Meteorite

The Student Journal of Philosophy
from the University of Michigan

Spring 2013

Meteorite gratefully acknowledges the support of Molly Mahony, Judith Beck, the staff of the Tanner Library, and the faculty and staff of the University of Michigan Department of Philosophy, without whom this publication would not be possible.

Meteorite c/o
Department of Philosophy
2215 Angell Hall
435 S. State St.
Ann Arbor, MI 48109
USA

Meteorite is typeset in Palatino Linotype.

ISSN: 1099-8764

CONTENTS

Meaning and Triviality: Brogaard and Smith on the Meaning of Life

Michael Garatoni
Trinity University

IN THEIR RECENT ARTICLE "On Luck, Responsibility and the Meaning of Life," Berit Brogaard and Barry Smith classify existing accounts of the meaning of life as 'internalist' or 'externalist,' and claim that all purely internal and all purely external types of analyses face fatal objections (Brogaard and Smith 2005, 446). In lieu of the seemingly defective purist accounts, Brogaard and Smith propose their own "mixed account" of the meaning of life, a prescription that combines internalist and externalist features and which includes the authors' "standard of non-triviality" (Brogaard and Smith 2005, 446). As the element most responsible for constraining meaning under the authors' account, the non-triviality standard demands that the individual deliberately pursue "difficult and ambitious" goals involving activities that enjoy "public measures of success" (Brogaard and Smith 2005, 446). In the first section of this paper, I consider the strength and essential character of Brogaard and Smith's objection to the pure internalist account and detail the authors' own account of meaning. I then conclude the section by briefly outlining my objections to those criteria that together constitute the authors' 'standard of non-triviality.' In the remaining sections of the paper I present various objections to each criterion of that standard of non-triviality. I hope to show that the pieces of Brogaard and Smith's 'non-triviality' criterion (hereafter referred to as "SNT") constitute a problematic and largely implausible precondition for one's life to have meaning.

A Dubious Problem and the Problematic Solution

The pure internalist account of meaning to which Brogaard and Smith refer makes the meaningfulness of one's life contingent purely on one's mental states, while the externalist account makes meaningfulness dependent exclusively on factors that are external to one's mental states. The authors note that the pure internalist account would allow meaning to arise through something like Robert Nozick's experience machine.[1] If the meaning of one's life depends strictly on the mental state of the individual, then presumably the individual could attain meaning simply by entering the experience machine or using drugs to create the desired psychological state, but this cannot confer meaning on a life according to Brogaard and Smith (Brogaard and Smith 2005, 445). Alternatively, the pure externalist account faces the threat of the Forest Gump figure who accidentally produces numerous external effects on the world but never accomplishes the effect deliberately, and also the Sisyphean figure, who labors against his will.

According to the authors, "a meaningful life is a life upon which a pattern has been imposed that relates not merely to what goes on inside the person's head, but which also involves, in serious ways, the person having an effect upon the world." (Brogaard and Smith 2005, 445). Against the pure internalist account of meaning, the authors prefer to countenance certain lives (such as that of Mozart, who was a depressive) which lack the intrinsically meaningful mental

[1] Following Robert Nozick (1981), who devised the "experience machine" thought experiment, Brogaard and Smith contend that meaning cannot accrue apart from external effects. Nozick, who gives an account of meaning that involves transcending limits, argues that "for a life to have meaning, it must connect with other things, with some things or values beyond itself," and thus the experience machine cannot give life meaning because, "the experience machine, though it may give you the experience of transcending limits, encloses you within the circle of just your own experiences." (*Philosophical Explanations,* 545).

states that could possibly make life meaningful according to internalism. They also seem interested in regulating the correspondence between the individual's mental states and the reality of his situation, particularly (though not exclusively) insofar as improper correspondence comes as a result of delusional, biased, or otherwise corrupted mental states (as in the case of the denizens of Nozick's machine). Brogaard and Smith thus spend the first half of their article forging a link in their theoretical framework that would properly ground the individual's mental states in objective reality. For Brogaard and Smith, in order to live a meaningful life:

You will need to decide how to shape your life and the world in which you live and set goals accordingly. These goals must be effective in giving rise to corresponding actions on your part, and they must culminate in a shape or pattern that is non-trivial… (Brogaard and Smith 2005, 446).

In response to the pill and Experience Machine counterexamples observed by the authors, many internalists note that very few people will claim meaning in such trivial and disconnected pursuits (Metz 2007). Perhaps the apparent repugnancy of this counterexample results directly from this fact that so few people know of anyone who would find meaning in wholly illusory accomplishments i.e. from life inside an experience machine, or otherwise from effects produced in some illusory fashion. However, the experience machine may only present a problem to those who prefer actuality to appearance, who premise the acquisition of meaning upon the realization of certain ends that pertain to the actual world. Certainly for most people, including those who would acknowledge the force of the examples involving the pill or the experience machine, life would seem meaningless if all the accomplishments therein became known as mere illusions instead of actualities, but, as with many of the intuitions that inform philosophical debates, the widespread disturbance

caused by the aforementioned thought experiments may reflect prejudice rather than some real problem. Still, even if one acknowledges the force of the pill/experience-machine counterexample, the particular remedy offered by Brogaard and Smith, specifically the SNT (Brogaard and Smith 2005, 446), fails to prevent the recurrence of experience-machine-type objections—because 'public measures of success' depend on the public's mental states, which are no less corruptible than the mental states of the individual.

Moreover, as I will argue later, the author's standard of 'non-triviality' introduces additional problems. The authors write:

> To capture the factor of non-triviality in a more substantial fashion, we need further to impose the criterion that the effort in question must be directed and calibrated in the relation to some independent standards of success and failure, standards which are 'objective' in the sense that they could be applied by some disinterested observer. A meaningful life is a life which consists in your making and realizing what are for you in your particular setting ambitious and difficult plans; but these must at the same time be plans in relation to which there exist genuine, public measures of success and therefore also the risk of failure. (Brogaard and Smith 2005, 446)

As this passage suggests, the authors' standard of 'non-triviality' consists of two criteria. On their view, something is not trivial if and only if:

(a) it is a plan of action that has public measures of success.
(b) it is a plan of action that is 'ambitious and difficult.'

Brogaard and Smith's particular construction of the non-triviality standard appears carefully designed to cover perceived

weaknesses in both the pure internalist and pure externalist accounts of meaning. For example, the PMS component creates "independent standards of success," (Brogaard and Smith 2005, 446) that effectively disempower the subject's mental states as measures of meaningfulness and which purportedly create the risk of failure that the authors wish to impose on the pursuit of meaning. However, Brogaard and Smith's "factor of non-triviality" (Brogaard and Smith 2005, 446) and its constituent parts, which are not wholly unprecedented in the literature on the meaning of life, effect much more than the elimination of certain defects in the aforementioned pure accounts of meaning, and thus the authors likely retain some other unidentified motivation for formulating and imposing this particular standard. While Brogaard and Smith do not proffer an explicit rationale for the SNT in their account, they might appeal to the work of their predecessors, which variously defends some version of the non-triviality criterion. For example, Cottingham, in his account of the meaning of life, rationalizes the need for a non-triviality standard by claiming that "to call an activity or life meaningful implies a certain profundity or seriousness," and thus that "to appraise something as meaningful excludes it being trivial or silly" (Cottingham 2003, 21). Of course, Cottingham defends a much different view than the one presented by Brogaard and Smith. However, with regards to the possibility that Brogaard and Smith might appeal to the aforementioned rationale (for the purposes of their own account), I doubt that the apparent "profundity or seriousness" of a given activity precludes the possibility that the activity may also seem somewhat "trivial or silly," especially when the attributions come from different perspectives. What the individual recognizes as "profound or serious" may seem utterly absurd from the perspective of someone else.

