person because of its physical characteristics; the heart can retain those properties, while detached, that did not derive solely from composing a human being. E.g., muscular cells are still capable of moving under electric shocks, and the whole heart is capable, when reattached, of pumping blood (*pace* Pawl and Spencer 2016, 144). Nothing about Aquinas's position requires that the heart, when it is *in via* during a transplant, will not be a thing "suited to beat and pump blood."[14] All that is required, on Aquinas's metaphysics, is that my heart has undergone some intrinsic, essential change when it ceased to compose my body, such that it is a distinct thing when it is a part of me and when it is not. What we want to know is why, if it is a distinct thing, the heart outside of my body has apparently very similar properties. On Aquinas's view, the answer is that my heart was just the kind of part that could become such a substance—acquire *that* kind of substantial form—when it was detached from my body, given the proximate matter of which human being from which it was taken was composed essentially included a heart.

In Oxygen, the relation among the substance kinds to which the oxygen, hemoglobin, and oxyhemoglobin belong explains the potentiality of the oxygen atom to become a part of oxyhemoglobin. Other things could not compose hemoglobin unless they both underwent some suitable external stimulus *and* were suitable to have the potentialities to compose oxyhemoglobin. However, in fact, all of these conditions were met in Oxygen. Thus, when actualized by some stimulus conditions, the proximity of the oxygen atom to the

14 I am speaking generally because a detached heart is likely not a substance but a collection of individual substances (cells).

hemoglobin initiated a chemical reaction of oxidization of that protein, and oxyhemoglobin was composed from those other substances. Similarly, after it comes to composing oxyhemoglobin, that oxygen atom substance becomes an atomic part typical of oxyhemoglobin and has a certain set of chemical bonds with the protein. Not every oxygen part of any molecule is of such a type as *this* oxygen part of oxyhemoglobin: not every oxygen part bonds with the particular geometry involved in an oxygen part's bond to the rest of the oxyhemoglobin molecule (i.e., an "end-on bent" configuration in bonding with the Fe_2 parts of that molecule).[15] Yet some other kinds of molecules could have oxygen in the same type of configuration as the oxygen parts of oxyhemoglobin.

An objector might point out that the homonymy principle entailed, for Socrates, that Socrates's eye is no longer an eye after he dies. The objector could then argue that we should not think that "atom" is being used homonymously of the atom substance and the atomic part of oxyhemoglobin: "These are 'atoms' in just the same sense, whether or not they compose anything! Whereas it might be plausible that 'eye' is a functional term for a certain kind of part, and we can imagine it ceases to apply to an eye when it is separated from its functional system, surely atoms are not a functional part of that sort."

In response, first, it seems likely to me that Aquinas and Aristotle treat the aforementioned "transplant" cases as the organ ceasing to have any biological properties merely because organ transplants were not then medically possible, and they did not know that an organ's cells do not immediately cease to be alive on detachment. Yet, even if we *were* committed to the homonymy principle for all parts, this can be plausible when we specify the nature of the kinds in question. If we assume that kinds are kinds *of substances*, and substances are those objects composing no other, then a kind such as "oxygen atom" cannot apply to the oxygen in oxyhemoglobin. An oxygen atom as a substance is, by stipulation, something that does not compose anything else, and the oxygen in hemoglobin clearly composes it. As an integral part that essentially characterizes molecules of the kind, the atomic part now belongs to the kind "oxyhemoglobin" in virtue of composing the whole.

The only thing further the objector might be looking for, as we saw with Jaworski, is the *numerical identity* of the thing having the property at every time it has the property (whether as a part or a whole). But numerical identity strikes me as something we cannot just *see*, because we characterize the

15 See the case study of carbon monoxide poisoning in Gaffney and Marley (2018, 233–234).

doi: 10.48106/dial.v77.i1

substances (even for Jaworski) in terms of what is essential to them. For something to be numerically identical is to say that it underwent no change in what it is essentially. To say that the atom is "numerically identical" whether it composes the molecule or not entails both that the oxygen differs in *no* way when it composes oxyhemoglobin and that oxyhemoglobin is not a kind of molecule. Both seem empirically false.

The atoms in oxyhemoglobin have distinct shapes, properties, and powers from the oxygen atoms composing O_2. In addition, if oxyhemoglobin were not a kind of molecule, then the example could not undermine Aquinas's overall thesis, as the atom only gets spatially located very close to the other. Similarly, if the objector were to insist that numerically identical *properties* characterize the oxyhemoglobin and the oxygen atom that results from its decomposition, even though it is conceded that they are exactly similar, such a response would appear to beg the question against Aquinas that only substances (e.g., atoms and molecules) bear properties. The properties, to be "numerically identical," would have to be substances, in Aquinas's sense. If they were substances, however, they could not compose a material object without, necessarily, ceasing to be substances. Thus, it is not clear how to make sense of either attempt to cash out "numerical identity" in a way that does not beg the question against Aquinas's position.

3 The Plausibility of the Solution

The ontologically relevant payoff of distinguishing between prime matter and the proximate matter of material parts is that it allows Aquinas to draw a distinction such that he can affirm both that, even though these parts potentially could constitute something else, these parts actually compose a substance's essential parts. On the one hand, distinguishing a substantial form as a particular, a metaphysical part of a composite, is to say that the substantial form is not identical with those material parts or the whole they compose (*pace* Scaltsas 1994; cf. *SLM*, 1674, translated by John P. Rowan in 1961b; compare: Keinänen and Hakkarainen 2017; Keinänen 2018). Substantial forms are *particulars* that, in virtue of characterizing some set of material parts, account for why those parts constitute a whole of some kind. It is not a feature of our counting or conceptual schemes that the material *x*s are such that they compose *y*, but an extra-mental fact about the *x*s that they compose *y* since they do so by reason of the substantial form that is intrinsic to that material

object and all its parts.[16] This is what it is to say that the substantial form is the actuality *of* that substance *y* and its parts, the *x*s, is that each of the *x*s (and the substantial form) are such that they *actually* compose *y* in virtue of something that is essential to them (the substantial form[17]), even though the substantial form is not identical with the *x*s or the *y*.

On the other hand, there remains a sense in which *those* material things, the *x*s, could have composed *z* instead of *y*. Insofar as the *x*s are adequately characterized by being "material" (i.e., composed of prime matter as a metaphysical part), Aquinas holds that material objects, in general, are essentially such that they can undergo a change of kind—one material object can serve as matter *from which* we generate the matter of another material object of a different kind, because prime matter is just the potentiality of any material object to come to constitute a distinct kind of material object under the relevant conditions (Brown 2005, 79–83). It is not essential to their matter that the *x*s compose *y*, and thus *z*'s parts could be composed *from* the *x*s.

Aquinas does not hold that matter is fundamentally or essentially particulate, and it is apparent now why he cannot think it is. If matter were fundamentally particulate, it would be essential to those particles that they are mereologically simple, and it would consequently be false that they could ever compose a whole object. Aquinas's claim that matter is not essentially particulate does not merely constitute medieval empirical speculation lacking knowledge of the existence of fundamental physical particles but follows directly from the assumption that there are material composites (and that any arbitrary two or more material things do not compose an object)—that is, that material composition occurs only under some restricted circumstances.

An account of material composition would be involved in an infinite regress if it only specifies the conditions under which some things come to compose a whole without explaining what the things are that *get* composed. This would be akin to explaining what it is to be a bearer of properties, a thing defined in terms of being what bears properties, by appeal to a distinct property of that thing (see the discussion in Loux 2006, 84–120). Aquinas's earlier objection to a plurality of substantial forms in one substance is that it involves one in an infinite regress of exactly the same sort; it infinitely defers the question as to what has the potentiality to be composed, or it assumes an ultimately

16 "An act is in that of which it is the actuality: wherefore the soul must exist in the whole body and in each of its parts" (*ST* I, q. 76, a. 8, resp.).

17 By reflexivity, every part is a part of itself. The form is thus a metaphysical part of itself, itself that in virtue of which it composes the whole as a metaphysical part.

doi: 10.48106/dial.v77.i1

atomistic account of reality as a necessary truth (viz., Peter van Inwagen's account of composition—compare: Renz 2018). If matter were essentially particulate, then this would seem to mean that *to be a material object* is just to be one of the particles; that is, it is *necessary* for the material objects that they be mereologically simple. But this position would involve a confusion between two different senses of what it is to be a material object: being the sort of thing that essentially has integral parts and being the sort of thing that essentially has dimensions, occupying a spatiotemporal location. But it is false that having the sort of thing with dimensions and a spatiotemporal location necessarily entails that all the material objects are essentially mereologically simple.

To put it another way, even though prime matter characterizes every material object, it is only a way to describe the potentiality to be a material object and has *no* essential characteristics at all. Prime matter is thus an explanatory principle in virtue of which it is a contingent matter whether *any* material substance exists—i.e., it is not essential to any material thing, merely in virtue of being material, that it be actual. Consequently, whatever constitutes "materiality" cannot be something that has any essential properties (neither a property nor a property-bearer), but rather that special sort of potentiality that corresponds to the potential to be a material substance: namely, that no material object exists necessarily, but only contingently. Prime matter has to be "pure potentiality" in this way in order to thread the needle between the views that composition among two or more material things occurs of necessity (universalism) and that material objects are essentially mereologically simple (nihilism).

In fact, Aquinas argues that if there are things having dimensions and spatiotemporal location, they are *by that very fact* composite objects—composites precisely inasmuch as they possess spatial parts: "From the fact that matter has corporeal existence through forms, it immediately follows that there are dimensions in matter whereby it is understood to be divisible into different parts, so that it can receive different forms corresponding to its different parts" (*QDA*, a. 9, resp.; translated by John Patrick Rowan in 1949). Inasmuch as material objects have parts that are spatially distinguished, these are *integral parts*, and we find that distinct integral parts can bear distinct properties or evince different structures (that is, spatial parts of one object can bear distinct accidental forms). Aquinas's point is that spatial parts of a substance are the right kind of thing themselves to have properties in various ways. We can, for

example, characterize one spatial region of the same substance as hot and another as cold because spatial parts can have distinct properties.

John Heil denies this and argues, to the contrary, that if substances can only have spatial or temporal parts, then that is enough to claim that they are "mereologically simple" (2012, 18–19). The only realistic candidate substances would be particles or fields, whereas macroscopic entities like humans are not (Heil 2012, 19–22; see also 2003, 177–192). This is because fields or particles would not have parts that bear distinct properties from the whole. The whole field has one set of properties borne directly by the field, and all of its spatial parts (considering the field's extension in space as a way to divide it into spatial parts) have the same properties. Whereas a substance like a field or a particle can have many properties, properties are not parts of those things. Properties are not parts of their substances (Heil 2012, 107). The argument is to the effect that if Aquinas admits that a material object has integral parts, and these parts can bear properties, then those parts must be substances. Aquinas would therefore be contradicting the shared assumption that only substances are property-bearers.

But Aquinas has not assumed that integral parts of a substance, among which are that object's spatial parts, are bearers of properties *in their own right*. Aquinas just denies this implicit premise of Heil's argument. What is required is a distinction between the fact that some things essentially bear properties and that other things bear properties in a derivative way. That is, there is no contradiction if integral parts bear properties only *accidentally*, i.e., only in virtue of composing something that is *essentially* a property-bearer, a substance. When my hand is white, then *I* am white with respect to my hand. My hand is not a property-bearer in its own right but bears properties only in virtue of being a part of me. Yet, understanding integral parthood in this way, such that integral parts bear properties in an only accidental way, we not only can divide an object according to its spatial dimensions—top half, right half, etc.—but also in terms of the way in which each distinct part can bear distinct properties, or have a power, or be structured.

4 Conclusion

The reason that Heil restricts the parts of objects to merely spatiotemporal parts, however, seems independently motivated. At root, the difference lies in how to identify or classify which things are genuine material objects and, so, which things are genuine parts or properties of them. Stump's elaboration of

doi: 10.48106/dial.v77.i1

the phenomena associated with "emergence" in contemporary metaphysics aims to make it plausible, with appeal to empirical data, that there exist properties that come to be seated in one whole rather than merely a collection of substances. These properties are not such that they could be merely the properties of a complex of substances but must be of an emergent whole. By contrast, Heil is inclined to hold that the scientific data shows that the world is perhaps fundamentally composed only of fields or that they are the only things that qualify, empirically, as having properties.

Further, Aquinas's substances are emergent wholes in the sense that the whole is not identical with the ingredients that lead to their emergence because he believes substantial change of one substance into another is possible.[18] Heil would likely disagree with the terminology of "emergence" because, e.g., it is not clear under what circumstances two fields could come to compose a distinct object. One field does not appear to be the thing that could become another field. Heil's vision then seems to entail that there is only accidental change among the fields that exist.

Yet Aquinas is not claiming that it is necessarily the case that there are material objects, but only that there would be no problem of material composition if there were no material objects. From Aquinas's perspective, if the scientific data shows us that the world is built entirely from fields, and fields are things that entirely lack spatial parts and spatiotemporal location, then it would merely be the case that the existence of spatiotemporally-located objects is merely *apparent*. (Of course, one would need to explain how the

18 Even if Aquinas were wrong about the substantial change of macroscopic entities, the claim becomes far more plausible in the subatomic world. The identities of some physical particles appear inseparable from the physical systems they form—for example, when electrons become "entangled" in a quantum state. Aquinas could hold that these states are, in fact, hylomorphic composite substances without any parts other than spatial and temporal parts. Electrons appear to cease to exist (except "virtually") in these states they compose. As with the way Aquinas elsewhere treats an Aristotelian homogenous "mixture"—cf. *De Mixtione Elementorum*—electrons still relate to the subsequent entangled state in virtue of a mathematical "structural" correspondence between the individual quantum state (and powers/properties) of the electrons before their being entangled and that state after they are entangled. This way of considering particles allows us to hold that particles are substances in certain circumstances, even if these particles can come to compose entangled quantum states. This would entail realism about the quantum wave function. The relation of particles as substances to the quantum states would be composition, and of the quantum state to the distinct particles that result from a "collapse" of the wave function would be decomposition. The issues here are obviously highly simplified. See further Ney and Albert (2013). And it seems to me that there are philosophical accounts of the metaphysics of the wave function already being proposed that are compatible with my loose characterization: cf. Gao (2017). Robert Koons (2018) appeals to Aquinas in this way.

fields exist and how these fields generate the apparent spatiotemporal world.) The point is that once we admit that there are objects that have spatial dimensions, we are committed to the fact that there exist material composites, and, subsequently, we can pose questions as to the way in which their composite parts compose one whole. Consequently, if there are material objects with parts, they at least bear the property of being a *y* such that the *x*s compose it, and so these are all ipso facto going to be the kind of thing that can bear (one or more) properties by reason of the kind of thing they are.

But this points to an advantage of Aquinas's account of substances. What fills out the account, beyond the claims he makes about matter and form composition, is the way in which Aquinas thinks we discover and identify the natural kinds. Picking out the things with substantial forms thus requires us to identify *the causal process* by which some things are modified in order to become a whole. There are real differences among the causal processes that might produce distinct kinds of substances or parts. The account, in fact, rests the nature of what resolves the problem of material composition on the extra-mereological considerations that should lead us to hold that some causal processes produce (or lead to the destruction of) instances of a natural kind. A causal process produces a substance, for instance, when that process brings some things together as parts, such that the things that are the parts of the resultant substance cannot be described as essentially the same as the things that became those parts. Conversely, the process that leads to the destruction of a substance is one in which that substance "loses" some essential part or property over the course of the change and, ipso facto, ceases to exist. Even a "brute" theory of composition on which there is an infinite series of conditions under which the *x*s compose *y* can be hylomorphic if those conditions are interpreted as corresponding to different natural kinds of object; there would be no conflict with hylomorphism's account of the nature of material composition if there were infinitely many natural kinds, even if there might be some other good reason to think it is impossible for there to be truly be an "infinite" number of such kinds (Markosian 2008, 19). The account of natural kinds, therefore, tells us which are the things that have substantial forms and, so, are the "genuine material objects."

This Thomistic way of approaching material composition makes an empirical, scientific method for identifying the substances by appeal to facts about causal processes a promising one. What it *means* for a thing to be unified is to be a member of a natural kind, to be a material object of a determinate

doi: 10.48106/dial.v77.i1

nature.[19] Whether an atom comes to compose a molecule, for example, is determined by whether that atom's essential properties changed in virtue of becoming an apparent part. These give us those criteria in virtue of which there is unity among the molecule's atomic elements, quarks, and electrons. And those criteria for molecules are of a very different sort from the criteria for other substances. Different natural kinds exemplify different kinds of unified causal powers or activities among the parts; it would be impossible to give a general characterization of what that unity consists in, in general, without a minimal account of the natural kinds. Different natural kinds exemplify different essential activities or properties or powers or structures, and so their parts are unified according to different criteria.