In any case, Brogaard and Smith's piecemeal construction of the 'non-triviality' criterion displays several weaknesses that render it

implausible. First, the institution of public measures may actually handcuff individuals who accept the author's prescription of 'ambitious and difficult plans.' Further, the PMS criterion, which supports a number of absurd conclusions, does not necessarily introduce, as the authors claim, a substantially greater 'risk of failure,' nor does it adequately fulfill any other identifiable purpose that would vindicate it. In what follows, I examine the components of Brogaard and Smith's "factor of non-triviality" (Brogaard and Smith 2005, 446) and detail those problems that render the standard implausible.

"Difficult and Ambitious Plans" Versus "Public Measures of Success"

While the 'difficult and ambitious plans' requirement (hereafter referred to as DAP) directly bars meaning from unchallenging activities (e.g. grass-blade counting[2]), it also effectively handles the aforementioned counterexamples - neither drug-induced illusory mental states nor slavish rock-rolling, nor the dumb luck of Forest Gump would represent genuine accomplishments for the individuals in question because these effects do not occur as the result of "ambitious and difficult" plans. However, in addition to the DAP criterion, Brogaard and Smith impose a standard which requires the use of public measures of success in order to establish a risk of failure (Brogaard and Smith 2005, 446). Evidently, Brogaard and Smith so distrust the individual's ability to render an accurate assessment of his success (or lack thereof) that they require an appeal to the judgment of the public. As previously stated, their view eliminates the possibility of meaning from easy, 'trivial' accomplishments. But what if none of the activities that have public measures of success present a challenge for the individual in question? Brogaard and Smith do not es-

[2] Example borrowed from: John Rawls, *A Theory of Justice* (United States of America: President and Fellows of Harvard College, 1971).

tablish any necessary concurrence between what counts as 'difficult or ambitious' activities and those sorts of activities that happen to enjoy public measures of success, and thus nothing prevents the two standards from eventually coming into conflict. Of course modern humans are highly social animals living in a highly connected and social world, and one may find it difficult to imagine that any one of these individuals might pursue meaning in an activity that does not somehow reflect the individual's relationship to society and thereby carry some 'public measure of success.' However, the current state of affairs can only reflect some degree of coincidental concurrence between challenging activities and activities with public measures of success and thus there remains some possibility that certain individuals will not find anything challenging in the endeavors that enjoy public measures of success. For example, suppose someone invents a new sport and only this sport presents a challenge for him. The standard imposed by Brogaard and Smith in order to establish a mere risk of failure could in fact preclude certain individuals from ever attaining meaning.

Of course, the authors themselves very much desire to impose restrictions on the possibility of meaning, so they might argue that the de facto preclusion of a few individuals only distresses theorists who would prefer to view the pursuit of meaning as an equal opportunity enterprise.[3] However, while the authors' independent reasons for separately imposing these two conditions certainly suggests that they intend to delegitimize things like mental states and grassblade counting as potential avenues toward meaning, nothing in the rationale for either of these standards suggests that the authors intend to punish those exceptional individuals who could and would satisfy either one of the criteria if not for the interdiction of the other. Nevertheless, the exceptional man of integrity- who satisfies the DAP criterion by refusing to seek meaning in a practically infallible (and therefore unchallenging and riskless) pursuit (e.g. grass counting) - necessarily fails, due, most ironically, to the fact that his successes actually transcend all public measures. The confluence of these two criteria not only excludes agents that the authors never intended to capture, but it excludes them specifically for adopting the standard endorsed by one (or the other) of the criteria.

[3] I owe credit to Steve Luper for this identifying this possible counterargument.

"Public Measures of Success"

Perhaps the most unsettling element of Brogaard and Smith's account comes with their suggestion that "public measures of success" need to involve some "widely disseminated culture of honest admiration" (Brogaard and Smith 2005, 446). This requirement entails that any activities that the community either condemns or simply does not appreciate cannot support meaning. Thus, "activities which have to be practiced in the dark, in secret (petty crime, for instance) are," according to Brogaard and Smith, "lacking such public measures of success" (Brogaard and Smith 2005, 447). On the basis of this criterion Brogaard and Smith strip all meaning from activities such as "genocide and gratuitous torture," - activities they have already found counterintuitive as sources of meaning (Brogaard and Smith 2005, 447). According to the authors, "these activities will also be found to be such that they need at least to some extent to be practiced in the dark," and thus they too fail to create meaning.

Do Brogaard and Smith really mean to suggest that public disapproval necessarily forces the individual out of the public eye? Personally, I doubt that crimes against humanity are necessarily meaningless, but I also deny that any undertaking of this sort needs necessarily to occur 'in the dark, in secret,' as Brogaard and Smith claim for the sake of saving their criterion from the counterexample of

genocide and torture. If, as Brogaard and Smith suggest, "genocide and gratuitous torture cause some problems for this criterion," (Brogaard and Smith 2005, 447) then so must many acts of terror that actually demand very little deception and which *intentionally* occur in the "public light of day" (Brogaard and Smith 2005, 446)! It seems very unlikely that the authors would grant the possibility of meaning to homicidal terrorists since they have already denied meaning to the homicidal dictator.

The authors might reply that terrorism too needs, 'at least to some extent to be practiced in the dark.' However, what factors then do Brogaard and Smith think necessarily compel a criminal to practice an activity 'in the dark, in secret?' I find nothing that distinguishes their criterion from a policy that would strip meaning from any activity that faces significant physical opposition in a public setting, and yet no human activity enjoys constant immunity from such opposition. In each of the examples supplied by the authors, the activity in question faces public disapprobation and penal codes that render the activity impracticable in the public sphere. However, if Brogaard and Smith think that a certain degree of public hostility or effective impracticability forces a person to act 'in the dark, in secret,' then human history offers numerous counterexamples to that criterion which predicates meaning on the practicability of the activity in public. Socrates and Jesus Christ, for example, both fell victim to a community that became so increasingly hostile to their subversive activities that any continuation of those activities became at least as impracticable in public as petty theft (we have the death sentences to prove it). Still, Socrates and Jesus continued to carry out their activities to their deaths, and today almost anyone (including, I suspect, Brogaard and Smith) who admits of the possibility of any person leading a meaningful life would also admit that Jesus and Socrates led non-trivial and extraordinarily meaningful lives. Even if one accepts the authors' intuition regarding 'genocide and gratuitous torture,' the criterion applied by Brogaard and Smith to filter out these cases of extreme violence seems both too broad to capture all of the unacceptable cases and also too narrow to permit meaning to accrue in some of the most benign and admirable cases. Because it lacks a clear limiting principle, the criterion threatens to remove the possibility of meaning from merely taboo practices or even from some of the most admirable activities.

As I intimated earlier, the text strongly implies that public measures of success can only exist where one finds a "widely disseminated culture of honest admiration" (Brogaard and Smith 2005, 446). Immediately following their discussion of PMS Brogaard and Smith describe daydreaming as a meaningless activity based on the presumption that daydreaming has "no standards of better and worse and no widely disseminated culture of honest admiration" (Brogaard and Smith 2005, 446). Additionally, the authors' first introduction of the 'public-light-of-day' criterion appears to follow directly from their notion of PMS which further suggests that public measures of success, as the authors understand them, should place a particularly restrictive limitation on the range of activities that retain the potential to confer meaning on a life. Meanwhile the example activities supplied by the authors- raising children, playing chess, participating in athletics- do not contradict but rather concur with the interpretation I have so far adopted. Obviously, this sort of requirement banishes any and all activities that the public does not care about, and thus so many important historical figures and some of the most consequential activities all lose their capacity for meaning.

Nevertheless, Brogaard and Smith do not explicitly derive the 'public-light-of-day' criterion from their broader PMS requirement so perhaps the authors use the phrase "public measures" independently, to refer to a more accommodating criterion that could ef-

fectively evaluate not only those activities with a 'widely disseminated culture of honest admiration,' but also some more obscure or less widely appreciated activities. Although the text strongly suggests that only a relatively small number of possible activities actually have accompanying public measures of the success, the authors still might have in mind some 'form' of success with more general applicability. However, even if the authors intend to affirm a significantly more accommodating definition of PMS, one must then begin to wonder if the criterion itself becomes rather trivial.[4] In this scenario, the authors still nominally achieve their stated goal by calibrating the individual's activities to something external to the individual, but the standard itself becomes so overly broad that it fails to regulate the acquisition of meaning in any significant way. Meanwhile the 'public' becomes largely irrelevant, a vestigial factor in the equation that determines meaning.

Ironically, the stricter version of the PMS may preclude many activities which obtain their license from extreme nonconformist philosophies, despite the fact that attempting to transcend cultural standards has a long and storied history as an ideal form of human behavior. Have those who have practiced individualism and freethinking always played dangerously at the edge of meaninglessness? What about those who spurn public measures of success to stake their entire existence on the possibility that they might actually alter those public measures of success? Brogaard and Smith's theory instructs these rebels to submit to the status quo or else lead a meaningless life. Strangely, those who most help to create the standards by which all others will receive their measure may very well lead a 'meaningless' life simply because their exceptional actions lack "relevant public measures of success." How exactly does a meaningless

[4] I owe this point to Steve Luper.

achievement in a meaningless individual life become a standard for meaningful achievement?[5] This paradox alludes to a deeper problem with the appeal to 'public measures': Brogaard and Smith select external object- public measures of success- to ground individual subjectivity despite the fact the object in question depends on subjectivity.

"Whether a person leads a meaningful life depends in every case not on that person's, or other people's, beliefs or feelings," say Brogaard and Smith, "but on what the person did as a consequence of his or her own decisions, as evaluated (actually or potentially) against the relevant public measures of success" (Brogaard and Smith 2005, 447). Of course, the authors may plausibly describe "relevant public measures of success," as something like an external object, but the measures themselves depend directly on the 'beliefs or feelings,' or even the attitudes and interests of the public. Instead of relying on those same subjective 'beliefs and feelings,' that suffice as the warrant and impetus for the individual's action, Brogaard and Smith prefer to supplant individual prerogative with the intersubjectivity of the masses for the purposes of determining not only the success or failure of the individual's activity but the entire domain of potentially meaningful activities. Unfortunately, the authors appeal to intersubjectivity does not actually dissolve the difficulties that Brogaard and Smith associate with individual subjectivity.

[5] Note that the way meaning is structured in this scenario is not consistent with the basic structure of Nozicks' account (which Brogaard and Smith cite). According Nozick's account, a life becomes meaningful by connecting with "things or values beyond itself," which already have meaning by virtue of connecting with some other "things or values beyond itself," and the regress only ends with the infinite God who is the giver of all meaning. In contrast, Brogaard and Smith's account allows finite and meaningless agents to contribute independently to developing standards of meaning- the individual can obtain meaning by 'connecting' with meaningless standards.

Recall that Brogaard and Smith object to the pure internalist account because it makes the individual's mental state the sole arbiter of meaning. The authors require external, objective events to correspond properly with the individual's mental states in order for meaning to obtain. Brogaard and Smith imply further (by insisting on "independent standard of success" (Brogaard and Smith 2005 446) that the general fallibility of the mind (which would presumably include a self-serving bias, a tendency toward self-delusion, hallucinations, and basic misperception characteristic of the individual mind) causes a degree of interference that disqualifies the individual's mental state from judging his success or failure. If the individual's own mental assessment of his success, now in conjunction with objective activities, could ultimately determine meaning, then the individual would only have to become convinced of the relative successfulness of his real activities in order to attain meaning, and his own biases might obscure and even diminish the possibility of his failure to attain meaning- hence the authors' desire for 'independent' standards of success. The authors propose the PMS standard as a remedy that would, allegedly, ensure that each individual life incurs a legitimate trial complete with a nonbiased, 'independent' standard of success and failure.

However, contrary to their contention that public measures, "make meaningfulness something objective," and that they thus impose a substantial 'risk of failure,' the author's appeal to public measures merely makes 'meaningfulness' indirectly dependent upon the intersubjectivity of a number of individuals whose collective consciousness has no privileged knowledge of what constitutes 'success' and which cannot, therefore, impose an absolute, constant risk of failure. Theoretically, any individual could add a psychoactive drug to the water supply and thus artificially generate the "relevant public measure of success" needed for his favored activity to carry

the potential for meaning. Manipulating the mental states of the public at large could guarantee "success" for the individual, irrespective of his real actions. Thus, even if one accepts their criticism of the pure internalist account, Brogaard and Smith have not, so far, eliminated the alleged problem. Unless Brogaard and Smith subscribe to the fallacious notion that sharing an opinion or belief with the largest number of other individuals somehow validates one's position, then their appropriation of 'public measures' seems rather like an arbitrary endorsement of an essentially impotent criterion for meaning. And yet, even as the PMS merely generates a subtle and ineffective distinction between the authors' conceptual model and the allegedly untenable internalist account, the criterion nevertheless creates several additional problems that would not arise in the internalist system.

For example, suppose that Brogaard and Smith have concluded that "public measures of success" are those measures used by a simple majority.[6] What happens then if the referendum on the meaningfulness of one's life ends in a tie? Is the meaningfulness or lack thereof simply indeterminate? What happens if a demographic fluctuation brings a previously meaningful life or activity down to one or two votes below the required majority threshold? Has the individual's life now become conclusively meaningless? Does it become meaningful again if public support fluctuates back upward? Suppose alternatively that Brogaard and Smith merely require the support of a significantly large number of individuals. When an individual or activity begins to lose support, at what point during the decline does the number of supporters cease to be significant? In addition to its imposition of arbitrary restrictions on meaning, this hy-

[666] Thaddeus Metz, "New Developments in the Meaning of Life." *Philosophy Compass* 2, 2 (2007):196–217. 10.1111/j.1747-9991.2007.00061.x. Metz suggests that Brogaard and Smith, "render what is worthwhile hostage to majoritarianism."

pothetical majoritarian system might also makes meaningfulness vague and uncertain.