In the Thomistic picture, then, the role that the form has in unifying some parts into a whole depends on *the right kinds of changes in virtue of which some object changes kind membership* and thereby becomes a part of another thing of a distinct kind. That is, if substances can never compose other substances as parts, things must cease to be members of one kind and become members of another kind at the time they become parts.[20] For human beings, the unity among our parts is explained by facts about our organic chemistry as animals of a special sort. Captain Hook, as a human, is a living, organic thing of a particular natural kind *human being*, where his parts are unified by biochemical bonds and various bio-organizational interactions. That in virtue of which all of his material parts are of the same kind, *human being*, is what makes those parts belong to Hook, but Hook is neither identical with his kind (the essence of *human being*) nor is Hook merely that which makes him a member of the kind or all his parts human (his soul). Rather, Hook is a substance formed of his material parts, among which is his hand, suitably informed and united by his soul. Captain Hook loses his hand to the crocodile *because* that action causes his hand to cease to have the right biochemical bonds and interactions so that Hook's substantial form ceases to be something Hook shares in common with that piece of matter that formerly was his hand. When Captain Hook acquires a hook made of iron and wood, and puts the hook in the place of his hand, that hook is not a part of Captain Hook because the hook is not changed into the right *kind* of thing that could be a part of his

19 One should note the way this account is similar to the way in which Jansen criticizes and modifies Kit Fine's account of embodiments around the notion of a sortal: see Jansen (2019).

20 Koslicki recognized this implication, where she notes that accepting Aristotle's homonymy principle requires that, if a substance becomes a part, "any such transformation would essentially involve a change in *kind membership*" (Koslicki 2008, 147).

living, organic body. His substantial form cannot overlap the hook because the substantial form just is that in virtue of which Hook is a biological organism of a kind, and the hook is not the right kind of thing to be part of a biological organism.[21]

In sum, on the Thomistic theory, the only way to determine the way in which a substantial form accounts for the unity of the parts of a substance is to determine the nature of the substance in question. This seems the best reason to commend the Thomistic account of substantial forms. Questions about the unity of material objects can be resolved to a certain degree at an abstract, metaphysical level but are fundamentally a matter to be resolved through empirical investigation. Hylomorphism of the Thomistic sort appeals to forms to explain material composition, but what and how a form accounts for the composition of a substance depends on the *kind* of substance it informs. This reliance on natural kinds grounds answers to "what it is in virtue of which the *x*s compose *y*" soundly on empirical concerns; the question of what kinds there are, or their properties, can be given a fully satisfactory answer only in tandem with scientific investigation.[22,23,*]

Fr. James Dominic Rooney, OP
0000-0003-0087-3218
Hong Kong Baptist University, Hong Kong SAR
jdrooney@hkbu.edu.hk

References

Brown, Christopher M. 2005. *Aquinas and the Ship of Theseus: Solving Puzzles About Material Objects*. London: Continuum International Publishing Group.

21 For example, a human being *could* digest iron and make it part of its organism, but the hook as-is has not been digested or appropriately modified to form part of Hook's body. The same would be true of any sophisticated prosthetic; as long as these are such that they are not "biologically continuous" with the human organism, they are not parts of that organism except in perhaps an extended or metaphorical sense. By contrast, if we created a biological replica of Hook's hand, growing a cloned set of Hook's cells and structuring them in an appropriate way, then attaching such a biological prosthetic *would* be able to become part of Hook's body (if it was not rejected by his immune system, etc.).

22 Permission has been given to reprint material adapted from James Rooney (2022).

23 All Latin editions obtained from those hosted at Corpus Thomisticum, courtesy of the University of Navarre.

∗ THANKS

doi: 10.48106/dial.v77.i1

Dumsday, Travis. 2021. "Can a Relational Substance Ontology Be Hylomorphic?" *Synthese* 198(suppl. 11): 2717–2734, doi:10.1007/s11229-019-02173-1.

Gaffney, Jeffrey and Marley, Nancy. 2018. *General Chemistry for Engineers*. Amsterdam: Elsevier Science Publishers B.V., doi:10.1016/C2015-0-05956-1.

Gao, Shan. 2017. *The Meaning of the Wave Function: In Search of the Ontology of Quantum Mechanics*. Cambridge: Cambridge University Press, doi:10.1017/9781316407479.

Heil, John. 2003. *From an Ontological Point of View*. Oxford: Oxford University Press, doi:10.1093/0199259747.001.0001.

—. 2012. *The Universe as We Find It*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199596201.001.0001.

Jansen, Charles M. 2019. "De-Fining Material Things." *Dialectica* 73(4): 459–477, doi:10.1111/1746-8361.12280.

Jaworski, William. 2016. *Structure and the Metaphysics of Mind: How Hylomorphism Solves the Mind-Body Problem*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780198749561.001.0001.

Jenkins, Frank and Van Kessel, Hans. 2002. *Nelson Chemistry 12*. Toronto: Thomson Nelson.

Keinänen, Markku. 2018. "Instantiation and Characterization: Problems in Lowe's Four-Category Ontology." in *Studies in the Ontology of E.J. Lowe*, edited by Timothy Tambassi, pp. 109–124. Neunkirchen-Seelscheid: editiones scholasticae.

Keinänen, Markku and Hakkarainen, Jani. 2017. "Kind Instantiation and Kind Change: A Problem for Four-Category Ontology." *Studia Neoaristotelica* 14(2): 139–165, doi:10.5840/studneoar20171427.

Koons, Robert C. 2018. "Hylomorphic Escalation: An Aristotelian Interpretation of Quantum Thermodynamics and Chemistry." *American Catholic Philosophical Quarterly* 92(1): 159–178, doi:10.5840/acpq2017124139.

Koslicki, Kathrin. 2008. *The Structure of Objects*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199539895.001.0001.

—. 2018. *Form, Matter, Substance*. Oxford: Oxford University Press, doi:10.1093/oso/9780198823803.001.0001.

Loux, Michael J. 1998. *Metaphysics: A Contemporary Introduction*. Routledge Contemporary Introductions to Philosophy. London: Routledge, doi:10.4324/9780203438244.

—. 2006. *Metaphysics: A Contemporary Introduction*. 3rd ed. Routledge Contemporary Introductions to Philosophy. London: Routledge. Revised edition of Loux (1998), doi:10.4324/9780203968871.

Markosian, Ned. 2008. "Restricted Composition." in *Contemporary Debates in Metaphysics*, edited by Theodore Sider, John Hawthorne, and Dean W. Zimmerman, pp. 341–364. Contemporary Debates in Philosophy n. 10. Malden, Massachusetts: Wiley-Blackwell.

MARMODORO, Anna and PAGE, Ben. 2016. "Aquinas on Forms, Substances and Artifacts." *Vivarium* 54(1): 1–21, doi:10.1163/15685349-12341310.

MCKENZIE, Kerry. 2011. "Arguing Against Fundamentality." *Studies in History and Philosophy of Science. Part B: Studies in History and Philosophy of Modern Physics* 42(4): 244–255, doi:10.1016/j.shpsb.2011.09.002.

—. 2014. "Priority and Particle Physics: Ontic Structural Realism as a Fundamentality Thesis." *The British Journal for the Philosophy of Science* 65(2): 353–380, doi:10.1093/bjps/axt017.

MCKENZIE, Kerry and MULLER, F. A. 2017. "Bound States and the Special Composition Question." in *EPSA15 Selected Papers: The 5th Conference of the European Philosophy of Science Association in Düsseldorf*, edited by Michela MASSIMI, Jan-Willem ROMEIJN, and Gerhard SCHURZ, pp. 233–241. European Studies in Philosophy of Science n. 5. Cham: Springer, doi:10.1007/978-3-319-53730-6_19.

NEY, Alyssa and ALBERT, David Z., eds. 2013. *The Wave Function: Essays on the Metaphysics of Quantum Mechanics*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199790807.001.0001.

PASNAU, Robert. 2011. *Metaphysical Themes 1274–1671*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199567911.001.0001.

—. 2012. "Mind and Hylomorphism." in *The Oxford Handbook of Medieval Philosophy*, edited by John MARENBON, pp. 486–504. Oxford Handbooks. Oxford: Oxford University Press, doi:10.1093/oxfordhb/9780195379488.013.0022.

PAWL, Timothy and SPENCER, Mark K. 2016. "Christologically Inspired, Empirically Motivated Hylomorphism." *Res Philosophica* 93(1): 137–160, doi:10.11612/resphil.2016.93.1.6.

RENZ, Graham. 2018. "Form as Structure: It's Not So Simple." *Ratio* 31(1): 20–36, doi:10.1111/rati.12155.

ROONEY, James Dominic. 2022. *Material Objects in Confucian and Aristotelian Metaphysics: The Inevitability of Hylomorphism*. London: Bloomsbury Academic, doi:10.5040/9781350276376.

SCALTSAS, Theodore. 1994. *Substances and Universals in Aristotle's* Metaphysics. Ithaca, New York: Cornell University Press.

STUMP, Eleonore. 2003. *Aquinas*. The Arguments of the Philosophers. London: Routledge, doi:10.4324/9780203928356.

THOMAS AQUINAS. 1912–1936. *Summa Theologica*. London: Burns, Oates & Washbourne. Translated by the Fathers of the English Dominican Province, http://www.domcentral.org/summa/.

—. 1926. *De ente et essentia*. Opuscula et textus. Series scholastica n. 1. Münster i.W.: Monasterii; Aschendorf. Edidit Ludovicus Baur.

—. 1949. *The Soul: A Translation of St. Thomas Aquinas'* De anima. St. Louis: B. Herder Book Co. Transl. by John Patrick Rowan, ark:/13960/t87jo3ros.

—. 1950. *Sententia libri Metaphysicae*. Roma: Domus Editorialis Marietti Taurini. Translated and edited by Roberto Busa, https://www.corpusthomisticum.org/cmp00.html.

—. 1953. *Quaestiones disputatae de anima*. Roma: Domus Editorialis Marietti Taurini. Translated and edited by Roberto Busa, https://www.corpusthomisticum.org/qda01.html.

—. 1956. *Quaestiones Quodlibetales*. Roma: Domus Editorialis Marietti Taurini. Translated and edited by Raimondo Spiazzi, https://www.corpusthomisticum.org/qda01.html.

—. 1961a. *Commentary on the Metaphysics of Aristotle, Volume 1*. Library of Living Catholic Thought. Chicago, Illinois: Henry Regnery Company. Transl. by John Patrick Rowan, ark:/13960/t5gb5411f.

—. 1961b. *Commentary on the Metaphysics of Aristotle, Volume 2 (Books VI-XII)*. Library of Living Catholic Thought. Chicago, Illinois: Henry Regnery Company. Transl. by John Patrick Rowan, ark:/13960/t5gb5411f.

—. 1968. "De substantiis separatis." in, pp. D1–D87. Leonina, XL. Roma: Ad Sanctae Sabinae. Ed. H.-F. Dondaine, https://www.corpusthomisticum.org/ots.html.

—. 1976. "De principiis naturae [ad Fratrem Sylvestrum]." in *Opuscula IV*, pp. 1–47. Leonina, XLIII. Roma: Editori di San Tommaso. Ed. H.-F. Dondaine, based on the previous work of P.J.M. Perrier.

—. 1983. *Quodlibetal Questions 1 and 2*. Mediaeval Sources in Translation n. 27. Toronto: Pontifical Institute of Mediaeval Studies. Translated with an introduction and notes by Sandra Edwards.

—. 1996a. *Quaestiones disputatae de anima*. Leonina, XXIV, 1. Paris: Éditions du Cerf. Ed. B.C. Bazán.

—. 1996b. *Quaestiones de quolibet. Préface, Quodl. VII, VIII, IX, X, XI*. Leonina, XXV, 1. Paris: Éditions du Cerf. Ed. R.-A. Gauthier.

—. 1996c. *Quaestiones de quolibet. Quodl. I, II, III, VI, IV, V, XII*. Leonina, XXV, 2. Paris: Éditions du Cerf. Ed. R.-A. Gauthier.

—. n.d. *Summa Theologica*. Leonina, IV-XII. Roma: Typographia poliglotta S.C. de Propaganda Fide.

Toner, Patrick. 2008. "Emergent Substance." *Philosophical Studies* 141(3): 281–297, doi:10.1007/s11098-007-9160-6.

Wippel, John F. 2000. *The Metaphysical Thought of Thomas Aquinas: From Finite Being to Uncreated Being*. Washington, D.C.: The Catholic University of America Press.

—. 2011. "Thomas Aquinas and the Unity of Substantial Form." in *Philosophy and Theology in the Long Middle Ages: A Tribute to Stephen F. Brown*, edited by Kent Emery Jr., Russell L. Friedman, and Andreas Speer, pp. 117–154. Studien und Texte zur Geistesgeschichte des Mittelalters n. 105. Leiden: E.J. Brill, doi:10.1163/ej.9789004169425.i-1006.32.

A Generalization of the Reflection Principle

Wolfgang Spohn

This paper generalizes (probabilistic) auto-epistemology by amending the original forward-looking reflection principle of van Fraassen (1984), which is about learning or favorable epistemic changes in general, by a new but similar backward-looking principle, which is about forgetting and other unfavorable epistemic changes. The generalization is argued to be a completion by defending what is called the no-neutrality condition. Due to the similarity, analogous consequences are provable for both principles. This fact is utilized for a plausibility check of the new principle. Finally, it is argued that this generalization should not be considered as a special case of the expert principle.

Van Fraassen (1984, 244) has introduced the *reflection principle*, as he called it: *Given* your tomorrow's probabilities are such and such, your present conditional probabilities should be the very same. Hence, this principle may be called *forward-looking*. It is *the* basic principle of probabilistic or Bayesian auto-epistemology. Various interesting consequences have been derived from it. It has met diverse criticisms, several variants have been offered in response, and it is well-known by now that it holds only under restrictions. However, the literature does not offer clear ideas on whether and how it may be suitably amended or even completed. This paper tries to do better by proposing what I call the full reflection principle. It has certainly been suggested in one way or another. But I am after precise statements allowing strict inferences.

Here is how the paper proceeds: Section 1 briefly recapitulates van Fraassen's original principle. Section 2 suggests an equally strong and formally analogous *backward-looking* principle. Both combine to what I call the full reflection principle. Thereby, I propose to double, as it were, and arguably complete the range of auto-epistemology. In section 3, the completeness claim is supported by what I call the no-neutrality condition. However, the completeness claim does not go so far as to offer an account

of how to iteratively apply the full principle. Section 4 looks at some well-known consequences of van Fraassen's principle, which carry over to the backward-looking principle. This is intended to serve as a plausibility check of the proposed generalization. Section 5 discusses what happens if the reflection principles are referred to as auto-epistemic propositions themselves; this is often not strictly distinguished. There, we will discover a slight disanalogy between the forward- and the backward-looking principle. Section 6, finally, defends my generalization against the objection that it is already contained in the familiar generalization of van Fraassen's principle, known as the expert principle.

1 Van Fraassen's Reflection Principle

This paper intends to focus on the philosophical and not on the formal aspects of its topic. Still, the main claims should be stated in a formally precise way. For this purpose, we need a modicum of notation. We will refer to a prior time, 0, and a posterior time, 1. This may be today and tomorrow, or yesterday and today; we will use both readings. Everything indexed in this way will refer so as well. Thus, P_0 is to represent your actual prior and P_1 your actual posterior credences, where I suppose you to satisfy all rationality constraints we think to be pertinent. These are at least the synchronic axioms of probability and some diachronic learning rules we need not fix. Without an index, the temporal reference may be any. I use π, with or without indices, as a variable for your possible probability functions. Sets of those functions represent *auto-epistemic* propositions about your own probabilities; e.g., $\{\pi_0 \mid \pi_0(A) = x\}$ represents the proposition that your prior probability for A is x, $\{\pi_1 \mid \pi_1 = Q_1\}$ represents the proposition that your posterior probability function is Q_1, etc.

These possible probability functions are about some fixed algebra of propositions concerning some worldly affairs. Not all worldly affairs; the range may be quite restricted but need not be made explicit. The label "worldly" is to mean that these propositions are about external matters and not about your epistemic states. (In section 5, we will consider dropping this restriction.) It also means that the propositions are about empirical, not abstract or formal, matters. The epistemology of formal sciences is a very different topic not amenable to the present methods in my view.

However, your actual P_0 and P_1 are not only about these worldly propositions but also about all the auto-epistemic propositions just introduced and all

algebraic combinations thereof. Thus, P_0 and P_1, but not the possible π, reflect on how your prior and posterior probabilities might be.

Now, we are ready to explicitly state van Fraassen's principle; I will call it the *forward-looking reflection principle* for quite obvious reasons:

(1) $P_0(A \mid \{\pi_1 \mid \pi_1 = P_1\}) = P_1(A)$ for all worldly propositions A.

That is, *given* your posterior credence is P_1 and, in particular, $P_1(A)$ for A, your prior for A is also $P_1(A)$.[1] Why do I state the condition in (1) as $\{\pi_1 \mid \pi_1 = P_1\}$ and not in the usual way as $\{\pi_1 \mid \pi_1(A) = P_1(A)\}$? Because the usual version is weaker and derivable from the present version, which is clearly the intended one.[2] Note, moreover, that "$\pi_1 = P_1$" cannot be literally true, given that π_1 is only about worldly propositions, while P_1 is also about autoepistemic propositions. Strictly speaking, I should refer to P_1 as restricted to worldly propositions. This would be too cumbersome, though. This little sloppiness will do no harm.