Still, one must wonder whether such an inelegant mechanism even actually fulfills the author's desire to ground the meaning of life in the will of the public at large. On a practical level how does the public develop, interpret, and sustain 'public' or 'cultural' standards ('either actually or potentially'), and do these standards actually represent the opinion independently formed and shared by the most people? I suspect that these standards arise when a few of the most powerful and influential individuals tip off a trend that reaches critical mass (meaning that a bandwagon effect is achieved) long before the standard in question becomes endorsed by the majority/plurality and that each of the individual opinions do not arise rationally and dispassionately, as if in a vacuum, but amid a myriad of internal and external influences. (A healthy mob often seems equally adept at developing distortions as the most unsound of individual minds.) PMS would thus seem rather undemocratic; far from representing some eternal truth about the meaning of life or even a majority opinion, Brogaard and Smith's 'public measures' might merely represent the warped and transient view of the few who manage to incite the passions of the many. So, whatever virtue Brogaard and Smith find in applying PMS for the purposes of the meaning of life (Perhaps the authors find virtue in democratic process, perhaps they see virtue in simply forcing the individual into association with the real constituents of his community, or perhaps they would rely on an appeal to the (dubious) potential for the advancement of 'standards of success' through a mass exchange of ideas), it seems that the virtue in question would not survive under the circumstance in which some nominally 'public' standards of success actually represented something entirely different.

A Possible Appeal to Susan Wolf

In the text Brogaard and Smith describe public standards of success as "'objective' in the sense that they could be applied by a disinterested observer," but perhaps they would prefer to append their criteria by appealing to the type of objectivity proposed by Susan Wolf. Wolf roughly attributes 'meaning' to the act of loving (where loving means being "gripped, excited, and engaged" by) that which is objectively worthy of one's love (Wolf 2010,). On her view, "not any object will do," (Wolf 2010, 9) for meaning, because not all objects carry the objective value that makes them worthy of love. According to Wolf, "this conception of meaning invokes an objective standard" (Wolf 2010, 9). However, much like Brogaard and Smiths' mixture of the internal and external account of meaning, Wolf wants to combine two popular views-the one which says, "it doesn't matter what you do with your life as long as you love it,"- with the view that "one needs to get involved with something 'larger than oneself'" (Wolf 2010, 10). Wolf regards the latter sentiment "as a way of gesturing toward the aim of participating in or contributing to something whose value is independent of oneself" (Wolf 2010, 11). Her formulation of the 'objective' standard merely requires that "the project or activity must possess a value whose source comes from outside oneself." According to Wolf, this formulation of objectivity, "has the advantage of being minimally exclusive" (Wolf 2010, 39).

Given the weakness of the intersubjective standard, Brogaard and Smith could conceivably wish to assign Wolf's objective ('independent') value to the range of activities that survive their criteria. For example, the authors might argue that 'ambitious and difficult' plans have the kind of independent value that makes them worthy of the individual's time and labor. Additionally, the authors might

recognize the independent/objective value in faring well in the court of public opinion or they might take the view that public opinion constitutes the best source of independent value. However, to this point I think I have sufficiently exposed those defects in the court of public opinion that would make it particularly difficult to accept public approval, either as an independent value or as a creator of independent value. Just like the high court of the United States, the court of public opinion produces deeply divided and warped opinions and creates a climate of vagueness and uncertainty that causes many reasonable people to question its legitimacy.

Furthermore, if Brogaard and Smith were to fully equate the PMS criterion with Wolf's 'independent' sources of value, then the standard would again seem overbroad and ineffective. If, as Wolf says, the standard of 'independent' value may accommodate the artist who labors "for an only dimly conceived posterity," (Wolf 2010, 30) then any advantage that the criterion gains by being 'minimally exclusive,' becomes a disadvantage in the sense that the rule no longer excludes very many people and no longer regulates the pursuit of meaning in any significant way. The individual's 'dimly conceived posterity' could, conceivably, measure the success of any activity, and thus any activity could conceivably pass this criterion.

Finally, as I argued earlier, and as Wolf admits, "Neither I, nor any group of professional ethicists or academicians—nor, for that matter, any other group I can think of—have any special expertise that makes their judgment particularly reliable," (Wolf 2010, 39) and thus neither Wolf nor Brogaard and Smith provide any compelling reason to make 'independent value' an integral part of the formula for meaning. Wolf does suggest that most people have an interest in "being able to see [their] lives as worthwhile from some point of view external to [themselves]…" (Wolf 2010, 32) but this alleged evidence for the standard of 'objectivity' appeals to a feature of individual subjectivity and thus the argument would imply that the applicability of the standard actually depends on a feature of individual subjectivity. If Wolf said to the subjectivist theorist, "most people have an interest in seeing their lives as worthwhile from some point of view external to themselves," then the theorist would only have to retort, "some don't," and this statement would adequately defend his position. It is one thing to appeal to one's intuitions as theorist, but quite another to cite majority sentiment as evidence of a normative criterion for meaning.

Difficult and Ambitious Plans Revisited

Brogaard and Smith's DAP criterion (which calls upon the individual to pursue "difficult and ambitious plans") represents the last remaining pillar of the SNT that might still stand unscathed despite the foregoing criticisms. This requirement places demands on the individual's subjective relationship to the object of his pursuit. Presumably Brogaard and Smith view stretching one's own abilities as an essential aspect of creating meaning, but what makes the DAP criterion uniquely qualified as an essential part of the formula for meaning? Obviously, ambition or difficulty does not, by itself, change or magnify the objective result of individual action in the world- if I accomplish x, then x stays the same whether or not I regard x as a difficult and ambitious goal, so Brogaard and Smith must have designed this requirement to have a desired effect on the individual's mental state. But what mental phenomena arise always and only from the individual's completion of 'difficult and ambitious plans'?

Perhaps most people receive some additional satisfaction or sense of achievement from the completion of difficult and ambitious plans, but this additional sense of satisfaction (or lack thereof) seems rather dependent on some prior value judgment or disposition. Only

an individual who would like to set for himself a challenge would absolutely need to carry out difficult and ambitious plans in order to achieve some additional sense of satisfaction. Without this presupposition, the individual's completion of "difficult and ambitious plans" (Brogaard and Smith 2005, 446) does not necessarily affect his mental state any differently than his participation in seemingly silly activities that he nevertheless holds in high esteem and pursues with passion. If the individual does not value a challenge and does not otherwise retain some requisite mind-dependent relationship to challenging activities in general, then no challenging task can, in virtue of providing a challenge, reliably produce whatever specific effect the authors want to impose on the individual's mental state.

Further, even if all people had the same relationship to challenging activities and thus all experienced the same mental phenomena upon completing difficult plans, what compelling reason does one have to eliminate the possibility that maintaining a similarly powerful relationship to different sorts of pursuits (ones that do not qualify as 'difficult or ambitious') could produce the same mental effect that Brogaard and Smith view as essential for meaning? If, for example, a great virtuoso creates beautiful music with ease but nevertheless values making that music above all else, then does he necessarily fail to attain mental states (that contribute to meaning) sufficiently similar to those experienced by another artist who actually struggles mightily to produce the same music?

As an aspect of the SNT, this criterion describes, most basically, which activities the authors would consider worthwhile, which kinds of engagements they believe deserve any given individual's energies- for Brogaard and Smith, one does not have worthwhile plans if he does not have difficult and ambitious plans. This view, I think, overestimates the uniqueness of the power of ambition's intrinsic features, and oversimplifies the human values and motivations necessary to make any activity 'worthwhile,' either for individual in question or for the community to which the individual belongs.

Review

For some theorists, the 'non-triviality' requirement may represent a plausible and satisfying addition to the formula for meaning, one that would at least compel the individual to focus his energy and reach beyond himself in some form or fashion. As Cottingham argues, use of the term meaning does suggest a certain degree of profundity or seriousness (Cottingham 2003, 21). However, even in the relatively short period of time during which humans have inhabited the earth, nature has produced such an immense variety of individuals, cultures, and perspectives that one must surely struggle to circumscribe, by adding an element of objectivity to his theoretical framework, all and only the activities and aspirations which the individual could plausibly describe as "serious" or "profound." In my view, the nonessential nature of humanity and the inexorable evolution of human striving prohibit any such determinations or exactions. Whatever the purpose, Brogaaard and Smith's use of the SNT, demonstrates the difficulty of defending standards for meaning that depart from the individual's own subjective interests.

Contemporary theorists who generate these mixed accounts of meaning sometimes appeal to objectivity in the abstract. For example, Susan Wolf provides for meaning by positing objectively valuable objects, but she fails to name all of these objects or provide reliable parameters for identifying them. To their credit, Brogaard and Smith appear to have composed a substantive 'mixed' account of meaningfulness that grounds individual subjectivity in concrete, particular mechanisms called "public measures of success." Unfortunately, by almost any measure, the centerpiece of their 'non-

triviality' requirement exhibits weaknesses that vitiate their entire account. Further, the other minor criteria that constitute the authors' 'standard of non-triviality' either interfere with one another, support absurd conclusions, or simply fail to serve any legitimate purpose. However, ambiguities in the text have handicapped the present assessment, and future critical analyses will likely benefit from any additional details that the authors may provide.

Michael Garatoni
Trinity University

References:

Berrit Brogaard and Berry Smith. 2005. "On Luck, Responsibility and the Meaning of Life." *Philosophical Papers.* 44, (3): 443-458

Cottingham, John. 2003. *On the Meaning of Life.* New York: Routlege.

Metz, Thaddeus. 2008. "The Meaning of Life." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford, CA: The Metaphysics Research Lab.

Metz, Thaddeus. 2007. "New Developments in the Meaning of Life." *Philosophy Compass* 2, no. 2 (Feb 2.): 196–217, 10.1111/j.1747-9991.2007.00061.x

Nozick, Robert. 1981. *Philosophical Explanations.* Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Rawls, John. 1971. *A Theory of Justice*. United States of America: President and Fellows of Harvard College.

Wolf, Susan. 2010. *Meaning in Life and Why It Matters*. Princeton, New Jersey: Princeton University Press.

Rescuing Habermas' Knowledge Constitutive Interests

Alexander Meehan
Brown University

IN HIS 1965 ESSAY, "Knowledge and Human Interests: A General Perspective," Jürgen Habermas introduces a key idea in his epistemology and his general philosophy[7]: the idea of knowledge constitutive interests (KCIs). Broadly speaking, KCIs are fundamental human interests that make objective knowledge possible by defining what is meant by, and what counts as, objectivity for a particular form of inquiry. In other terms, KCIs are the interests that constitute the conditions for the possibility of objectivity (CPOs). I will outline some of Habermas' motivations for introducing KCIs, describe their role in his conception of theory, and compare his view on CPOs with that of Immanuel Kant. I will then present a critique of KCIs and attempt to defend Habermas' view from that critique, focusing primarily on the first KCI.[8]

Habermas' 1965 essay can be seen as both a restatement and a revision of Max Horkheimer's 1937 essay, "Traditional and Critical Theory," in which Horkheimer examines and critiques the 'traditional conception of theory.' The distinctive features of the traditional conception can be summarized as by the following three features:

1) Theory has freed itself from dependence on human interests.

7 Many of Habermas' views have changed throughout his career. This paper only deals with views as presented in 1965.

8 By this I mean that most examples I use will refer to the interest in technical control. The purpose of this paper is not to give a close examination of each individual KCI, but to assess the role of KCIs in their broader context as conditions for the possibility of objectivity.

2) In freeing itself from human interests, theory allows us to grasp a domain of pre-existing facts.

3) Grasping this domain of pre-existing facts allows us to devise a proper form of action that will be rooted in truth.

This conception, according to Horkheimer and Habermas, has been held systematically since the Greeks, and has been applied widely to historical and social phenomena in addition to natural phenomena.[9] Habermas follows Horkheimer's footsteps[10] by also developing his version of critical theory—an idea of what theory is that differs from the traditional conception.[11]

Habermas situates the development of his critical theory in an evaluation of Edmund Husserl's 1937 book "The Crisis of the European Sciences" which, according to Habermas, was Husserl's attempt to restore the traditional conception as it came under threat from positivism, a view which denies the possibility of Feature 3 but maintains that Feature 1 and 2 are correct.[12] Max Weber's influential 1918 book, "Science as a Vocation," was a prime expression of this positivistic self-understanding: theory can tell us objective facts about the world but, since the world is itself disenchanted, or devoid

9 In all three cases, the goal is to arrive at the laws that will explain the fundamental features of the subject matter, uncovering the universe as it really is. Then, Feature 3 can be accomplished: "through the soul's likening itself to the theory of the cosmos, theory enters the conduct of life" and spurs "…a process of cultivation of the person".

10 I do not give the details of Horkheimer's critical theory here, as it is out of the scope of this paper.

11 Jürgen Habermas, "Knowledge and Human Interests: A General Perspective," in *Knowledge and Human Interests*, trans. Jeremy J. Shapiro (Boston: Beacon Press, 1971), 304.

12 In other words, it asserts that the theoretical attitude can enter the conduct of life and ground this conduct in truth, but maintains that there is indeed a realm of pre-existing facts of which theory can acquire knowledge, and that we can set aside the pressures of human interests in order to register the truth about those facts.

of purposes and goals, uncovering these facts cannot tell us what we should value or how to live our lives.[13] Husserl, in defending from this positivist intrusion, argues that the sciences have overlooked their rootedness in the life-world—the world as we normally experience it in our everyday lives. Science, especially natural science and physics, illegitimately claims that its sets of causal laws and physical-mathematical theories are "actual" or "true" nature, and substitutes this idealization for the life-world.[14] In fact, says Husserl, knowledge of this "objective" world of facts has its transcendental basis in the pre-scientific world; what counts as a possible object of scientific knowledge is constituted a priori in the "self-evidence" of our life-world.[15] Thus the sciences, as they stand, are not true theories—they have not freed themselves from the interests of the life-world. Only phenomenology, by bringing this fact to consciousness, can transcend these interests and achieve Feature 1. And by these means we achieve Feature 3, for it is the interest-independent quality of the theoretical attitude that gives it practical efficacy. Thus phenomenology is the only pure theory, able to accomplish Feature 1 and Feature 3, while Feature 2 is, in an important sense, an objectivist illusion that does not take into account the priority of the life world.