Let me emphasize that I am using probabilistic terminology here only for convenience; it is the most familiar one. However, the reflection principle is not restricted to Bayesian epistemology. It holds for any kind of epistemic format that allows for conditional epistemic states. Binkley (1968) has proposed a qualitative version: if you now believe that you will believe p tomorrow, you should believe p already now. This principle played a crucial role in his account of the surprise examination paradox. In Spohn (2010, 125–127), I have stated the reflection principle in terms of ranking theory. One may state the reflection principle in terms of imprecise or interval probabilities. And so on.[3] Here, I will simply assume that epistemic states are represented in a fixed formal format, and then I choose the most familiar one. I do not want

1 In order to keep things simple, this statement is predicated on the assumption that there are only finitely many possible π under consideration. Generally, though, there are uncountably many π in play, and then the condition is likely to have probability 0 so that the conditional probability in (1) is undefined. Then we must replace (1) by Constraint 2, as Skyrms (1980, 163) calls it, which says for all intervals I: $P_0(A \mid \{\pi_1 \mid \pi_1(A) \in I\}) \in I$. This solves the problem because now the condition will usually have a positive probability. See also Goldstein (1983).

2 For instance, from (1), but not from the usual version, one can derive $P_0(A \mid \{\pi_1 \mid \pi_1(A) = x \,\&\, \pi_1(B) = y\}) = x$, which is desirable, too.

3 Schoenfield (2012) argues that the reflection principle *entails* probabilities to be precise—and is, therefore, to be rejected because other arguments require probabilities to be imprecise. If so, one should state, I think, a reflection principle for interval probabilities: $P_0(A \mid \{\pi_1 \mid \pi_1(A) = I\}) = I$, where P_0 and the π_1 would now be interval probability functions. (Observe the difference to "Constraint 2" in footnote 1.)

doi: 10.48106/dial.v77.i1

to discuss the more basic issue of the most adequate representation of epistemic states. Hence, the entire paper will move within a standard probabilistic representation of epistemic states, though only *pars pro toto*.

I shall not repeat here the grounds on which (1) is widely accepted; I shall only refer to some criticism below. I should mention, though, that (1) has been considered before. Spohn (1978, 161–162) and Goldstein (1983) have proposed the iteration principle (which is almost equivalent; see sections 4 and 5). Quite generally, one might say that the reflection principle is implicit or well-nigh explicit in de Finetti's (1937) philosophy of probability. One incurs considerable philosophical costs if one wants to abandon it.

Gaifman (1986) has suggested a more general reading of (1). He defines P_1 to be an expert for you at the prior time 0 (regarding A) if and only if (1) holds for your P_0. In this reading, (1) turns into what is called the *expert principle*, saying that you should trust the experts—though this is now tautological because an expert has been defined as one you can trust in this way. Still, the maneuver provides quite a graphic reading of the original (1), namely, as saying that you should accept your posterior opinion as an expert for you.[4] In section 6, I shall discuss whether the amendment of (1) proposed in section 2 may simply be conceived as an instance of this expert principle.

This makes (1) plausible. Surely, if your posterior probabilities have learned something, have gathered evidence, are better informed, etc., then your prior probabilities should consider them to be an expert for you. At the same time, this points to the restricted applicability of (1). Your future self is not always better informed. You may forget things, you may be tired, brainwashed, confused by drugs, your judgment may be obfuscated by prejudices against better knowledge, etc. In all those cases, your future opinion is not a trustworthy expert for you. This restriction has often been noticed—e.g., by Christensen (1991) (referring to epistemic change due to psychedelic drugs), by Talbott (1991) (referring mainly to memory loss), or by Spohn (1978, 166) (with re-

4 The striking similarity between the reflection principle and Lewis's Principal Principle has often been noticed; see, e.g., Spohn (2010). In these terms, the Principal Principle says that chance at t is your best expert at t, and truth = chance at the end of time is your best expert at all. Christensen (2010b) discusses a principle of rational reflection, which is about your probabilities being conditional not on your future probabilities but on your current probabilities as they should rationally be. This differs from the original reflection principle only in case you suspect your future probabilities to diverge from how the current probabilities should rationally be. In any case, I stick to the original principle and won't discuss such subtly related principles. The real challenge raised by Gaifman is, of course, what to do when I am impressed by several (diverging) experts. Again, peer disagreement is not my topic here.

spect to the iteration principle). Briggs (2009, 64ff.) presents a nice list of exceptions.[5]

In my view, this restriction does not diminish the importance of the reflection principle (1). Whenever your epistemic changes, be they rational or not, do not just occur to you and you rather take a reflective attitude towards them, (1) serves as a fundamental meta-principle. It does not specify how learning precisely works. There, we may consider rules like simple or Jeffrey conditionalization, minimizing relative entropy, etc. But it says, however learning works, it must satisfy (1) if reflected.

Surely, though, (simple) conditionalization is the basic Bayesian learning rule. Since van Fraassen (1984) proposed a Dutch book argument in favor of his reflection principle, critical discussions like those of Christensen (1991) and Talbott (1991) focused on the tension of this argument with the familiar Dutch book justification of conditionalization. Indeed, the relation between the reflection principle and the conditionalization rule is delicate, as displayed also in Weisberg (2007). However, Hild (1998a, 1998b) has already shown that the reflection principle is equivalent to a rule he calls auto-epistemic conditionalization, and he has fully stated the conditions under which auto-epistemic conditionalization and simple and Jeffrey conditionalization may come apart. Therefore, I shall not pursue this connection any further.

2 A Generalization of Van Fraassen's Reflection Principle

If the forward-looking principle (1) has only restricted validity, we should think about whether there are auto-epistemic principles governing the cases not covered by (1). These may be irrational epistemic changes violating diachronic principles of rationality that are supposed to be governed by the reflection principle (1), whatever they are. E.g., whenever your prejudices get a hold on you, although you are sure that they are irrelevant, this is arguably a case of epistemic irrationality. However, there may also be arational epistemic changes, which do not violate diachronic principles but simply fall outside the scope of such principles, such as fatigue or clouding one's judgment by getting drunk (where the only irrationality may have been to get drunk), etc. Forgetting is perhaps the most familiar case of such an arational change. However, it consists not only in unlearning a fact. It can take many

5 Van Fraassen (1995) takes the strange recourse to say that in such cases, your epistemic self ceases to exist, thus recovering the general validity of (1) for *you*.

doi: 10.48106/dial.v77.i1

shapes. E.g., one may forget live possibilities so that one becomes sure of the remaining possibilities. Or one can forget about evidential relations, say, that a characteristic smell is a sign of a poison. And so on.

Such changes seem to be a matter of empirical psychology and not of a normative theory of epistemic rationality. Isn't it an empirical question what we are prone to forget, how our beliefs are influenced by prejudices or drugs, etc.? Yes. However, this does not mean that a theory of epistemic rationality cannot say anything about how we should *rationally deal* with such arational and irrational changes when we, possibly falsely, suspect them to occur. On the contrary, if we are thinking about how to extend the reflection principle (1), this is the challenge we positively face. Hence, the question is: Is there an auto-epistemic rationality principle dealing also with those irrational and arational changes?

Yes, there is. I follow the idea of Titelbaum (2013, chap. 6): just reverse the temporal perspective![6] Place yourself at your posterior P_1 and ask whether you should consult your prior P_0, whatever it was. Certainly not in van Fraassen's cases where you are better informed in P_1. But surely in those cases where you have forgotten something, are foggy-brained, etc. in P_1. Therefore, I propose the following *backward-looking reflection principle*:[7]

(2) $P_1(A \mid \{\pi_0 \mid \pi_0 = P_0\}) = P_0(A)$ for all worldly propositions A.[8]

6 I have first proposed an essential part of the following considerations in Spohn (2017, sec. 10) in a more convoluted way at the end of a long investigation on indexical belief. The connection was Sleeping Beauty, which seems to be not only a problem about indexical belief but also about auto-epistemology. It has been observed, e.g., by Arntzenius (2003) or Mahtani (2017), that the proposition on which the probabilities in the reflection principle are conditioned must egocentrically refer to *me* and *my future* probabilities, not to a's probabilities at t, where a happens to be me and t happens to lie in the future. Here, I ignore this line of thought. The subject's potential uncertainty about her own location is not our issue.

7 Christensen (2000, 352ff.) also speaks of taking a backward-looking perspective, but he thereby means something else. He is interested in diachronic coherence in the sense of an epistemic conservativism, which seeks to preserve as much of the old beliefs as possible while *learning* something new.

8 As just mentioned, Titelbaum (2013) has already observed the symmetry between the forward- and the backward-looking case. He captured both in his principle of Generalized Conditionalization (Titelbaum 2013, 127). I should have known and noticed this in Spohn (2017). However, he refers his observation to simple conditionalization and conceives of the prior state as the conditionalization of the posterior forgetful state with respect to the forgotten proposition. On p. 133, he also considers a generalized reflection principle, but only as derived from simple conditionalization. By contrast, I think of the reflection principles as having an independent role and apply the backward-looking principle to all kinds of arational and irrational epistemic

(2) is as important and fundamental a meta-principle as (1). Whatever the multifarious changes of our epistemic position to the worse, if we reflect on them, (2) must be obeyed in all such cases. In a decision-theoretic perspective, which I do not unfold in this paper, this meta-principle is at the bottom of our efforts to fight forgetting, e.g., by building museums and archives and even inventing scripture, and of our attempts to preserve our epistemic integrity wherever we can, e.g., by banning brainwashing.

It is clear, however, that not both (1) and (2) can be applied generally. Their applications cannot even overlap. In a case of such overlap, your prior probability should trust your posterior one, but your posterior one should reversely follow your prior one; so, if both are mutually envisaged, they must be the same. In other words, only in the case of non-change can (1) and (2) apply simultaneously. In section 5, I will provide a formal proof of this claim.

Hence, the fields of application of (1) and (2) must be disjoint (with the possible exception of non-change). Indeed, this is how I have explained (1) and (2). In order to have uniform labels, let's say that (1) applies in the case of *favorable* changes, such as learning, receiving information, acquiring evidence, etc., and that (2) applies in the case of *unfavorable* changes, such as forgetting, drinking too much, being influenced by prejudices one takes to be unjustified, etc. In a case of a favorable change from P to P' or an unfavorable change from P' to P, let's say that P' is *superior* to P and P is *inferior* to P'.

The next question is: Who is to judge changes as favorable or unfavorable and epistemic positions as superior or inferior? We who decree these principles from outside? No, I think it is preferable to subjectivize the application conditions of (1) and (2). The subject herself must assess the epistemic changes she is considering: Is my change from P_0 to P_1 favorable and P_1 superior? Then apply (1)! Is my change from P_0 to P_1 unfavorable and P_0 superior? Then apply (2)! The first instance to assess this is the subject herself.

Of course, this does not preclude that, in a second step, we have a normative argument about this assessment. Presumably, we all agree that experience is favorable and forgetting is unfavorable. But what about hunches and gut feelings? Gigerenzer (2007) is a strong plea for respecting gut feelings not only as a psychological fact but also as a guideline to rational decision-making. Prejudices may not always be bad. What about epiphanies? Those claiming to have had them feel to be in a superior position. Surely, these are difficult

changes, not just to the forgetting of previous certainties, as Titelbaum does by only dealing with simple conditionalization. Insofar my approach is more general.

doi: 10.48106/dial.v77.i1

and possibly contested issues that we may and must discuss. However, the reflection principles as such are independent of that discussion, and, therefore, we should keep matters separate. The intent of my subjectivizing move was precisely not to get involved in that discussion.

Let's take a slightly more general perspective for stating it. The temporal relations do not really seem to matter. The point is rather that in whatever epistemic position I am, I would trust a superior position and mistrust an inferior position (where it is up to me what I take to be superior and inferior). This seems to be the gist of the principle. If so, we arrive at the following *full reflection principle*, in which the temporal location of the probabilities referred to is left open (hence no indices):

(3) for all worldly propositions A $P(A \mid \{\pi \mid \pi = P'\}) = Q(A)$, given that Q is taken to be the superior one of P and P'.[9]

This is the generalization of van Fraassen's auto-epistemology I would like to propose.

3 The No-Neutrality Condition and the Iteration Problem

Is the generalization a completion? This raises two issues. First, the principle (3) reflects only upon a single possible change. But we may certainly reflect on iterated change. (3) is silent on this and insofar still incomplete.[10] I will not be able to offer a solution, but I will comment on the issue below.

Secondly, if we attend only to a single change, the reflection principle (3) is complete only if the *no-neutrality condition* holds, which states that there are no neutral and no incomparable changes; there are only either favorable or unfavorable changes and nothing else. Then, but only then, there would

9 The "is taken" always refers to the assessment of the subject we are talking about, not our own. In Spohn (2017), I have emphasized this by making the condition in (3) part of the conditional probability statement. However, this raises awkward questions. Are propositions of the form "P' is superior to P" part of the auto-epistemically extended algebra of propositions? Do they receive probabilities? Are these probabilities subject to change? We better avoid such questions. These propositions are outside the epistemic game we are considering. We may rather assume that the subject's superiority assessments are stable within our dynamic scenario. Therefore, I now state this condition outside the probability statement, though still in a subjectivized form.

10 The importance of this issue is underscored by the parallel case in belief revision theory, which was initially restricted to treating only single revisions and thus plagued by the iteration problem, too. I have first raised it in Spohn (1988, 112ff.). It turned out to virtually be an anomaly in the Kuhnian sense; see, e.g., Rott (2009).

always be *the superior one* of P and P' (with the irrelevant exception of non-change), and the application condition of (3) would be complete. Does this condition hold?

Yes, I think so. I welcome favorable changes and seek superior epistemic positions (if they are not too costly), and I try to prevent unfavorable changes and avoid inferior positions (if that is not too costly, either). At least this is so by purely epistemic standards; moral standards, e.g., may tell otherwise in special cases. Thus, a change that is neither favorable nor unfavorable would be one I don't care about. I would say then: it's nice to have the present prior P_0, and it's equally nice to have the posterior $P_1 \neq P_0$ later on; both are fine, and none of them is inferior or superior. This sounds very strange to me. This makes the change from P_0 to P_1 appear arbitrary and without good reason, and then I can't stay indifferent about the change; it must appear unfavorable to me.

To illustrate: Today, I think I will be in good health next year, and tomorrow, just after sleeping and without any new information whatsoever, I think I won't. Usually, this would not be taken as a change of mind but as an expression of continued uncertainty. But say, today, I am firmly convinced that I will be in good health next year. From this perspective, it must appear arbitrary when I would have changed my mind by tomorrow to equally firmly believe the contrary. It would be odd to presently be neutral about such a change; I should rather reject and not trust it.[11]

This is not a cogent argument. It is only to say that I cannot imagine how the no-neutrality condition could be violated. In any case, one must be aware that this condition is a crucial and substantial normative principle. If we accept it, then (3) indeed deserves the label "full reflection principle," at least regarding single changes.[12]

I think, though, that there is a deeper reason behind the no-neutrality condition. It is that, ultimately, there is only one standard for our epistemic states: truth. We try to approach the truth and avoid veering away from the truth, however we measure the distance here. The point is that there is only one "scale" to measure. If epistemic states had to meet many standards on different scales, then indifferences or even incomparabilities might easily arise. Such more complex situations would certainly be relevant when we were to more generally think about what kind of person we want to be. There

11 See also the arguments against arbitrary switching in White (2014, 318ff.).

12 As mentioned, the case of non-change from P to P' may be ruled arbitrarily. We may say then that P is superior, or P' is, or both are. It doesn't make any difference for (3).

doi: 10.48106/dial.v77.i1

are many aspects in which we change for the better or the worse, and we will often have indeterminate preferences about possible personal changes or none at all. But in the case of epistemic change, our judgments seem to be unambiguous.

So far, I have only argued that there are no neutral changes. And I have excluded the possibility of incomparabilities due to a multitude of epistemic standards. However, there are easier ways for incomparabilities to arise. Surely, there are complex changes that are favorable in some respects and unfavorable in others so that, overall, the result is neither superior nor inferior but incomparable. For instance, I learn that I have a date with the president next Friday and simultaneously forget that I have already agreed to meet the vice president at the very same time. I propose to treat this as two changes, first a favorable one and then an unfavorable one—or the other way around; it is not guaranteed that this comes to the same. Often, a temporal succession can be discerned within such a complex change, and sometimes, e.g., in my example, this move may be artificial. However, my proposal seems feasible; it avoids the need to refer favorability and unfavorability to aspects of complex changes, and it saves the no-neutrality condition. So, in any case, it is theoretically beneficial.

However, this move makes the first issue of extending the full reflection principle (3) to iterated change more pressing. To my knowledge, this issue has not been considered in the literature. Perhaps the reason is that it seemed trivial in the case of the original reflection principle. If my first epistemic state trusts the second, and the second trusts the third, already the first state can trust the third. Reversely, in the case of iterated forgetting.[13] However, there are also mixed cases, and I have just alluded to them.[14] The difficult case is the one where my epistemic state first changes in an unfavorable way and then in a favorable way; e.g., first, I forget some things, and then I learn other things. In this case, my initial state can neither trust the final state in the sense of principle (1) nor dismiss the final state in the sense of principle (2). Rather, it seems that I have to engage in a counterfactual consideration. In this case, I can only trust that the epistemic state that *would have* emerged had I not incurred the first unfavorable change (forgotten the one thing) but still

13 Titelbaum (2013, 140) seems to be able to treat the iterated case with the help of his principle of suppositional consistency. But if so, this is due to the fact that the only epistemic changes he considers are the gain and loss of certainties.

14 I have discussed the various cases and their problems a bit more extensively in Spohn (2017, 408–409).

experienced the second favorable change (learned the other things). That is, I would have to speculate not only about my actual epistemic states and their change but also about my counterfactual epistemic states and their change. Hence, a general solution of this problem seems to require quite different theoretical means. It is not a task we can pursue here.

Still, it should be pursued. To emphasize its urgency: As far as I see, the issue of so-called second- or higher-order evidence is closely related. Christensen (2010a) gives a wide variety of examples. A salient structure of them (not all of them) is this: I receive a lot of ordinary (first-order) evidence on a certain matter, and I seem to draw my conclusions from it in the usual rational way. At the same time, I receive higher-order evidence (perhaps falsely) indicating that my cognitive abilities are somehow hampered. I am overly tired, I am told to have consumed a fancy drug, I am instructed that I regularly tend to overoptimism, I may be suffering from hypoxia (a realistic example from Christensen 2010b, 126), etc. So, maybe I should correct my inferences?

In such cases, the higher-order evidence indicates that I should not trust the epistemic state I have reached. But, neither can I simply rely on my prior epistemic state before the change, as the backward-looking reflection principle (2) would have it. As mentioned above, such cases are mixtures of two different epistemic movements. On the one hand, there is the first-order evidence, which I should trust. On the other hand, there is my alleged epistemic handicap, which suggests an accompanying unfavorable change and should make me think about what my inferences would have been without the handicap. So, the issue of higher-order evidence would also profit from solving the iteration problem. However, for the reasons indicated, I don't address this problem further. We should be content with treating the pure cases.[15]

4 Are the Consequences of the Full Principle Acceptable?

The full reflection principle (3) seems to be intuitively plausible and philosophically important. If so, we should also check for its consequences, or at least some of them. The consequences of the original principle (1) are well known. We may follow here Hild (1998a), but need not do it very far.

15 In response to such examples, Briggs (2009, 71) proposes a principle of distorted reflection: $P_0(A \mid \{\pi_1 \mid \pi_1(A) = x\}) = x - D$, where D is a factor measuring the "expected departure from conditionalization on veridical evidence" (regarding A). This may be a correct qualification. But again, it takes two steps at once, and we should first get clear about the single steps.

doi: 10.48106/dial.v77.i1

Formally, the consequences of the generalized (3) are obviously analogous. So, the strategy in this section will be to develop the formal analogy and to check whether the results are also intuitively acceptable.

The first thing to do, perhaps, is to unpack again what we have packed into the condensed abstract statement of (3). It contains, in fact, five different principles, depending on the temporal and the superiority relations between P and P' (where, once and for all, each principle quantifies over all *worldly* propositions A).

One case is where $P = P_0$ is the prior and $P' = P_1$ is the posterior. If P_1 is superior to P_0, we get:

(3a) $P_0(A \mid \{\pi_1 \mid \pi_1 = P_1\}) = P_1(A)$, given that P_1 is taken to be superior to P_0.

This is our original forward-looking reflection principle (1). Since the proviso takes care of the main objections against the principle, there is no need to discuss it further.

However, P_1 may also be inferior to P_0. Then we get something we have not yet explicitly stated:

(3b) $P_0(A \mid \{\pi_1 \mid \pi_1 = P_1\}) = P_0(A)$, given that P_1 is taken to be inferior to P_0.

Given tomorrow's inferior opinion, I stick to my prior opinion. E.g., today, I believe that I have a date with my dentist on Tuesday next week. It is only reasonable, then, to stick to this belief, given I am confused tomorrow and think the date is next Wednesday. Christensen (1991, 234) imagines an agent having swallowed a hefty dose of a certain drug and then being asked: "What do you think the probability is that you'll be able to fly in one hour, given that you'll then take the probability that you can fly to be .99?" He answers in place of the agent: "The sane answer to the above question is clearly one that gives a very low probability to the agent's ability to fly one hour from now, even on the supposition that she will at that time give it a very high probability" (Christensen 1991, 235). This is clearly an instance of (3b). Hence, Christensen may be said to have anticipated the intention of the full principle (3).

Another case is where $P = P_1$ is the posterior and $P' = P_0$ is the prior. Again, this splits up into:

(3c) $P_1(A \mid \{\pi_0 \mid \pi_0 = P_0\}) = P_0(A)$, given that P_1 is taken to be inferior to P_0,

which is our backward-looking reflection principle (2). And into:

(3d) $P_1(A \mid \{\pi_0 \mid \pi_0 = P_0\}) = P_1(A)$, given that P_1 is taken to be superior to P_0,

which we have not yet explicitly stated, either. It says that, given my prior opinion, I stick to my posterior opinion if it has been acquired through a favorable change. Again, this seems to go without saying.

There is, finally, the case where P and P' refer to the same time, so that $P = P' = P_i$ $(i = 0, 1)$, where we may, as mentioned, define the superiority relation either way. Thereby, we get a *synchronic reflection principle*, which is independent of the previous diachronic principles:

(3e) $P_i(A \mid \{\pi \mid \pi = P_i\}) = P_i(A) \quad (i = 0, 1)$.

This may look odd. But it simply says that your present opinion, whether prior or posterior, is presently an expert for you. You presently don't know better than you actually know. Of course, this does not preclude you from accepting other hypothetical experts as well. Since it seems to differ from the other principles, it has also received a separate discussion. We need not go into it now.[16]

What is the relation between the five parts (3a–e) of the full reflection principle (3)? As far as I see, they are independent. As already observed, the synchronic principle (3e) must be independent of the other diachronic principles. Moreover, the principles (3a + d) for favorable changes and the principles (3b + c) for unfavorable changes are independent as well, simply because they refer to disjoint conditions. Maybe the two principles about favorable changes are related? And likewise, those about unfavorable changes? However, I have not discovered any relation, and I think that the five parts (3a–e) are indeed independent. In the next section, however, I shall indicate how things may change.

Let's look at some consequences of the original reflection principle (1), or (3a), in order to check whether their formal generalization is also intuitively plausible. The first is the *iteration principle* already mentioned:

(4a) for all worldly propositions A $P_0(A) = \Sigma\, \pi_1(A) \cdot P_0(\pi_1)$, where $P_0(\pi_1)$ is the subject's prior auto-epistemic probability for π_1 being her posterior,

16 Van Fraassen (1984, 248) takes it to be "uncontroversial." However, Christensen (2007, 322–323), after calling (3a) a principle of epistemic self-respect and quoting a lot of support for it, puts forward putative counter-examples. They trade, I think, on a subtle ambiguity of the inner and outer P_i in (3a). There, the outer P_i has some (second-order) information about the inner P_i, and, thus, the two may come to diverge.

doi: 10.48106/dial.v77.i1

where the sum is taken over all her possible posteriors π_1 taken to be superior to P_0, and where $\Sigma\, P_0(\pi_1) = 1$, i.e., P_0 is sure to undergo a favorable change.

(4a) is entailed by (3a);[17] for a possible reversal, see below. In other words, your prior opinion is always a weighted mixture of all the posterior opinions you may favorably reach, where the weights are given by your prior opinion about reaching these posteriors.[18] This is, I think, a deep epistemological insight.

In the same way, the backward-looking reflection principle (2), or (3c), entails the following *reverse iteration principle*:

(4b) for all worldly propositions A $P_1(A) = \Sigma\, \pi_0(A) \cdot P_1(\pi_0)$, where the sum is taken over all possible priors π_0 taken to be superior to P_1, and again $\Sigma\, P_1(\pi_0) = 1$, i.e., P_1 is sure to have undergone an unfavorable change.

The proof is analogous to the one of (4a). Is (4b) plausible? Yes. If, in your posterior P_1, you have forgotten about something, you will usually not remember, either, what your past opinion about that thing has been. Still, you might autoepistemically wonder what your past opinion has been. And this guesswork is coherent only if it satisfies (4b). For instance, you cannot coherently say, "Oh, I have forgotten my date with the dentist; I guess it's next Wednesday. But I think that yesterday I was still quite sure that it was next Tuesday." You may reversely take this as support for the backward-looking reflection principle (2).

One may think that there is a difference between (4a) and (4b). (4a), but not (4b), is grounded, as it were, in experience. The standard instantiation of (4a) is simple conditionalization: You learn exactly one member of a partition of evidential (worldly) propositions about which you have some prior expectations. And since your possible posteriors are just the conditionalization of your prior with respect to these evidential propositions, you have the very same expectations about these posteriors. Such a grounding is, however, entirely missing in the case of (4b). You may remember your past probabilities, but if

17 If we suppress the additional condition about superiority and stick to the shorter notation used in (4a), (3a) says $P_0(A \mid \pi_1) = \pi_1(A)$. The formula of the total probability says that $P_0(A) = \Sigma\, P_0(A \mid \pi_1) \cdot P_0(\pi_1)$. By inserting the first equation into the second, we get (4a).

18 Here, de Finetti's heritage is particularly salient. Recall his famous representation theorem, saying that your (prior) probabilities are symmetric or exchangeable (as they should be) if and only if they are a unique mixture of all the statistical hypotheses they might converge to.

you have forgotten them, your present opinion about them is mere guesswork without such grounding.

However, the case is not as asymmetric as it seems. (4a) is not only made for simple conditionalization. It holds as well, e.g., for Jeffrey conditionalization, where learning results in some new posterior probabilities for the partition of evidential propositions. And then the posterior is not grounded in a specific evidential proposition but in your possibly vague seemings concerning this evidential partition. Then, however, your expectations about these seemings are not much better off than in the past-oriented case. So, (4a + b) is auto-epistemic business justified by full reflection (3). Such specific grounding is welcome but not required.

5 Reflection Applied to Auto-Epistemic Propositions

To check out further consequences, we must attend to the way Hild (1998a) presents the principles. He tacitly assumes an innocent-looking generalization; i.e., from the outset, he applies van Fraassen's principle (1) also to auto-epistemic (and mixed) propositions A. Let's briefly consider in this section what happens when we thus generalize our principles (3a–e) and drop their restriction to worldly propositions. Note that this also requires us to drop the restriction of the possible probability measures π to worldly propositions. Hence, equations like $\pi = Q$ can now be literally and not only sloppily true.

A first consequence would be that the synchronic reflection principle (3e) turns out to be equivalent to what Hild calls *auto-epistemic transparency*:[19]

(5) $P_i(\{\pi \mid \pi = P_j\}) = 1 \quad (i = 0, 1)$.[20]

In other words, in each second-order epistemic state, also reflecting on your first-order state, you know, or are sure, what your present first-order state is. In doxastic logic, this is sometimes called "positive introspection" or the BB thesis, "If you believe that p, then you believe that you believe that p," first discussed in Hintikka (1962, 123ff.), amply attacked and amply defended. Let us not engage in this discussion now.[21]

19 Christensen (2007) and Weisberg (2007) split this up into two principles called confidence and accuracy by Christensen and luminosity and transparency by Weisberg.

20 *Proof*: For the one direction, take A in (3e) to be the auto-epistemic proposition $\{\pi \mid \pi = Pi\}$. Reversely, P_i is identical to P_i conditional on a proposition with probability 1.

21 Besides the arguments referred to in footnote 15, Christensen (2007) casts doubt on auto-epistemic transparency by questioning that we have certain knowledge about our precise subjective probabil-

doi: 10.48106/dial.v77.i1

Another consequence of the reflection principles (3a + e) thus extended is *perfect memory*:

(6) $P_1(\{\pi_0 \mid \pi_0 = P_0\}) = 1$, given that P_1 is taken to be superior to P_0.[22]

This is a suspiciously strong consequence. However, given the extension of the principles to auto-epistemic propositions, one could say that whenever I have become uncertain about my former epistemic state, I have forgotten my former attitude towards some proposition. So, my uncertainty must be the result of an unfavorable change.

By a similar proof, the extended (3c) + (3e) implies an analogous principle of perfect foresight. My prior knows for sure what my inferior posterior will be. This is obviously absurd. Hence, the auto-epistemic extension of the backward-looking principle (3c) must be rejected. The proof of (6) displays where the analogy breaks down. Inserting $\{\pi_1 \mid \pi_1 = P_1\}$ for A in (3c) would mean that I would trust my former superior P_0 concerning my present inferior state. But regarding my own present state, I am always in an optimal epistemic position, as confirmed by auto-epistemic transparency (5); in this respect, I need no lessons from my better self. My position may be inferior only with respect to worldly propositions (and auto-epistemic propositions referring to other times). This is why we must not extend (3c) to auto-epistemic propositions (or at least not to the simultaneous ones).

The envisaged extension also helps a bit regarding the relations among the five parts of (3). That is, we may see now that the extended (3a + e) entails not only (6) but also (3d) simply because (6) says that the condition of (3d) has probability 1. For the same reason as before, we must not exploit this observation for a corresponding derivation of (3b) from (3c + e); (3b) seems to remain independent.

ities. However, this rather questions the representation of epistemic states by precise probabilities, which we have assumed at the outset of this paper. That is, I tend to assume that beliefs and epistemic states, in general, are conscious mental states (in the sense of intentional or higher-order consciousness), similar to phenomenally conscious pains. So, if I do not know my precise probabilities, I don't have them, just as I don't have pains when I am not aware of them. It is this assumption, I think, that is the motivation behind the BB thesis and its kin.

22 *Proof*: Take A in (3a) to be the auto-epistemic proposition $\{\pi_0 \mid \pi_0 = P_0\}$. Thus, $P_0(\{\pi_0 \mid \pi_0 = P_0\} \mid \{\pi_1 \mid \pi_1 = P_1\}) = P_1(\{\pi_0 \mid \pi_0 = P_0\})$, given that P_1 is taken to be superior to P_0. Auto-epistemic transparency (5) says that $P_0(\{\pi_0 \mid \pi_0 = P_0\}) = 1$. Hence, $P_0(\{\pi_0 \mid \pi_0 = P_0\} \mid \{\pi_1 \mid \pi_1 = P_1\}) = 1$ as well. So, finally, $P_1(\{\pi_0 \mid \pi_0 = P_0\}) = 1$, given that P_1 is taken to be superior to P_0 (cf. Hild 1998a, 353).

Moreover, we may note that Hild's extension strengthens the relation between reflection and iteration. We observed already that (3a) and (3c), respectively, entail (4a) and (4b). With the extension, we may reverse the entailment: given auto-epistemic transparency (5) or the equivalent synchronic reflection principle (3e), the iteration principle (4a) implies the forward-looking reflection principle (3a).[23] Thus, under the same assumptions, the reverse iteration principle (4b) entails the backward-looking (3c), as seems unobjectionable.

Finally, this extension helps us to provide a formal proof of my informally justified claim in section 2 that the forward- and backward-looking principles (1) and (2) can apply simultaneously only in the case of non-change. Given that (1) and (2) apply also to auto-epistemic propositions, we have:

(7) Given auto-transparency (5), if (a) $P_0(A \mid \{\pi_1 \mid \pi_1 = P_1\}) = P_1(A)$ and (b) $P_1(A \mid \{\pi_0 \mid \pi_0 = P_0\}) = P_0(A)$ hold for all propositions A, then $P_0 = P_1$.[24]

In conclusion, we have found a slight divergence among our principles in this extension, a divergence we could justify. In the main, however, the parallel between the forward-looking and the backward-looking perspective and thus between the parts of the full reflection principle (3) stands. We have not discovered any incoherence.

6 The Full Reflection Principle and the Expert Principle

In section 1, I pointed already to the expert principle, which is the most common generalization of van Fraassen's principle (1). It may seem that the full reflection principle (3) is just another special case of the expert principle.[25] Yes, almost. At least, this holds for the backward-looking principle (2). Here, my better-informed past self may be taken to be an expert for my present forgetful self. However, not all cases of (3) are special cases of the expert principle. Let P in (3) be my probability measure and P' that of my neighbor. When I take my neighbor to be better informed, to be in a superior epistemic

23 *Proof*: We have $P_0(A \text{ and } \{\pi_1 \mid \pi_1 = P_1\}) = \Sigma\, \pi(A \text{ and } \{\pi_1 \mid \pi_1 = P_1\}) \cdot P_0(\pi)$ (by the extended (4a), where the sum is taken over all possible posteriors π) $= P_1(A) \cdot P_0(\{\pi_1 \mid \pi_1 = P_1\})$ (by auto-transparency, because $\pi(\{\pi_1 \mid \pi_1 = P_1\}) = 1$ only for $\pi = P_1$, and otherwise $= 0$) (cf. Hild 1998a, 354).