In this context Habermas introduces KCIs, and presents his own critical theory. He endorses Husserl's view that Feature 2 is an ob-

13 Weber argues that there is a deep divide between factual judgments and value judgments, which accounts for the denial of Feature 3. Theory can tell us how to manipulate variables in order to achieve certain goals or certain values, and it can tell us about the goals and values that have driven humans in the past. However, theory cannot possibly tell us what goals we ought to pursue, or what values we ought to value. This matter is simply a matter of decision—there is no 'correct' choice to which reason can lead us. This divide between the prescriptive and descriptive exists because we live in a disenchanted world.

14 Edmund Husserl, *The Crisis of the European Sciences and Transcendental Phenomenology* (Northwestern University Press: 1954), 221.

15 Habermas, "Knowledge and Human Interests: A General Perspective," 304.

jectivist illusion, for the same reasons.[16] However, Habermas does not think that phenomenology frees itself of human interests, simply in virtue of its awareness of the interdependence of scientific knowledge and interests. Habermas' proposed improvement on Husserl's view[17] is to reject both Feature 1, the assumption that theory can or should be interest-independent, and the idea that interest-independence achieves Feature 3. Theory seemed, all along, to have practical efficacy because it was discovering these pre-existing normative facts. However, theory in any kind of intellectual inquiry, including philosophy, that exhibits Feature 1 cannot possibly have any bearing on human interests given that Feature 2 is an illusion. In reality, theory did not free knowledge from interests—in fact, it had practical efficacy because of its concealed interests. These, until now, concealed interests are the KCIs. KCIs do not compromise objectivity, but rather make objective knowledge possible by defining what is meant by objectivity for a particular area of intellectual inquiry. What counts as a fact is constituted by these fundamental interests.

Habermas assigns a KCI to three main areas of intellectual inquiry: the natural sciences, the historical-hermeneutical sciences, and the critical social sciences, which includes philosophy. The KCI that underlies the natural sciences is the interest in technical control. The natural sciences' aim is to find causal laws which are regular and

16 Ibid.

17 Husserl, argues Habermas, is pulling the rug out from under his own feet: he maintains that Feature 2 is an objectivist illusion, but then claims that accomplishing Feature 1 is what leads to Feature 3. However, accomplishing Feature 3 by way of Feature 1 is only possible because of Feature 2. In other words, the endeavor of 'discovering how humans should act through an interest-independent theoretical attitude' was only considered possible because it was thought that interest-independent, objective facts about how humans should act do exist in the cosmos. In insisting on the intimate connection between Feature 1 and Feature 3 for phenomenology, Husserl is indirectly assuming the existence of these facts, and thus succumbing to aspects of the objectivist illusion contained in the ontological assumptions of the traditional conception.

predictable. This aim is achieved by manipulating one variable and observing the effects on another variable. It so happens that humans, as evolved mammals, have a fundamental interest in controlling their environment and making it predictable.[18] This interest shapes our idea of causal law as we think of it in the natural sciences. For example, Aristotle recognized four types of causes: material cause (physical constitution), formal cause (essence), final cause (purpose), and efficient cause (cause as in cause-and-effect). Claims about efficient cause are predictable and testable by way of manipulation of the environment, while claims about formal and final cause are not. Thus our interest in technical control is the reason that laws regarding efficient cause—as opposed to formal or final cause—count as 'objective' in the modern natural sciences, and the reason that this predictive knowledge is possible in the first place.

The KCI that underlies the historical-hermeneutical sciences is the interest in mutual understanding. Here, the aim is not to arrive at causal laws, but to understand the meanings of historical events and texts. It so happens that humans live together in societies where effective communication, and therefore the ability to understand and interpret, is essential for survival and progress. This interest has, over time, been channeled into more particular avenues of inquiry: the study of religious and legal texts and the analyses of past human events. Habermas outlines how this interest constitutes knowledge in this area of inquiry. He proposes, like Horkheimer, that understanding fundamentally involves a fusion of horizons, or the act of applying the content of a text to one's own situation and pre-

18 For example, the invention of hunting techniques: we need the skill to build weapons that successfully killed large animals. This task requires the manipulation and use of physical objects, with an understanding that certain actions would have certain consequences (e.g. a sharper arrow will better penetrate the animal's skin and cause harm).

understanding. Through this union of interpretation and application, we achieve Feature 3 and thus escape positivism.

The KCI that underlies the critical social sciences, which include psychoanalysis, philosophy, and the critique of ideology, is the interest in emancipation or freedom. Here the aim is to cast off false conceptions and see more truly what the world is like—to liberate ourselves from illusions.[19] This interest has grown out of the fact that humans live with power relations, and that much of human history has consisted of conflicts between oppressed and dominant groups.[20]

Thus what counts as objective in the three areas of scientific inquiry is determined, respectively, by the three KCIs. In framing the KCIs as CPOs, Habermas is in an important sense conducting a transcendental analysis similar to that of Kant's in his 1781 Critique of Pure Reason. Kant was interested in the transcendental CPOs for various forms of inquiry; he wished to lay out the conditions that reality had to meet in order to become knowable. He argued that these conditions come from a subject—a mind—that lies outside the world of experience, and that constitutes the world of experience in virtue of the fundamental features of its thinking. This mind is, metaphorically, an 'offstage eye' that 'sees' the world, but cannot 'see' itself since it transcends the world. It imposes certain a priori concepts or categories which constitute the conditions of possible experience. Since knowledge does not extend beyond experience, nothing can count as knowledge for us unless it is shaped by these categories.[21]

19 Ibid., 310.

20 What counts as a valid proposition in the critical social sciences, says Husserl, is determined by a methodological framework that is based on the concept of self-reflection. And self-reflection is, in turn, determined by our interest in emancipation, for it is self-reflection that "releases the subject from dependence on hypostatized powers."

21 These categories endow a judgment about experience with its objectivity and generality. To get to these categories we must ask ourselves what must be the case in or-

Habermas does not understand his endeavor, however, as a repeat of Kant's. The key difference is that, for Habermas, the CPOs are both transcendental and empirical: transcendental in the sense that they are still prior to the possibility of objectivity, and empirical in the sense that KCIs—unlike Kant's forms of thinking or categories—are part of the real world. Thus Habermas' transcendental (or, perhaps more accurately, quasi-transcendental) conditions are rooted in us as animals in an empirical world, not in us as an 'offstage eye' that transcends the world.

Habermas' critical theory and his idea of KCIs as CPOs can be subjected to a strong criticism. Habermas claims that we only see nature as ruled by causal law because we approach nature with an interest in technical control. However, it is deeply implausible that the reality of causal laws depends on our interests. Causal laws would still exist in nature even if there were no humans alive to learn about them. For example, the fact that 'electrons repel electrons' was true before it was discovered.[22] And the pre-existence of these facts is a good explanation for why we are successful in discovering them through manipulation of variables. The KCI for the historical-hermeneutical sciences has a similar problem. We cannot engage in a fusion of horizons unless we know what it is we are applying to our own situation. Thus we must distinguish between the meaning of a text and its significance. While comprehending the significance of a text may necessarily involve a fusion of horizons, understanding its meaning does not. As with the natural sciences, there is a realm of facts about history and hermeneutics that exist independently, and

der for experience to be possible in the first place. For example, Kant argues that anything that can count as experience for us will contain causal connections.