24 *Proof*: We have $P_1(A) = P_0(A \mid \pi_1 = P_1)$ (due to (a)) $= P_0(A \mid \pi_1 = P_1, \pi_0 = P_0)$ (due to auto-transparency (5)) $= P_1(A \mid \pi_0 = P_0)$ (by applying (a) conditional on $\pi_0 = P_0$) $= P_0(A)$ (due to (b)).

25 I am grateful to a reviewer for raising this issue.

doi: 10.48106/dial.v77.i1

position (concerning a certain field), then I listen to her (in the sense of obeying (3)); these are the cases (3a) and (3c), where the temporal relation between me and my neighbor is irrelevant. But when I take her to be less well informed or in an inferior epistemic position, I do not listen to her; these are the cases (3b) and (3d). This seems to go without saying. Christensen (2000, 358) takes this for granted, too. Strictly speaking, though, it is not part of the expert principle, which says only how to deal with people taken to be at least as well-informed.

Of course, it would be no problem to pair the expert principle with a "non-expert principle," saying that, rationally, we don't listen to persons we take to be in an inferior epistemic position. However, this would still leave us with very incomplete principles. The no-neutrality condition—which was plausible in the intrasubjective case, at least when we can divide up complex epistemic changes into unidirectional steps—has no analog with respect to experts. Most of my fellow humans are neither better nor less well-informed than me; their epistemic state is just incomparable to mine. And then both the expert and the non-expert principle are silent. This is not an objection. It is hard to give any recommendations for the incomparable cases. But it is our daily business to somehow deal with them.

One may think[26] that we can apply the treatment of intrasubjective incomparabilities suggested in section 3 also to the interpersonal case of experts. However, this is not so easy. In the intrapersonal case, we had to refer, it seemed, to counterfactual epistemic states that the subject would be in had certain unfavorable changes not occurred. This might be manageable. In the case of my incomparable neighbor, however, the corresponding counterfactual question would be which favorable changes he would have to undergo and which unfavorable changes to avoid till I could acknowledge him to be an expert, i.e., to be in an equal or superior epistemic position concerning the issue at hand. This is a much more sweeping and indeterminate counterfactual question. The point is that the superiority and inferiority of epistemic positions are clearly assessable on the basis of intrasubjective favorable or unfavorable changes. But it is very hard to assess as such, as an interpersonal comparison would require.

So, we have a tension here. While van Fraassen's principle (1) was clearly a special case of the expert principle, the subsumption of the full principle (3) is at least doubtful. In fact, such a subsumption was not intended in the begin-

26 Suggested by the same reviewer.

ning. Gaifman (1986) proposed the expert principle as a formal generalization of the reflection principle, not as a substitute for the latter's epistemological role. As such it operates only in so-called time-slice epistemology (Moss 2015) or time-slice rationality (Hedden 2015). A basic assumption of this approach is called impartiality:[27] "In determining how you rationally ought to be at a time, your beliefs about what attitudes you have at other times play the same role as your beliefs about what attitudes other people have" (Hedden 2015, 9). If so, it is clear that reflection principles are unnecessarily restrictive and that the expert principle completely takes over the epistemological role of the full reflection principle (given an additional "non-expert principle").

In his defense of time-slice rationality, Hedden (2015, chaps. 8–9) makes crucial use of an assumption called uniqueness (by Feldman 2007): "Given a body of total evidence, there is a unique doxastic state that it is rational to be in" (Hedden 2015, 130). This reveals an entirely different picture of normative epistemology than the one pursued here. It is that there is a unique prior—*the ur*-prior, as it were—and then all epistemic change is due to a change of the body of total evidence.[28] Given this, we do not need any diachronic rules for epistemic change. We always refer back to the *ur*-prior and then consider how the given body of total evidence operates on it. All change is in that body and only there; the body may get larger (through learning) or smaller (through forgetting).[29] My fellow humans are in the very same situation. They rationally proceed from the very same *ur*-prior, and they differ from me only in their total evidence. This is why they count just as much as my future or past epistemic states. Or rather, nobody counts; only the bodies of total evidence count. Certainly, this would simplify our epistemological business considerably.

The assumption of uniqueness also makes the notion of an expert very easy. Among rational subjects, a is an expert for b simply if a has at least as much evidence as b. If a's and b's evidence only overlap, they are in incomparable states. But joining their evidence would result in an expert for both. With

27 The other basic assumption is synchronicity: "All rationality requirements are synchronic" (Hedden 2015, 9). See also Moss (2015, 177) for a statement of the two principles. The principle of impartiality seems to have been stated first by Christensen (2000, 363–364). Moss (2015, 178) and Hedden (2015, 56) happily observe that van Fraassen's reflection principle satisfies the assumption of synchronicity insofar as it speaks only about P_0 and its conjectures about P_1. Of course, this observation carries over to the other versions.

28 Or the total evidence need not appeal to any prior at all. But then the *ur*-prior is unique as well, namely, empty.

29 This picture also motivates Titelbaum's (2013) framework of gain and loss of certainties.

doi: 10.48106/dial.v77.i1

this easy notion of an expert, the above idea of paralleling intersubjective incomparabilities with intrasubjective incomparable changes might be less problematic.

The alternative is to deny uniqueness. Hedden (2015, 129) calls this permissiveness. But what is the dialectic situation here? Pace Hedden, it is not that one must defend uniqueness and the other permissiveness (by showing two *ur*-priors to be equally rationally acceptable). In my view, the burden of proof lies with the defender of uniqueness. And a proof should constructively indicate what the unique *ur*-prior looks like. The literature is not so promising. The only positive attempt I know of is objective Bayesianism, as proposed by Williamson (2005).[30] By contrast, I tend to take centuries of skepticism to suggest that such a proof will fail.[31] Obviously, this is too big an issue to be discussed now.[32] The point is only this: In the absence of such a proof, we should not proceed from the assumption of uniqueness. A positive defense of permissiveness is not really required.[33]

Kelly (2014) usefully distinguishes statements of uniqueness that have interpersonal import, as he calls it, from those that have not. Intrasubjective uniqueness only requires that, given my background or my personal *ur*-prior, I have only one rational way of responding to the evidence. I have no quarrel with this. However, uniqueness with interpersonal import requires that there is only one and the same rational way for everybody to respond to the evidence. This is the version intended and critically discussed above.

The question, then, is how to pursue normative epistemology without the assumption of uniqueness. Just in the way it is done traditionally, and here

30 To be precise, Williamson does not need to refer to an *ur*-prior. He rather proposes a unique way of responding to any given total body of evidence, which does not depend on an underlying *ur*-prior. Rather, the *ur*-prior would be the response to the empty evidence.

31 Hedden (2015, 134) himself (as well as Kelly 2014, 309) points to the alleged failure of Carnap's (1962) project of inductive logic, which also started out searching for the unique prior. I should add, though, that in chap. 8, Hedden admits that uniqueness may force one to allow for indeterminate and/or imprecise probabilities. However, considering other epistemic formats shifts the discussion still further. I have explained why I focus here on precise probabilities only.

32 I admit that the issue can also be discussed in the abstract without constructive proposals for an *ur*-prior. See, e.g., the exchange between White (2014) and Kelly (2014). It is clear that my sympathies lie here with Kelly. However, he is still too obliging, I find; he does not raise the point about the burden of proof.

33 Recall also this: In the final section of his (1980), Lewis discovers a tension between his Principal Principle and Humean Supervenience. He considers resolving the tension by assuming what is now called uniqueness. But he shies away from this solution, which he finds "not very easy to believe." As is well known, he modified the Principal Principle later on.

as well, namely, by stating synchronic principles of epistemic rationality and diachronic principles. The latter can only refer to a subject's prior and posterior and a piece of total evidence in between, but *not* to an *ur*-prior and a body of total evidence reaching back to the indefinite time of the *ur*-prior.[34] In this conception, the reflection principles, as discussed here, have a natural place, and the goal of stating a complete dynamic is important, while atemporal expert principles do not directly add to it and need not be complete. Given uniqueness, we can also distinguish inferior and superior epistemic positions simply by looking at the size of the bodies of total evidence underlying them. However, when looking at epistemic dynamics in the traditional way, then, as argued, we must also classify single changes as favorable and unfavorable (and if we cannot do this objectively, we leave it to the subject herself, as proposed here). As mentioned, favorable and unfavorable changes do not only consist in gaining and losing evidence or certainties; they may take various other forms not easily subsumed under the picture motivated by uniqueness. In the perspective pursued here, expert principles become relevant only because we take listening to experts to induce favorable change, unlike listening to non-experts. This is why I think that the reflection principles have an independent value. They can be substantially subsumed under the expert principles only within a questionable epistemological picture.*

Wolfgang Spohn
0000-0002-3213-8907
Department of Philosophy, University of Konstanz
wolfgang.spohn@uni-konstanz.de

References

ARNTZENIUS, Frank. 2003. "Some Problems for Conditionalization and Reflection." *The Journal of Philosophy* 100(7): 356–370, doi:10.5840/jphil2003100729.

34 When we model a dynamic process, in physics, meteorology, or wherever, we do it in the form of a law saying how one state of the system modeled changes into the subsequent state, possibly under the influence of external factors, and we can do this in discrete or in continuous time. So, this is also the natural format for normative epistemology as well, where we try to say what a rational epistemic dynamic should look like.

* I am deeply indebted to three reviewers, one from the *Journal of Philosophy* and two from *Dialectica*. They considerably helped me in various respects. I also acknowledge support by the German Research Foundation, Grant EXC 2064/1, No. 390727645, and Grant SP279/21-1, No. 420094936.

BINKLEY, Robert W. 1968. "The Surprise Examination in Modal Logic." *The Journal of Philosophy* 65(5): 127–136, doi:10.2307/2024556.

BRIGGS, Rachael A. 2009. "Distorted Reflection." *The Philosophical Review* 118(1): 59–85, doi:10.1215/00318108-2008-029.

CARNAP, Rudolf. 1950. *Logical Foundations of Probability*. Chicago, Illinois: University of Chicago Press. Second edition: Carnap (1962).

—. 1962. *Logical Foundations of Probability*. 2nd ed. Chicago, Illinois: University of Chicago Press. First edition: Carnap (1950).

CHRISTENSEN, David. 1991. "Clever Bookies and Coherent Beliefs." *The Philosophical Review* 100(2): 229–247, doi:10.2307/2185301.

—. 2000. "Diachronic Coherence Versus Epistemic Impartiality." *The Philosophical Review* 109(3): 349–371, doi:10.1215/00318108-109-3-349.

—. 2007. "Epistemic Self-Respect." *Proceedings of the Aristotelian Society* 107: 319–337, doi:10.1111/j.1467-9264.2007.00224.x.

—. 2010a. "Higher-Order Evidence." *Philosophy and Phenomenological Research* 81(1): 185–215, doi:10.1111/j.1933-1592.2010.00366.x.

—. 2010b. "Rational Reflection." in *Philosophical Perspectives 24: Epistemology*, edited by John HAWTHORNE, pp. 121–140. Hoboken, New Jersey: John Wiley and Sons, Inc., doi:10.1111/j.1520-8583.2010.00187.x.

FELDMAN, Fred. 2004. *Pleasure and the Good Life: Concerning the Nature, Varieties, and Plausibility of Hedonism*. Oxford: Oxford University Press, doi:10.1093/019926516X.001.0001.

—. 2007. "Précis of Feldman (2004)." *Philosophical Studies* 136(3): 405–408, doi:10.1007/s11098-006-9039-y.

DE FINETTI, Bruno. 1937. "La prévision, ses lois logiques, ses sources subjectives." *Annales de l'Institut Henri Poincaré* 7: 1–68. Translated as de Finetti (1964).

—. 1964. "Foresight: Its Logical Laws, Its Subjective Sources." in *Studies in Subjective Probability*, edited by Henry E. KYBURG Jr. and Howard E. SMOKLER, pp. 93–158. Hoboken, New Jersey: John Wiley and Sons, Inc. Translation of de Finetti (1937), reprinted in Kyburg and Smokler (1980, 53–118).

VAN FRAASSEN, Bas C. 1984. "Belief and the Will." *The Journal of Philosophy* 81(5): 235–256, doi:10.2307/2026388.

—. 1995. "Belief and the Problem of Ulysses and the Sirens." *Philosophical Studies* 77(1): 7–37, doi:10.1007/bf00996309.

GAIFMAN, Haim. 1986. "A Theory of Higher Order Probabilities." in *TARK 1986. Theoretical Aspects of Reasoning about Knowledge: Proceedings of the First Conference*, edited by Joseph Y. HALPERN, pp. 275–292. San Francisco, California: Morgan Kaufmann Publishers. Reprinted in Skyrms and Harper (1988, 191–219), doi:10.1016/B978-0-934613-04-0.50023-5.

GIGERENZER, Gerd. 2007. *Gut Feelings: Short Cuts to Better Decision Making*. London: Penguin Books.

GOLDSTEIN, Michael. 1983. "The Prevision of a Prevision." *Journal of the American Statistical Association* 78(384): 817–819, doi:10.2307/2288190.

HEDDEN, Brian. 2015. *Reasons Without Persons: Rationality, Identity, and Time*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780198732594.001.0001.

HILD, Matthias. 1998a. "Auto-Epistemology and Updating." *Philosophical Studies* 92(3): 321–361, doi:10.1023/a:1004229808144.

—. 1998b. "The Coherence Argument Against Conditionalization." *Synthese* 115(2): 229–258, doi:10.1023/a:1005082908147.

HINTIKKA, Jaakko. 1962. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca, New York: Cornell University Press.

KELLY, Thomas. 2014. "Evidence Can Be Permissive." in *Contemporary Debates in Epistemology*, edited by Matthias STEUP, John TURRI, and Ernest SOSA, 2nd ed., pp. 298–311. Contemporary Debates in Philosophy n. 3. Malden, Massachusetts: Wiley-Blackwell. First edition: Sosa and Steup (2005), doi:10.1002/9781394260744.ch12.

KYBURG, Henry E., Jr. and SMOKLER, Howard E., eds. 1980. *Studies in Subjective Probability*. 2nd ed. Huntington, New York: Robert E. Krieger Publishing Co.

LEWIS, David. 1980. "A Subjectivist's Guide to Objective Chance." in *Studies in Inductive Logic and Probability: Volume II*, edited by Richard C. JEFFREY, pp. 263–294. Berkeley, California: University of California Press. Reprinted, with a postscript (Lewis 1986b), in Lewis (1986a, 83–113), doi:10.1525/9780520318328-009.

—. 1986a. *Philosophical Papers, Volume 2*. Oxford: Oxford University Press, doi:10.1093/0195036468.001.0001.

—. 1986b. "Postscript to Lewis (1980)." in *Philosophical Papers, Volume 2*, pp. 114–132. Oxford: Oxford University Press, doi:10.1093/0195036468.003.0004.

MAHTANI, Anna. 2017. "Deference, Respect and Intensionality." *Philosophical Studies* 174(1): 163–183, doi:10.1007/s11098-016-0675-6.

MOSS, Sarah. 2015. "Time-Slice Epistemology and Action Under Indeterminacy." in *Oxford Studies in Epistemology*, volume V, edited by Tamar Szabó GENDLER and John HAWTHORNE, pp. 172–194. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780198722762.003.0006.

ROTT, Hans. 2009. "Shifting Prorities: Simple Representations for Twenty-Seven Iterated Theory Change Operators." in *Towards Mathematical Philosophy: Papers from the Studia Logica Conference* Trends in Logic IV, edited by David C. MAKINSON, Jacek MALINOWSKI, and Heinrich Theodor WANSING, pp. 269–296. Trends in Logic n. 28. Dordrecht: Springer, doi:10.1007/978-1-4020-9084-4_14.

SCHOENFIELD, Miriam. 2012. "Chilling Out on Epistemic Rationality: A Defense of Imprecise Credences (and Other Imprecise Doxastic Attitudes)." *Philosophical Studies* 158(2): 197–219, doi:10.1007/s11098-012-9886-7.

SKYRMS, Brian. 1980. *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. New Haven, Connecticut: Yale University Press.

SKYRMS, Brian and HARPER, William L., eds. 1988. *Causation, Chance, and Credence: Proceedings of the Irvine Conference on Probability and Causation, Volume 1*. The University of Western Ontario Series in Philosophy of Science n. 41. Dordrecht: Kluwer Academic Publishers, doi:10.1007/978-94-009-2863-3.

SOSA, Ernest and STEUP, Matthias, eds. 2005. *Contemporary Debates in Epistemology*. 1st ed. Contemporary Debates in Philosophy n. 3. Boston, Massachusetts: Blackwell Publishers. Second edition: Steup, Turri and Sosa (2014).