22 Of course, our ability to discover that 'electrons repel electrons,' and our decision to classify this fact as scientific, may indeed depend on our interest in technical control. But the existence of the law that we discover is not at all dependent on our KCIs; these facts do independently exist, as Feature 2 would hold it.

are objective, regardless of our interests. This realist view is to be contrasted with Habermas' idealist position that what is in reality somehow depends on us. Rather, says the realist, the world has a structure independent of our minds, and that explains how our minds succeed to the extent that they do.

In defending Habermas from this criticism, I should first make two caveats. First, I endorse the realist position that the world has a structure independent of our minds. The realist versus idealist debate is a substantial one, and I will not attempt to defend my position here, except on an intuitive level. For example, it is intuitive that the causal law 'electrons repel electrons' was true before humans existed. Second, I agree that facts are not 'constituted' by our interests, as if our interests somehow altered reality. Many strands of Habermas' argument seem to suggest this, such as the argument that "facts are first constituted in relation to the standards that establish them".[23] I will not attempt to defend this idealist position as it stands, as it seems deeply implausible for reasons already outlined. I do think, however, that if we re-interpret and shift some of Habermas' more dubious ontological claims, then we are left with a compelling and defensible argument for KCIs.

Habermas erred when he turned his knowledge constitutive interests into what might be termed 'reality constitutive interests' or 'truth-value constitutive interests.' He mistakenly made KCIs the conditions for the possibility of the objective truth of a claim. But, I propose, there is still reason to think KCIs are the conditions for the possibility of objective knowledge of a claim. What makes objective knowledge possible and what counts as objective knowledge is more than a question of whether our beliefs are objectively true in the real

23 Ibid., 309.

world. It also matters whether our beliefs are justified[24]—which heavily depends on our minds insofar as they are able to make sense of the world. Thus the view that there are transcendental conditions for knowledge can be plausibly upheld, even with a realist conception of the pre-existence of facts.

We begin, then, by re-interpreting Habermas' idealist claim that 'what counts as facts or reality in an area of inquiry is constituted by fundamental human interest(s)' as the claim that 'what counts as knowledge in an area of inquiry is constituted by fundamental human interest(s).' In the case of the natural sciences, it is our interest in technical control that allows us to see nature as a realm of causal law—thus this interest is a condition for the possibility of natural scientific knowledge. If we could not, in the first place, cognize the fact that there exist these causal laws, then even if we did manage to make some true natural scientific claims they would have no basis or justification—without a methodological approach characterized by our interest in the manipulation of variables, we are left at best with a few 'lucky guesses' that do not fit into a coherent theory. Thus the KCIs define what counts as objective knowledge, not in the sense that they define what counts as objectively true or what counts as objectively false, but in the sense that they define what we can know to be objectively true or objectively false.

Given this adjustment, however, there is a now a large problem for KCIs. As it stands, this is the argument being presented:

i) Humans, as a biological species, have fundamental interests that were present long before the invention of the specialized sciences; we evolved cognitive abilities to meet these interests because this was crucial for individual survival and/or the development of a sustainable society.

ii) Owing to those cognitive abilities, we are able to discover scientific truths.

iii) Thus these fundamental interests make knowledge of scientific facts possible; they are the transcendental conditions for scientific knowledge.

The problem is that this argument is not particularly substantive, especially in its jump from ii) to iii). We might agree that without our evolved cognitive abilities, it would be very difficult to discover scientific truths. We might also agree that, without the KCIs, we would probably not have evolved these cognitive abilities. In this very historical sense, KCIs make objective scientific knowledge possible. However, it would be strange to claim that KCIs are the CPOs merely because, without them, we would not have evolved the cognitive abilities required to do science. For then we could just as consistently categorize the fact that 'humans have sensory organs' or the fact that 'the earth is habitable' as transcendental conditions. The KCI is thus reduced to an item on the long list of historical reasons that humans became epistemologically competent. It seems that in removing Habermas' idealist ontological views from his epistemology, we have rid the KCIs of their fundamental importance.

I claim that if we adopt an externalist account of justification, this really is the end for KCIs. If justification boils down to, for example, the percentage of times our cognitive processes produce a true belief or how the world causes our beliefs, then it is not meaningful to speak of the transcendental conditions for knowledge. If, however, justification depends on what is going on inside our

24 Or whatever we wish to call what it takes to supplement 'true belief' and make it knowledge. At this point, we do not need to restrict ourselves to the traditional 'justified true belief' conception of knowledge; we only need to agree that knowledge requires something more than true belief.

minds—for example on how we are organizing our beliefs and how they support one another—then it is sensible to ask about the conditions under which we can know something, given the world as it exists from our perspective. Therefore, to 'rescue KCIs,' I will assume an internalist account of justification, and conduct a transcendental analysis.

Kant made a division between the phenomenal world as-experienced-by-us and the real or noumenal world that contains 'things-as-they-are-in-themselves.' Since knowledge only extends so far as experience, we cannot know anything about these things-as-they-are-in-themselves. Let us make a similar division in the context of Habermas. Let us say that there is the life-world of facts-as-constituted-by-our-interests, and then there is the real-world of independently pre-existing facts. In the spirit of Kant's 'Copernican Revolution,' we wish to lay out the conditions that reality has to meet in order to become knowable. As Habermas argues, the KCIs "establish the specific viewpoints from which we can apprehend reality as such in any way whatsoever"[25]—and it is impossible to surpass these transcendental limits. Thus this knowable reality, the reality in which we perform theoretical inquiry, is the world of our own subjective perceptions—the life-world. It is like the real-world, except the structure of the human mind has been imposed upon it. Here, therefore, Habermas' transcendental idealism applies—the possible objects of scientific analysis are indeed "constituted a priori in the self-evidence of our primary life-world".[26] Hence we retain the importance of KCIs without making the dubious ontological claim that facts in the real-world have interest-dependent truth-values.

25 Ibid., 311.
26 Ibid., 304.

Nevertheless, it would seem that under this model, humans are 'stuck' in the life-world, without hope of gaining knowledge about the real-world—just as humans cannot hope to gain knowledge about Kant's noumenal world. Furthermore, without some metaphysical account, it seems a matter of coincidence that what counts as a fact in the life-world is also a pre-existing fact in the real world. This is analogous to one of Kant's problems: he wanted to make metaphysics possible by confining its task to the phenomenal world. But, in doing so, he inadvertently gave his own metaphysical view—that things-as-they-are-in-themselves somehow give rise to experience in us while also being free from the categories, i.e. being outside of space-time and not subject to causal law.[27] Kant's transcendental analysis left us with a far-fetched metaphysical picture of the real world, and an insurmountable epistemological divide between the real world and the world we experience.