SPOHN, Wolfgang. 1978. *Grundlagen der Entscheidungstheorie*. Kronberg/Ts.: Scriptor Verlag, urn:nbn:de:bsz:352-135468.

—. 1988. "Ordinal Conditional Functions: A Dynamic Theory of Epistemic States." in *Causation in Decision, Belief Change, and Statistics. Proceedings of the Irvine Conference on Probability and Causation, vol. 2*, edited by William L. HARPER and Brian SKYRMS, pp. 105–134. The University of Western Ontario Series in Philosophy of Science n. 42. Dordrecht: Kluwer Academic Publishers, doi:10.1007/978-94-009-2865-7_6.

—. 2010. "Chance and Necessity: From Humean Supervenience to Humean Projection." in *The Place of Probability in Science: In Honor of Ellery Eells (1953–2006)*, edited by James H. FETZER, pp. 101–131. Boston Studies in the Philosophy of Science n. 284. Dordrecht: Springer, doi:10.1007/978-90-481-3615-5_5.

—. 2017. "The Epistemology and Auto-Epistemology of Temporal Self-Location and Forgetfulness." *Ergo* 4(13): 359–418, doi:10.3998/ergo.12405314.0004.013.

STEUP, Matthias, TURRI, John and SOSA, Ernest, eds. 2014. *Contemporary Debates in Epistemology*. 2nd ed. Contemporary Debates in Philosophy n. 3. Malden, Massachusetts: Wiley-Blackwell. First edition: Sosa and Steup (2005), doi:10.1002/9781394260744.

TALBOTT, William J. 1991. "Two Principles of Bayesian Epistemology." *Philosophical Studies* 62(2): 135–150, doi:10.1007/bf00419049.

TITELBAUM, Michael G. 2013. *Quitting Certainties: A Bayesian Framework Modeling Degrees of Belief*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199658305.001.0001.

WEISBERG, Jonathan. 2007. "Conditionalization, Reflection, and Self-Knowledge." *Philosophical Studies* 135(2): 179–197, doi:10.1007/s11098-007-9073-4.

WHITE, Roger. 2014. "Evidence Cannot Be Permissive." in *Contemporary Debates in Epistemology*, edited by Matthias STEUP, John TURRI, and Ernest SOSA, 2nd ed., pp. 312–323. Contemporary Debates in Philosophy n. 3. Malden, Massachusetts: Wiley-Blackwell. First edition: Sosa and Steup (2005), doi:10.1002/9781394260744.ch12.

WILLIAMSON, Jon. 2005. *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780198530794.001.0001.

Our Naïve Representation of Time and of the Open Future

Batoul Hodroj, Andrew James Latham & Kristie Miller

It's generally thought that we naïvely or pre-theoretically represent the future to be open. While philosophers have modelled future openness in different ways, it's unclear which, if any, captures our naïve sense that the future is open. In this paper, we focus on just one way the future might count as being open: by being nomically open. We empirically investigate whether our naïve representation of the future as open is partly constituted by representing the future as nomically open. We also investigate the connection between our naïve representation of the future as open and our representation of time. One of the purported advantages of the growing block theory of time is that it captures our naïve sense that the future is open and the past closed. We investigate whether there is an explanatory connection between people representing the future as being nomically open and representing our world as a growing block. We also reflect on the implications of our findings for theorising about future openness and temporal ontology.

It's often thought that our intuitive or pre-reflective view of the world is one in which, in some sense or other, the future is open.[1] It has also been thought that our intuitive, pre-reflective, or folk view of the world is one in which the totality of our world grows as new being comes into existence in the present moment and then becomes past as yet more being comes into existence.[2] This latter view is the view that our world is a *growing block*.[3]

1 Callender (2017) takes this to be part of the manifest image; Ismael (2012) likewise.
2 See Forbes (2016). Latham, Miller and Norton (2021b) confirmed empirically that, of the ~70% of people who are temporal dynamists, the most popular view is the growing block view.
3 Defenders of this view include Broad (1923, 1938), Forbes (2016), Correia and Rosenkranz (2018), Tooley (1997), and Forrest (2004).

In what follows, rather than talking about pre-reflective or folk views, we will talk of *naïve representations* of the world. As we will understand them, naïve representations are contentful mental states, i.e., representations of various aspects of our world that are not informed by (or, at least, are largely not the product of engagement with) current science or philosophy. These are folk views, folk theories, or folk models of aspects of the world. These representations may be tacit in the sense that the people whose representations they are may not be able to specify the content of the representation when asked. Nevertheless, we take it that these representations guide people's behaviours (linguistic and otherwise) and that we can probe their content by giving people tasks that require them to use those representations.

We are interested in two sorts of naïve representations. The first is our *naïve representation of the future*; the second is our *naïve representation of time*. Ultimately, we will be interested in whether these representations are connected.

We will take the claim that our pre-reflective view of the world is one in which the future is open to be the claim that we naïvely represent the future as open. Philosophers have offered various accounts of the open future. In fact, we can (and should) distinguish at least two rather different projects with which philosophers are engaged. The first of these aims to model the open future. On one natural interpretation of such a project, which we will call *the capturing project*, the aim is to work out which model of, or theory of, the open future is the one that best captures our intuitive sense that the future is open. As we construe this project, the aim is to offer a model of the open future that best captures our naïve representation of future openness. The second project, which we will call *the explanatory project*, focuses on explaining various "open future" practices (conceived of very broadly) and attempts to explain why it is that we have such practices and what it is about our world that grounds our having such practices. These practices might include (but not be limited to) practices of deliberating about the future but not the past; taking ourselves to be able to causally intervene on the future but not the past; having a certain kind of phenomenology in which the future feels, or seems, to us to be open in the way the past does not; taking ourselves to have a kind of access to past states that we do not have to future ones; and so on.

These two projects might be connected or not. It might be that what explains why we have the open future practices we do is the very thing that, in fact, captures our naïve representation of the future. In that case, we will say that our naïve representation of the future is *vindicated*. Alternatively, it could

be that what explains our open future practices does not capture our naïve representation of the future as open. To see this, consider several of the views philosophers have put forward as models of the open future and suppose these are claims about our naïve representation of future openness.

The first view models future openness in terms of *alethic openness*. On this view, our naïve representation of the open future consists in, or at least includes, our representing that (some or all) future-tensed contingent statements fail to take a determinate truth-value (see, for instance, Markosian 1995; Williams 2008 (unpublished); MacFarlane 2003; and Tooley 1997). The second of these is *epistemic openness*. On this view, our naïve representation of the future being open consists in, or at least includes, our representing that we have epistemic access to the future only by making predictions and forming intentions and not by having records of what will happen (see, for instance, Lewis 1979). The third is *nomic openness*. On this view, our naïve representation of the future being open consists in, or at least includes, our representing that future-directed indeterminism is true. There are multiple ways the future could go, consistent with how it has already gone (Belnap 1992, 2005; MacFarlane 2003, 2008; and McCall 1994).

It could be that our naïve representation of the future as open consists in our representing the future as being open in some or all of these ways.[4] Suppose it were to turn out that our naïve representation of future openness consists entirely in representing the future to be alethically open. Suppose, however, that our world is not, in fact, alethically open. Still, *something* explains why we have the open future practices that we do. It might be that the fact that there is an epistemic asymmetry between past and future is what explains our having these practices. It might even be that the world being this way legitimises or makes those practices rationally permissible (or obligatory). Still, it will turn out that what explains our having the open future practices we do does not *vindicate* our naïve representation of the future as open.

This paper will have nothing to say about why we have the open future practices we do. We set aside the explanatory project and focus entirely on the question of what our naïve representation of future openness consists in. This is a vital first step if we are interested in the question of whether what it is that explains our practices (whatever that might be) vindicates our naïve representation of the future as open.

4 This is not to say that these are the only such ways. For a discussion of the ways in which we could model openness, see Torre (2011) and Markosian (1995).

doi: 10.48106/dial.v77.i1

Some work in this area has already been undertaken. Previous research by Hodroj et al. (2023) suggests that our naïve representation of the future as open, at least partly, consists in our representing the future to be alethically open. So, in this paper, we focus on nomic openness. We will suppose that a world is nomically open just in case that world is future-wise indeterministic. That is, a world, w, is nomically open just in case for any time t in w, it is not the case that a complete specification of the way the world is at t, in conjunction with the laws of nature of w, logically entails the way the world is at all times later than t. This leaves open that w may or may not be past-wise nomically open: that is, whether the way the world is at t, in conjunction with the laws of nature, logically entails the way the world is at all times earlier than t. Then, we are interested in whether our naïve representation of the future involves our representing the future to be nomically open.

We are also interested in the connection between our naïve representation of the future as open and our naïve representation of the temporal dimension. That is because it has been suggested that part of what explains why the growing block theory is intuitively plausible is that we naïvely represent the future as open, and the growing block theory better captures, or better accords with, this (see, for instance, Briggs and Forbes 2012; Forbes 2016; Grandjean 2021, 2022; and Correia and Rosenkranz 2018).

According to the growing block model of time, past events and objects exist, but future ones do not. There is a set of events that are objectively present, and these are the events that sit at the end of the block looking out into the non-existent future. The temporal passage consists in the coming into existence of a new being on the edge of reality, where that new being becomes the objective present until more being comes to exist (at which point it becomes part of the objective past). Hence, the growing block theory is a version of the A-theory on which there exists robust temporal passage: there is a fact of the matter as to which events are present and which those are changes. By contrast, the block universe theory is a version of the B-theory. On this view, past, present, and future events/objects exist on a four-dimensional manifold and bear unchanging relations of earlier-than, later-than, and simultaneous-with to one another.[5] None of these events is singled out as objectively present, and so time does not robustly pass since there is no change in which events are objectively present.

5 This, of course, is also true of the moving spotlight theory, which is a version of dynamism. However, on that view, unlike the block universe view, there is a single set of events singled out as objectively present.

Unlike other models of time, the growing block theory has a built-in asymmetry between the past and the future. The past exists and is located somewhere in space-time, whereas the future is yet to happen and does not exist. By contrast, presentism holds that *neither* the future nor the past exists, and the block universe theory holds that *both* the future and the past exist. The moving spotlight theory also holds that both future and past exist but holds that some events are objectively present (namely, those on which the spotlight of presentness shines, as it were) and that which events those are changes as the present moves.[6]

This asymmetry has been hypothesised to better capture people's intuitive sense that the future is open and the past is closed than do views that lack this asymmetry.[7]

Following Latham, Miller and Norton (2021b), we take a naïve representation of time to be a (probably tacit) representation of time and temporal ontology in our world. People's naïve representation of time might be closer to one or another of the models of time that philosophers engage with.

Following Hodroj et al. (2023), we can distinguish three aspects of the idea that the growing block theory better accommodates people's intuitive sense that the future is open.

First, according to the *vindication claim*, our naïve representation of future openness has a content that is vindicated if our world is a growing block. The narrow version of the vindication claim that will be of interest to us in this paper is the claim that our naïve representation of future openness has a content that is vindicated if our world is a growing block and is not vindicated if our world is a block universe. Henceforth, we will call this the *narrow vindication claim*.

One might be particularly interested in the narrow vindication claim if one thinks that if the growing block vindicates our naïve representation of the open future and the block universe view does not, this gives us a reason (albeit defeasible) to prefer the former over the latter.

Second, according to the *reason claim*, people believe, perhaps tacitly, that the fact that a world has an open future is a reason to think that that world is a growing block world rather than a block universe world.

6 For empirical research into people's naïve views of time, see Latham, Miller and Norton (2021a).
7 Something that Grandjean (2021, 2022) and Correia and Rosenkranz (2018) point to.

doi: 10.48106/dial.v77.i1

Third, according to the *explanation claim*, people naïvely represent our world to be a growing block because they naïvely represent the future to be open.

Our aim is not to investigate all these claims in their full generality but rather to investigate certain aspects of these claims as they pertain to nomic openness.

Consider, first, the narrow vindication claim. In order to evaluate the narrow vindication claim, we would need to know the content of our naïve representation of future openness. This paper will speak to the issue of whether our naïve representation of future openness is partly constituted by our representing it to be nomically open. So, it will provide the beginnings of the sort of account we would need to determine whether the narrow vindication claim (and indeed the vindication claim itself) is true.

Next, consider the reason claim. We investigate whether people take the fact that a world is nomically open to be a reason to think that it is a growing block world rather than a block universe world. We also investigate a particular view about what this reasoning might consist in. According to this view, people reason from their ability to deliberate and to act freely to the idea that the future is nomically open. They then reason from the nomic openness of the future to the idea that future events do not exist because they think that if future events did exist "out there in spacetime," then those events must be determined because facts about them already obtain. But in representing that future events do not exist and will later come to exist, one represents one crucial element of the growing block view. Thus, it might be that by representing the world as nomically open, people come to represent it as a growing block.

Now, to be clear, we are not endorsing either stage of this reasoning from freedom/deliberation to nomic openness or from nomic openness to the non-existence of future events (indeed, this last inference is clearly invalid). We are merely hypothesising that people (likely tacitly) reason in something like this manner, and so they take the presence of nomic openness in a world to be a reason to think that the world is a growing block world rather than a block universe world. We will call the claim that people reason in this way the *deliberative reasoning claim*.

Finally, according to the version of the explanation claim that we investigate here, the fact that people naïvely represent the future as nomically open is part of what explains why they represent our world to be a growing block. Notice that the reason claim and the explanation claim can come apart. It

could be that people naïvely represent our world as a growing block because they represent it as nomically open, even though they do not tacitly suppose that the latter is a reason to think our world is a growing block (perhaps there is a common cause of both representations). Equally, it could be that people *do* think that a world being nomically open is a reason to think it is a growing block rather than a block universe, but this does not, in fact, explain why people think our world is a growing block world (either because they don't think it is a growing block, or because they don't think our world is nomically open, or because other factors completely swamp this reason and do all the explanatory work).

In experiment 1, we seek to determine whether people's naïve representation of the future involves nomic openness. We present participants with two *nomic vignettes*: one that describes a nomically open world and one that describes a nomically closed world. Having seen the two vignettes, participants are then asked which world is most like our world (nomically open or closed). Our first hypothesis (H1) is that more people will judge that the nomically open world is more like our world than the nomically closed world. If most people naïvely represent the future as nomically open, then it seems reasonable to say that their naïve representation of the future as open consists, at least in part, in them representing the future in this matter.

Participants are then presented with two *time vignettes*, one describing a growing block world and one describing a block universe world. They are then asked which world is most like our world. We predicted (H2) that more people would judge that our world is like the growing block world than the block universe world. This hypothesis is motivated by previous work on the way that people naïvely represent time, including that of Latham, Miller and Norton (2021a, 2021b, 2023), and, if vindicated, would replicate these findings.

If the explanation claim is true, then we should find an association between people judging that the nomically open world is most like our world and judging that the growing block world is most like our world, and between people judging that the nomically closed world is most like our world and judging that the block universe world is most like our world. This was H3.

In order to investigate the reason claim, we present participants with just one of the nomic vignettes. Those who see the nomically open vignette are told that Katie is in a world just like that and then asked whether she is more likely to be in the growing block or the block universe world. Those who see the nomically closed vignette are told that Katie is in a world just like that and then asked whether she is more likely to be in the growing block or the

doi: 10.48106/dial.v77.i1

block universe world. If the reason claim is true, then people should judge that if Katie is in a nomically open world, then she is more likely to be in a growing block world as opposed to a block universe world, and if Katie is in a nomically closed world, then they should judge that she is more likely to be in a block universe world as opposed to a growing block world. This was our H4.

Experiment 2 tests the deliberative reason claim. Here, participants are presented with a single vignette that describes an interaction between two characters (George and Helena). George reasons from the fact that our world is deliberatively open to the conclusion that it is nomically open and, from there, to the conclusion that future events do not exist. Helena rejects George's reasoning and explains where she thinks it goes awry. Participants are asked which character is correct. If the deliberative reason claim is true, then we should find that more people will judge that George is correct. This is H5. The final part of this experiment focuses on whether people can see the inferential connection between accepting or rejecting this reasoning. Participants are asked which world (growing block or block universe) the two characters will take *themselves* to be in. We predicted that participants would judge that Helena would take herself to be in a block universe world while George would take himself to be in a growing block world (H6).

We begin, in section 1, by outlining our methodology and results. Then, in section 2, we consider the upshot of those results for understanding our pre-reflective views of the world and the connection between them.

1 Methodology and Results

Experiment 1 Methodology

1.1 Participants

856 people participated in the study. Participants were recruited and tested online using Amazon Mechanical Turk and compensated $2 for their time. 732 participants had to be excluded from the analyses. That is because they failed to answer all the questions (n = 80), failed one of the attentional check questions (n = 73), or failed to answer two out of three comprehension questions correctly for the openness vignettes or three out of four comprehension questions correctly for both time vignettes (n = 579). The remaining sample was composed of 124 participants (46 female; aged 21–72, mean age 38.98

(SD = 9.95)). Ethics approval for these studies was obtained from the University of Sydney Human Research Ethics Committee. Informed consent was obtained from all participants prior to testing. The survey was conducted online using Qualtrics.[8]

1.2 Materials and Procedure

Participants first see *both* of the following openness vignettes. The first vignette describes a world in which the universe is Nomically Open—which we called Universe A. The second vignette describes a world in which the universe is Nomically Closed—which we called Universe B.

> Nomically Open (Universe A):
> Imagine a universe (Universe A) in which not everything that happens is completely caused by whatever happened before it. In Universe A, there are multiple different ways the future could go, given that the past and present are as they are. Given the past, every event *does not have to happen* the way that it does. So, if we "ran" Universe A over again from its very first moment, events might unfold differently to the way they did unfold.
>
> For example, one day, Katie decided she wanted to have a cup of coffee with her breakfast. Like everything else, this decision is not completely caused by whatever happened before it. So, if everything in the universe was exactly the same up until Katie made her decision, it *did not have to happen* that Katie would decide to have a cup of coffee.

> Nomically Closed (Universe B):
> Imagine a universe (Universe B) in which everything that happens is completely caused by whatever happened before it. In Universe B, there are not multiple different ways the future could go, given that the past and present are as they are. Given the past, every event *has to happen* the way that it does. So, if we "ran" Universe B over again from its very first moment, events would unfold exactly the same way that they did unfold.
>
> For example, one day, Katie decided she wanted to have a cup of coffee with her breakfast. Like everything else, this decision was

8 22% of the remaining sample got every comprehension question correct.

doi: 10.48106/dial.v77.i1

completely caused by whatever happened before it. So, if everything in this universe was exactly the same up until Katie made her decision, then it *had to happen* that Katie would decide to have a cup of coffee.

After reading both vignettes, participants responded to three comprehension questions to which they could either respond (a) true or (b) false.

1. If we "reran" Universe [A/B] over and over again, we would always get the very same events occurring in the very same order.
2. In Universe [A/B], the way things are now could not have been any different from how they are unless the past had been different from how it is.
3. In Universe [A/B], there is only one way the future can unfold, given that the past and present are the way they are.

Participants who did not correctly answer two out of three of these questions for each vignette were excluded from the analyses.

Participants are then asked, "Which universe do you think is most like our universe?" and given two options: (a) Universe A or (b) Universe B.

Participants then see both of the following time vignettes. The first vignette describes a universe that is a growing block world—which we called Universe C. The second vignette describes a block universe world—which we called Universe D.

Growing Block (Universe C):
Imagine a universe (Universe C) where new events—such as the extinction of the dinosaurs, the launching of a ship, or the cutting of a birthday cake—and objects—such as the birth of a baby or the creation of a new car—constantly come into existence. The events and objects that come into existence remain in existence, so the sum total of reality grows as new events and objects come to exist. In this universe, the events and objects that have just come into existence are those that are in the objective present. As new events and objects come into existence, already existing events and objects become part of the past. No future events or objects exist. So, there is a real, objective fact of the matter about which events are present and which are past.

For example, in Universe C, there is the event of Suzy throwing the ball at the window and the event of Billy throwing the ball at the window. When Suzy throws her ball, Billy is still holding his ball; he has yet to throw it. When the event of Suzy's ball hitting the window comes into existence, it is in the objective present, and the event of Billy's ball hitting the window does not yet exist. It is still in the future. When the event of Billy's ball hitting the window comes into existence, it is in the objective present, and the event of Suzy's ball hitting the window exists in the objective past. So, in this universe, first, Suzy throws the ball, and it hits the window; then, later, the event of Billy's ball hitting the window comes into existence, at which time Suzy's throwing the ball at the window still exists but is in the past.

Block Universe (Universe D):
Imagine a universe (Universe D) where a single set of events—such as the extinction of the dinosaurs, the launching of a ship, or the cutting of a birthday cake—and objects—such as the birth of a baby or the creation of a new car—exist. All these events are equally real. The sum total of reality never grows or shrinks, so the totality of events that exist never changes. In this world, past, present, and future events all exist. If dinosaurs have ever existed, then dinosaurs exist somewhere in the universe. If there will ever exist sentient robots, then there exist sentient robots somewhere in the universe. In Universe D, other *times* are much like other *places*. Just as in our world, Singapore, Sydney, and Seattle all exist, even though they do not exist in the same place; in Universe D, dinosaurs and robots exist, even though they do not exist at the same time. So, in Universe D, every time is present from the perspective of those located in it, just as every place is "here" from the perspective of those located in it.

For example, in Universe D, there is the event of Suzy throwing the ball at the window and the event of Billy throwing the ball at the window. When Suzy throws her ball, Billy is still holding his ball; he has yet to throw it. In Universe D, the event of Suzy throwing her ball and the event of Billy throwing his ball both exist. But they do not exist at the same place in space-time: the event of Suzy's ball hitting the window is earlier than the event of Billy's ball hitting the

doi: 10.48106/dial.v77.i1

> window. So, in Universe D, there is a fact of the matter about which ball hits the window first, namely, Suzy's, and so there is a fact of the matter about in which order the two events occur. But there is no fact about which event *really is* present and which is past or future. The event of Suzy's ball hitting the window is *past* relative to people who are located at the time that Billy's ball hits the window, while the event of Billy's ball hitting the window is *future* relative to people who are located at the time that Suzy's ball hits the window.

After reading both time vignettes, participants responded to four comprehension questions to which they could respond (a) true or (b) false.

1. In Universe [C/D], the past and present exist, but the future does not.
2. In Universe [C/D], the past, present, and future exist.
3. In Universe [C/D], there is an objective fact as to which events are present.
4. In Universe [C/D], events are always past or future relative to other events.

Participants who failed to correctly answer three out of four of these questions for each vignette were excluded from the analyses.

Participants are then asked, "Which universe do you think is most like our universe?" and are given two options: (a) Universe C or (b) Universe D.

Finally, participants then see either the nomically open or nomically closed vignette again, along with both time vignettes, and respond to the following question: "Katie is in a universe just like A/B. Do you think that Katie is more likely to be in Universe C or more likely to be in Universe D?" and are given two options: (a) Universe C or (b) Universe D.

1.3 Results

Before presenting the statistical analysis, we will start by summarising our main findings. We first hypothesised that (H1) more people would judge that the nomically open world is more like our world than the nomically closed world. This hypothesis was supported. Participants were more likely to judge that our world is more like a nomically open world compared to a nomically closed world. We then hypothesised that (H2) most people would judge that our world is a growing block world rather than a block universe world. This hypothesis was not supported.

Next, we hypothesised (H3) that there would be an association between people judging that the nomically open world is most like our world and judging that the growing block world is most like our world and between people judging that the nomically closed world is most like our world and judging that the block universe world is most like our world. This hypothesis was not supported. While there was a significant association between people's judgements about nomic openness and time, the association we found was not the one we hypothesised. Instead, there was an association between judging that our world is nomically closed and judging it to be a growing block world. Participants who judged our world to be nomically open were roughly divided in their likelihood to judge our world to be a growing block world or a block universe world.

Finally, we hypothesised (H4) that participants who are told that a character (Katie) is in a nomically open world would be more likely to judge that she is in a growing block world than a block universe world (and participants who are told that she is in a nomically closed world would be more likely to judge that she is in a block universe world than a growing block world). We found evidence for this.

Separate one-way chi-square tests were performed to test whether (a) most participants judged that the nomically open world was more like our world compared to the nomically closed world and whether (b) most participants judged that our world is a growing block world rather than a block universe world. The results of those tests showed that the first hypothesis was vindicated. This means that participants are more likely to judge the world as nomically open (76, 61.3%) as opposed to being nomically closed (48, 38.7%; $\chi^2(1, N = 124) = 6.323, p = .012$). Our hypothesis that participants will judge that our world is more like a growing block world (69, 55.9%) as opposed to a block universe world (55, 44.4%; $\chi^2(1, N = 124) = 1.582, p = .209$) was not statistically significant, indicating that participants are equally likely to judge our world as being either a growing block world or a block universe world.

Table 1 below summarises the descriptive data of participants' judgements regarding which nomic vignette (Nomically Open; Nomically Closed) is most like our world and which time vignette (Growing Block world; Block Universe world) is most like our world. To test whether there was an association between participants who judged our world to be nomically open and their judging of our world to be a growing block world, we performed a chi-square test of independence. This hypothesis was not supported. Instead, there

doi: 10.48106/dial.v77.i1

was an association between participants judging our world to be nomically closed and judging it to be a growing block world ($\chi^2(1, N = 124) = 5.449$, $p = .020$). Participants who judged our world to be nomically open were divided between judging it to be a growing block world and a block universe world.

Table 1: Participants' judgements of which nomic universe and time vignette are most like the actual world.

World	Growing Block World	Block Universe
Nomically Open	(36) 29.0%	(40) 32.3%
Nomically Closed	(33) 26.6%	(15) 12.1%

Finally, we performed a chi-square test of homogeneity to test whether participants who are told that Katie is in a nomically open world would be more likely to judge that she is in a growing block world (and whether people who are told that she is in a nomically closed world would be more likely to judge that she is in a block universe world). There was a significant association, $\chi^2(1, N = 124) = 6.613$, $p = .010$. Participants who were told that Katie was in a nomically open world were more likely to judge that she was also in a growing block world. Meanwhile, participants who were told that Katie was in a nomically closed world were more likely to judge that she was also in a block universe world (see table 2).

Table 2: Participants' judgements of which universe Katie is more likely to be in based on associations between nomic openness and time.

World	Growing Block World	Block Universe
Nomically Open	(38) 65.5%	(20) 34.5%
Nomically Closed	(28) 42.4%	(38) 57.6%

Experiment 2 Methodology

Participants

856 people participated in the study. Participants were recruited and tested online using Amazon Mechanical Turk and compensated $2 for their time.

732 participants had to be excluded from the analyses. That is because they failed to answer all the questions (n = 124), failed one of the attentional check questions (n = 54), or failed to answer three out of four comprehension questions correctly for the discussion vignette or failed to answer three out of four comprehension questions correctly for the time vignettes (n = 554). The remaining sample was composed of 124 participants (49 female, 2 trans/non-binary; aged 20–78, mean age 36.58 (SD = 99.716)). Ethics approval for these studies was obtained from the University of Sydney Human Research Ethics Committee. Informed consent was obtained from all participants prior to testing. The survey was conducted online using Qualtrics.[9]

1.2 Materials and Procedure

In this study, participants first see a single vignette—the nomic discussion vignette—in which Helena and George present different views about the connection between nomic openness and the existence of the future.

> Nomic Discussion:
> Helena and George are standing outside a philosophy room having a heated discussion about the reasons there are to think that the future either exists or does not exist. If the future **does not** exist, then future events, such as the existence of a colony on Mars or the robot uprising, do not exist, although perhaps one day they will. If the future **does** exist, then if there will be a colony on Mars in the future, it is true right now that the colony exists out there in the universe somewhere. If the future exists, then future events (and places) are much like other places here and now. While Helena and George are located in Singapore, it's still the case that Sydney and London exist; they just don't exist *in Singapore*. In the same way, if the future exists, then the colony on Mars exists; it just doesn't exist *here and now*.
>
> According to George, one reason to think that the future does not exist is that if the future did exist, then there are not multiple different ways the future could go, given that the past and present are as they are. If the future exists, then given the past and present, every future event *has to happen* the way that it does. So if the future exists, then if we re-ran the universe over again from its very first moment,

9 16% of the remaining sample got every comprehension question correct.

doi: 10.48106/dial.v77.i1

events would unfold exactly the same way. But then Helena cannot be free to *choose* what to eat for breakfast tomorrow, since whatever she eats for breakfast tomorrow, it *had* to be that she would eat that thing.

Helena tells George that he is mistaken. That kind of reasoning, she says, gives us no reason to think that the future does not exist. Just because the event of my (Helena's) eating cereal exists out there in the future, it doesn't mean that my eating cereal was determined by the past and present. It doesn't mean that the future could not have gone some other way. It could be that if we reran the universe over again, then I would instead eat toast instead of cereal for breakfast. The mere fact that the event of my eating cereal is out there in the universe doesn't tell us that that event *had* to be out there. You, George, are located here in this office. But the fact that you are located here doesn't tell us that if the past and present had been the same, you *had* to be located in this office. Perhaps you could have been somewhere different! So, the fact that the event of my eating cereal is out there in the universe does not mean that I *had* to eat cereal. It just means that, in fact, I do eat cereal.

Participants then answered four comprehension questions to which they could answer either (a) true or (b) false.

(a) If Helena is right, then if the future exists, it can still be true that there are multiple ways the future could go, given that the past and present are as they are.
(b) If George is right, then if the future exists, it can still be true that there are multiple ways the future could go, given that the past and present are as they are.
(c) According to Helena, if the event of her eating cereal tomorrow exists, then it could still be that the past and present did not determine that she would decide to eat cereal.
(d) According to George, if the event of her eating cereal tomorrow exists, then it must be that the past and present determined that she would decide to eat cereal.

Participants who failed to correctly answer three out of four of these questions were excluded from the analyses.

Participants are then asked, "Which of the two parties, Helena or George, do **you** think is right?" and are given two options: (a) George or (b) Helena.

Participants then see both the time vignettes and associated comprehension questions (see experiment 1). Participants who failed to correctly answer three out of four of these questions for each vignette were excluded from the analyses.

Finally, participants then saw the NOMIC DISCUSSION vignette again, along with both time vignettes. They were then presented with two questions:

(1) "Which universe do you think *Helena* will think is most like the universe she is in?"
(2) "Which universe do you think *George* will think is most like the universe he is in?"

For each question, they were given two options: (a) UNIVERSE C or (b) UNIVERSE D.

1.2.3 Results

As in experiment 1, we also tested H2 by asking participants which world they believed was most like our world (i.e., growing block world or block universe world) and predicted that most people would judge that our world is a growing block world rather than a block universe world. Again, H2 was not supported. People were divided between judging that our world is most like a growing block world and a block universe world.

We hypothesised that (H5) if the deliberative reasoning claim is right, then most people should judge that George, rather than Helena, is right in the NOMIC DISCUSSION vignette. This hypothesis was not supported. Instead, contrary to our prediction, we found that most participants judged that Helena, rather than George, was right.

Finally, we hypothesised that (H6) people will judge that Helena will take herself to be in a block universe world and that George will take himself to be in a growing block world. This hypothesis was supported.

Separate one-way chi-square tests were performed to test whether (a) most participants will judge that our world is more like a growing block world, (b) most participants will judge that George was right in the nomic openness discussion, (c) most participants will judge that Helena will take herself to be in a block universe world, and (d) most participants will judge that George will take himself to be in a growing block world. The results of those tests

doi: 10.48106/dial.v77.i1

showed that (a) participants were divided between judging that our world is more like a growing block world (64, 51.6%) and a block universe world (60, 48.4%; $\chi^2(1, N = 124) = .124, p = .129$), which does not support H2. Further, (b) contrary to H5, more participants judged that Helena (87, 70.2%), rather than George (37, 29.8%), was right in the nomic openness discussion, $\chi^2(1, N = 124) = 20.161, p < .001$. H6 was vindicated: most participants (c) judged that Helena would take herself to be in the block universe world (80, 64.5%; $\chi^2(1, N = 124) = 10.452, p < .001$), and that (d) George would take himself to be in the growing block world (80, 64.5%; $\chi^2(1, N = 124) = 10.452, p < .001$).

2 Discussion

There are several notable aspects of our results. First, as predicted, we found that a majority of people judged our world to be nomically open rather than closed. These results are of interest to those aiming to model our naïve representation of future openness. Taken in conjunction with previous work in this area, they begin to paint a picture of people's naïve representation of the future.

Hodroj et al. (2023) found that a majority of people (66%) judged our world to be one in which the future is *alethically* open rather than closed. Latham and Miller (2023) report that a majority of people (87%) judged our world to be deliberatively open rather than deliberatively closed: that is, they judged the future to be one in which what we do in the future is the product of our earlier deliberations, so that had we deliberated differently, we would have made different choices and subsequently done different things. These results, taken together with our current results, suggest that people's naïve representation of the future probably involves at least a combination of representing the future to be deliberatively, alethically, and nomically open. It also suggests that it may be deliberative openness that is most important when it comes to capturing people's naïve representation of the open future (something Torre 2011 gestures towards).

These results may also suggest that there are several naïve representations of future openness, all or almost all of which include representing the future as deliberatively open but only some of which include representing it as nomically and/or alethically open. Perhaps this is not surprising, given the evidence regarding people's naïve representation of time. Baron, Miller and Tallant (2022) cite a range of experiments that they take jointly to show that

there is no single, shared, naïve representation of time. What is true of time might also be true of naïve representations of the open future.

Our results also have implications for the narrow vindication claim. According to that claim, recall, the growing block theory vindicates our naïve representation of the future as open, and the block universe theory does not. There is some support for this claim, given the results of this study, alongside those of Hodroj et al. (2023) and Latham et al., despite the fact that these studies jointly suggest that *most* aspects of our naïve representation of future openness (and the most important of these) are consistent with our world being a block universe world.

The study by Latham et al. suggests that a vast majority of people have naïve representations of the future according to which the future is deliberatively open. But the presence of deliberative openness is clearly consistent with our world being either a block universe or a growing block world. So, arguably, the most powerful aspect of our naïve representation of the future is one that can be vindicated by either view of time.

The current study found that a majority of people represent the future as nomically open, not closed. But, again, the future being nomically open is consistent with our world being either a block universe or a growing block. So, either view can vindicate this aspect of our naïve representation.

The only good news for the growing block theorist lies in the Hodroj et al. (2023) study, which found that a majority of people represent the future as alethically open. On standard (i.e., nonbranching) versions of the block universe, the future is not alethically open, while on standard versions of the growing block theory, it is. So, the growing block theory does vindicate *this* aspect of openness, while the block universe view does not.

Still, it's worth bearing in mind that according to the study by Hodroj et al. (2023), ~34% of people did not judge the future to be alethically open. So, it may be that a substantial minority of people have a naïve representation of the future that is equally vindicated by both the growing block and block universe theories. And, of course, even if the narrow vindication claim is true, it remains open to dispute whether it gives us much, if any, reason to prefer the growing block view to the block universe view. Still, these studies suggest that, insofar as growing block theorists want to try and argue for their view via something like the (narrow) vindication claim, they might do well to focus more on alethic openness than other forms of openness.

Moving on, we did not find that a majority of people represent our world as a growing block rather than a block universe. Instead, across both experiments,

doi: 10.48106/dial.v77.i1

people were evenly split between the two models. This should, perhaps, not be such a surprise. Latham, Miller and Norton (2021a) found that across two experiments, ~70% of people judged our world to be dynamical (either growing block, moving spotlight, or presentist), and of those, between ~35% and ~50% judged it to be a growing block. Even though in these studies only ~25% and ~35% of all people judged our world to be most like a growing block world, we expected that, given a forced choice between a growing block and a block universe world, most people would judge it to be *more like* a growing block world than a block universe world, given that most people judge our world to be temporally dynamical.

Our results suggest that although people are drawn to dynamical theories of time, their naïve representation of time might be less *strongly* dynamical than has otherwise been thought. This might explain why, given that the block universe and growing block views are very similar in a number of ways, when given a forced choice between the two, people tended to be roughly evenly divided in which world they thought was most like ours.

This brings us to the explanation and reason claims. Our results here are both startling and puzzling. Consider, first, the explanation claim. Our hypothesis here (H3) was not vindicated. While we did find an association, it was the opposite of the one we predicted. We found an association between judging a world to be nomically *closed* and judging it to be a growing block world. Amongst people who judged our world to be nomically open, people were evenly split between judging it to be a growing block or a block universe. While the latter absence of an association is not such a surprise (given that, *in fact*, nomically open words are no more likely to be growing block worlds as opposed to block universe worlds, it is perhaps heartening to see people's judgements in this regard), the presence of the converse association is very puzzling. It's hard to see why people who judge the future to be nomically *closed* would tend to judge it to be a growing block. The best we can come up with is that perhaps some people think that the laws of nature "push" the world along and cause it to grow, and they imagine this growth process must be deterministic (else the world would not know what to grow into). If this is the reason why (some) people judge our world to be nomically closed, then we would expect those people to judge that our world is a growing block. All we can really say is that further investigation of the association here would be useful.

Certainly, though, the lack of any association between people judging our world to be nomically open and judging it to be a growing block world suggests

that it is unlikely that the fact that people naïvely represent the future as nomically open is what even partially explains why they represent it to be a growing block. This finding is interesting, given our results regarding the reason claim. Our hypothesis in this regard was vindicated: participants judged that Katie was more likely to be in a growing block world than a block universe world if she was in a nomically open world and to be in a block universe rather than a growing block world if she was in a nomically closed world. Thus, people do seem to think that the fact that a world is nomically open is a reason to think it is a growing block world rather than a block universe world. The reason claim seems to be vindicated.

The vindication of the reason claim does suggest that there is some *sense* in which the growing block view of time better accords with our naïve representation of the future as nomically open. It accords in at least this sense: if the only thing someone knows about a world is that it is nomically open, they will think it more likely that the world is a growing block rather than a block universe world. So, there is some important connection between people's naïve representation of the future and their naïve representation of time. The former, we might say, *predisposes* them to thinking that our world is a growing block world, since if all they know about our world is that it is nomically open, people will tend to judge that it is a growing block world.

But of course, this is not all that people know about our world, and presumably, this explains why we found no association between people judging that our world is nomically open and that it's a growing block world. One thought about what might be going on here is that contemporary scientific knowledge is pushing people who judge that our world is nomically open to judge that it is a block universe world rather than a growing block world. If so, that could tend to eliminate the predicted association. But, first, we know from previous research by Latham, Miller and Norton (2021a) that levels of education and levels of scientific knowledge, especially in physics, have no effect on people's judgements about which view of time they think is true of our world. Second, in this study, we found that ~50% of people judged our world to be a growing block. So, it seems unlikely that this explains why we found no association.

Another possibility is that the reason at least some people judge our world to be nomically open is that they are aware of quantum mechanics rather than on the basis of their judgement on their naïve representation of the future. If so, it may be that those who *naïvely* represent the future as nomically open *are* more inclined to represent it as a growing block, but that many of those who represent the future as nomically open are employing a scientifically

doi: 10.48106/dial.v77.i1

informed representation of the future, and, perhaps, those people also tend to represent the world as a block universe. If so, that could eliminate the association. It would be useful to do follow-up work here that attempts to determine to what extent people's representation of the future as nomically open is naïve, as opposed to scientifically informed.

What we can say, though, is that at best, people are predisposed to represent our world to be a growing block in virtue of representing it to be nomically open, but that as a matter of fact, what explains why people represent the world to be a growing block is not that they represent it to be nomically open. This is further suggested by the results of our second experiment, in which only ~30% of people judged that George's reasoning was correct. Most people, then, do not endorse the deliberative reasoning claim we investigated.

In all, then, we think there is little evidence for the idea that part of what explains why people naïvely represent our world as a growing block is that they naïvely represent the future as nomically open. This will be of interest to A- and B-theorists alike. B-theorists have recently resisted what has become known as the argument from temporal phenomenology (Baron et al. 2015)—according to which we have reason to think our world is temporally dynamical because this is how it seems to us to be in perceptual experience—by denying that it does seem this way to us in experience (Hoerl 2014; Prosser 2016; Deng 2013, 2018; Bardon 2013; Miller, Holcombe and Latham 2020; Miller 2019, 2023; Latham, Miller and Norton 2020). Such views have often been deemed deflationist.

We know, however, that people naïvely represent our world as temporally dynamical (Latham, Miller and Norton 2021a, 2021b, 2023). If, as deflationists suppose, it does not seem to us, in experience, as though time is dynamical (and there is some suggestion from Latham, Miller and Norton 2020 that this might be right), then the question arises as to why we naïvely represent it that way. Deflationists, it seems, owe us some kind of explanation here.

One possibility, alluded to by Prosser (2016), is that part of what explains why we represent time as dynamical is that we represent the future as open. This study had the potential to show that part of what explains why we represent time as dynamical (by representing it as a growing block) is that we represent it as nomically open. Unfortunately for deflationists, we found no evidence of this.

Having said that, Prosser's suggestion is rather different from the one we investigated here. He hypothesises that because people represent the future as being objectively open (as opposed to merely perspectivally or subjectively

open), and because we represent that this openness moves (as what was once open becomes closed and part of the past), we must represent that there is a privileged and moving moment in time that is the border between the closed past and the open future. Further work, taking up the specific details of Prosser's view, would be welcome, given that we found no evidence in favour of the hypotheses we tested in this regard.

In all, we think that there is much more that can be learned about both our naïve representation of the open future and the ways in which this representation connects to our naïve representation of time. That work can shed light on the best way to model future openness (insofar as that modelling is attempting to capture some naïve representation of the future) and on whether what explains our open future practices also vindicates our naïve representation of the open future. It can, we hope, also shed light on the connection between our naïve representation of the future and of time and, hence, on extant debates in the philosophy of time.*

Batoul Hodroj
0000-0002-5227-392X
University of Queensland
batoul.hodroj@sydney.edu.au

Andrew James Latham
0000-0001-8639-8053
Aarhus University
andrew.latham@cas.au.dk

Kristie Miller
0000-0002-5092-8419
University of Sydney
kristie.miller@sydney.edu.au

References

Bardon, Adrian. 2013. *A Brief History of the Philosophy of Time*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199976454.001.0001.

Baron, Sam, Cusbert, John, Farr, Matt, Kon, Maria and Miller, Kristie. 2015. "Temporal Experience, Temporal Passage and the Cognitive Sciences." *Philosophy Compass* 10(8): 560–571, doi:10.1111/phc3.12244.

* THANKS

Baron, Sam, Miller, Kristie and Tallant, Jonathan. 2022. *Out of Time: A Philosophical Study of Timelessness*. Oxford: Oxford University Press, doi:10.1093/oso/9780192864888.001.0001.

Belnap, Nuel D., Jr. 1992. "Branching Space-Time." *Synthese* 92(3): 385–434, doi:10.1007/bf00414289.

—. 2005. "Agents and Agency in Branching Space-Times." in *Logic, Thought and Action*, edited by Daniel Vanderveken, pp. 291–313. Logic, Epistemology, and the Unity of Science n. 2. Dordrecht: Springer, doi:10.1007/1-4020-3167-X_14.

Braddon-Mitchell, David. 2004. "How Do We Know it is Now Now?" *Analysis* 64(3): 199–203, doi:10.1111/j.0003-2638.2004.00485.x.

Briggs, Rachael A. and Forbes, Graeme A. 2012. "The Real Truth About the Unreal Future." in *Oxford Studies in Metaphysics*, volume VII, edited by Karen Bennett and Dean W. Zimmerman, pp. 257–304. New York: Oxford University Press, doi:10.1093/acprof:oso/9780199659081.003.0009.

Broad, Charlie Dunbar. 1923. *Scientific Thought*. International Library of Psychology, Philosophy and Scientific Method. London: Kegan Paul, Trench, Trübner & Co.

—. 1938. *Examination of McTaggart's Philosophy. Volume II, Part I*. Cambridge: Cambridge University Press.

Callender, Craig. 2017. *What Makes Time Special?* Oxford: Oxford University Press, doi:10.1093/oso/9780198797302.001.0001.

Correia, Fabrice and Rosenkranz, Sven. 2018. *Nothing to Come: A Defence of the Growing Block Theory of Time*. Synthese Library n. 395. Dordrecht: Springer, doi:10.1007/978-3-319-78704-6.

Deng, Natalja. 2013. "On Explaining Why Time Seems to Pass." *The Southern Journal of Philosophy* 51(3): 367–382, doi:10.1111/sjp.12033.

—. 2018. "On 'Experiencing Time': A Response to Simon Prosser [on Prosser (2016)]." *Inquiry* 61(3): 281–301, doi:10.1080/0020174x.2017.1322674.

Forbes, Graeme A. 2016. "The Growing Block's Past Problems." *Philosophical Studies* 173(3): 699–709, doi:10.1007/s11098-015-0514-1.

Forrest, Peter. 2004. "The Real but Dead Past: A Reply to Braddon-Mitchell (2004)." *Analysis* 64(4): 358–362, doi:10.1111/j.0003-2638.2004.00510.x.

Grandjean, Vincent. 2021. "How Is the Asymmetry Between the Open Future and the Fixed Past to Be Characterized?" *Synthese* 198(3): 1863–1886, doi:10.1007/s11229-019-02164-2.

—. 2022. *The Asymmetric Nature of Time: Accounting for the Open Future and the Fixed Past*. Synthese Library n. 468. Dordrecht: Springer, doi:10.1007/978-3-031-09763-8.

Hodroj, Batoul, Latham, Andrew James, Lee-Tory, Jordan and Miller, Kristie. 2023. "Alethic Opennes and the Growing Block Theory of Time." *The Philosophical Quarterly* 73(2): 532–556, doi:10.1093/pq/pqac062.

Hoerl, Christoph. 2014. "Do We (Seem To) Perceive Passage?" *Philosophical Explorations* 17(2): 188–202, doi:10.1080/13869795.2013.852615.

ISMAEL, Jenann. 2012. "Decision and the Open Future." in *The Future of the Philosophy of Time*, edited by Adrian BARDON, pp. 149–168. Routledge Studies in Metaphysics n. 4. London: Routledge, doi:10.4324/9780203338315.

LATHAM, Andrew James and MILLER, Kristie. 2023. "Why Do People Represent Time as Dynamical? An Investigation of Temporal Dynamism and the Open Future." *Philosophical Studies* 180(5): 1717–1742, doi:10.1007/s11098-023-01940-8.

LATHAM, Andrew James, MILLER, Kristie and NORTON, James. 2020. "An Empirical Investigation of Purported Passage Phenomenology." *The Journal of Philosophy* 117(7): 353–386, doi:10.5840/jphil2020117722.

—. 2021a. "Is Our Naïve Theory of Time Dynamical?" *Synthese* 198(5): 4251–4271, doi:10.1007/s11229-019-02340-4.

—. 2021b. "An Empirical Investigation of the Role of Direction in Our Concept of Time." *Acta Analytica* 36(1): 25–47, doi:10.1007/s12136-020-00435-z.

—. 2023. "Do the Folk Represent Time as Essentially Dynamical?" *Inquiry* 66(10): 1882–1913, doi:10.1080/0020174X.2020.1827027.

LEWIS, David. 1979. "Counterfactual Dependence and Time's Arrow." *Noûs* 13(4): 455–476. Reprinted, with a postscript (Lewis 1986b), in Lewis (1986a, 32–51), doi:10.2307/2215339.

—. 1986a. *Philosophical Papers, Volume 2*. Oxford: Oxford University Press, doi:10.1093/0195036468.001.0001.

—. 1986b. "Postscript to Lewis (1979)." in *Philosophical Papers, Volume 2*, pp. 52–66. Oxford: Oxford University Press, doi:10.1093/0195036468.003.0002.

MACFARLANE, John. 2003. "Future Contingents and Relative Truth." *The Philosophical Quarterly* 53(212): 321–336, doi:10.1111/1467-9213.00315.

—. 2008. "Truth in the Garden of Forking Paths." in *Relative Truth*, edited by Manuel GARCÍA-CARPINTERO and Max KÖLBEL, pp. 81–102. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199234950.003.0004.

MARKOSIAN, Ned. 1995. "The Open Past." *Philosophical Studies* 79(1): 95–105, doi:10.1007/bf00989786.

MCCALL, Storrs. 1994. *A Model of the Universe: Space-Time, Probability, and Decision*. Clarendon Library of Logic and Philosophy. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780198236221.001.0001.

MILLER, Kristie. 2019. "Does It Really Seem to Us as Though Time Passes?" in *The Illusions of Time: Philosophical and Psychological Essays on Timing and Time Perception*, edited by Valtteri ARSTILA, Adrian BARDON, Sean Enda POWER, and Argiro VATAKIS, pp. 17–34. London: Palgrave Macmillan, doi:10.1007/978-3-030-22048-8_2.

—. 2023. "Against Passage Illusionism." *Ergo* 9(45): 1233–1263, doi:10.3998/ergo.2914.

MILLER, Kristie, HOLCOMBE, Alex and LATHAM, Andrew James. 2020. "Temporal Phenomenology: Phenomenological Illusion Versus Cognitive Error." *Synthese* 197(2): 751–771, doi:10.1007/s11229-018-1730-y.

Prosser, Simon. 2016. *Experiencing Time*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780198748946.001.0001.

Tooley, Michael. 1997. *Time, Tense, and Causation*. Oxford: Oxford University Press, doi:10.1093/0198250746.001.0001.

Torre, Stephan. 2011. "The Open Future." *Philosophy Compass* 6(5): 360–373, doi:10.1111/j.1747-9991.2011.00395.x.

Williams, J. Robert G. 2008. "Aristotelian Indeterminacy and the Open Future." Unpublished manuscript, https://philpapers.org/rec/WILAIA-4.

Published by *Philosophie.ch*

Verein philosophie.ch
Fabrikgässli 1
2502 Biel/Bienne
Switzerland
dialectica@philosophie.ch

https://dialectica.philosophie.ch/

ISSN 0012-2017

ISBN 1234-5678

Dialectica is supported by the Swiss Academy of Humanities and Social Sciences.

Abstracting and Indexing Services

The journal is indexed by the Arts and Humanities Citation Index, Current Contents, Current Mathematical Publications, Dietrich's Index Philosophicus, IBZ — Internationale Bibliographie der Geistes- und Sozialwissenschaftlichen Zeitschriftenliteratur, Internationale Bibliographie der Rezensionen Geistes- und Sozialwissenschaftlicher Literatur, Linguistics and Language Behavior Abstracts, Mathematical Reviews, MathSciNet, Periodicals Contents Index, Philosopher's Index, Repertoire Bibliographique de la Philosophie, Russian Academy of Sciences Bibliographies.

Contents