I argue, however, that my transcendental analysis does not meet the same problem. Because Habermas' CPOs are empirical facts, not features of a transcendent mind, knowledge of the real-world is possible while we are trapped in experiencing the life-world. As a biological species, we evolved to survive in the real-world. We had to be able to manipulate our environment in a way that was actually successful, not just successful according to our own perceptions. Thus, evolution had to "stamp" our minds with perceptions of reality that were actually accurate.[28] These perceptions, and the intuitions and rationalizations that result from them, shape the more sophisticated cognitive processes that bring us knowledge. So, except in the

27 Keith Campbell, *Metaphysics: An Introduction* (Encino, California: Dickenson, 1976), 1-22.
28 For example, it stamped our minds with the ability to see cause and effect precisely because cause and effect is real—an inability to perceive it would entail an inability to survive.

areas where evolution has seriously misled us, when we discover a 'fact' in the life-world we also discover a fact in the real-world. As we intuitively expect, our ability to gain knowledge is limited by the flaws of the evolutionary process. Nevertheless, for the most part, the facts we can come to know while in the life-world are facts that identically obtain in the real-world—this is precisely because the interests that constitute what counts as fact and knowledge are rooted, themselves, in the real-world. Therefore we can still accomplish Feature 2, but we must clarify that the domain of facts that we can know, i.e. the facts constituted by our KCIs in the life-world, is always a subset (and, most likely, a smaller subset) of the full 'domain of pre-existing facts' that obtain in the real-world.

This model, then, attempts to consolidate a realist ontology with Habermas' transcendental idealism. There are several motivations for this seemingly ad hoc approach. To me, Habermas is correct on at least two accounts. First, that knowledge and fundamental human interests are not epistemologically severed, but are intimately connected. Second, that when it comes to deciding what counts as objective knowledge (but not when it comes to deciding what is objectively true), the subjective life-world takes priority over the real world: "representations and descriptions are never independent of standards".[29] Unlike truth, knowledge depends on a human holding a belief in her mind, and that belief being internally[30] justified. The mind's eye is subjective, and imposes a structure on what it perceives. Thus, it is appropriate to 'zoom in' on the mind and find out what could count as an object of knowledge, from its perspective. In this way, we 'rescue' KCIs, while simultaneously rejecting Habermas' ontological claim that facts do not independently pre-exist in the real-world.

29 Habermas, "Knowledge and Human Interests: A General Perspective," 312.
30 Again, I assume an internalist account of justification.

Alexander Meehan
Brown University

References:

Edmund Husserl, *The Crisis of the European Sciences and Transcendental Phenomenology* (Northwestern University Press: 1954).

Jürgen Habermas, "Knowledge and Human Interests: A General Perspective." In *Knowledge and Human Interests*, trans. by Jeremy J. Shapiro, 301-17. Boston: Beacon Press, 1971.

Keith Campbell, *Metaphysics: An Introduction* (Encino, California: Dickenson, 1976), 1-22.

Poincaré's Philosophy and the Development of Special Relativity*

Emily Adlam
Oxford University

AMONG HENRI POINCARÉ'S MANY and diverse scientific achievements were a number of contributions to the theory of special relativity. In this paper I compare the approaches taken by Poincaré and Einstein, aiming to clarify questions about the true extent of Poincaré's contribution, and I investigate the relationship between Poincaré's philosophy and his science, arguing that his approach to the development of special relativity can be traced back to his acceptance of a 'convenience thesis' about conventions. This provides a useful insight into the way in which the adoption of some philosophical position can influence scientific progress, as the divergence between Einstein and Poincaré can be understood as owing in part to a difference in philosophical viewpoints about the methodology of science.

1. Did Poincaré Discover Special Relativity?

Poincaré's work introduces many ideas that subsequently became important in special relativity, and on a cursory inspection it may seem that his 1906 paper 'Sur la dynamique de l'electron,'[31] (written before Einstein's landmark 1905 paper[32] was published) already contains most of the major features of the theory: he had the

*Author's appendix available online at bit.ly/12Jswpa

[31] Poincaré, *Sur la dynamique de l' electron*. 129-175

[32] Einstein, *On the Electrodynamics of Moving Bodies*, 891-921.

correct equations for the Lorentz transformations[33], articulated the relativity principle[34], derived the correct relativistic transformations for force and charge density[35], and found the rule for relativistic composition of velocities[36]. His conceptual approach also seems to have much in common with Einstein's; for instance, in 1898 he gave a derivation of Lorentz's 'local time' coordinate that closely mirrors Einstein's derivation of the Lorentz transformations, particularly in its use of an almost identical convention for synchronising spatially separated clocks[37]. Nonetheless, a closer reading of Poincaré's papers reveals that his understanding of both the relativity principle and the Lorentz transformations is significantly more limited than Einstein's.

1.1 The Relativity Principle

One element which links the work of both Poincaré and Einstein is a preoccupation with the principle of relativity. But it is important to be aware that Einstein and Poincaré were not working with precisely the same principle. Compare their two formulations:

[33] A set of equations describing how measurements of space and time made in different frames of reference are related.

[34] See section 1.1

[35] In classical physics, charge density is the same in all frames of reference, but in special relativity, if the charge density in the frame in which the charges are stationary is ρ and frame 2 is moving at speed v relative to this frame, then the charge density in frame 2 is $\gamma\rho$, where γ is the Lorentz factor $\sqrt{1-\frac{v^2}{c^2}}$, where c is the speed of light . Again, in classical physics force density is the same in all frames of reference, but in special relativity it transforms similarly to charge density.

[36] In classical physics, if the velocity of A relative to B is \textbf{v} and the velocity of B relative to C is \textbf{u}, then the velocity of A relative to C is the vector sum \textbf{v} + \textbf{u}. But velocities are no longer additive in special relativity; if \textbf{v} and \textbf{u} are parallel to each other, then special relativity tells us that velocity of A relative to C is given by $\frac{v+u}{1+\frac{vu}{c^2}}$, where c is the speed of light. (The relationship is more complex if the velocities are not parallel).

[37] Poincaré, *The Measure of Time*, 224-234.

Poincaré: 'the laws of physical phenomena must be the same for a motionless observer and for an observer experiencing uniform motion along a straight line.'[38][39]

Einstein: 'The laws by which the states of physical systems undergo change are not affected, whether these changes of state are referred to the one or the other of two systems of co-ordinates in uniform translatory motion.'[40]

The crucial difference is that Poincaré finds it necessary to refer to an observer, while Einstein does not. The distinction has been noted by Katzir: 'In contrast to Einstein, who denied the existence of absolute motion, Poincaré denied the possibility to detect it.'[41] As a result, Einstein's principle leads to stronger constraints: for Einstein, there can be no difference at all between the forms of the laws of nature in different inertial frames, whereas Poincaré can accept that the laws of nature take one form relative to a privileged frame and a more complicated form relative to all other frames, provided they work together in such a way that this difference between frames does not have any observable consequences. However, the importance to be attached to this distinction turns upon our understanding of what is meant by a 'law of nature.' Poincaré was famously a conventionalist, and therefore held a view of lawhood which

[38] In his 1902 essay 'Relative and Absolute Motion', Poincaré gave a different formulation of the principle of relativity, omitting the reference to an observer: 'the movement of any system whatever ought to obey the same laws, whether it is referred to fixed axes or to the movable axes with are implied in uniform motion in a straight line.' (1902, p.111) But this version appears in a philosophical paper rather than a scientific one, and as we shall see, Poincaré's scientific views must be kept separate from his philosophical ones.

[39] Poincaré, *L'etat et l'avenir de la Physique mathematique*, 302-324.

[40] Einstein, *On the Electrodynamics of Moving Bodies*, 891-921.

[41] Katzir, *Poincaré's Relativistic Physics*, 268-292.

may appear to make the two principles equivalent after all. He claimed that the modern notion of a law is 'a constant relation between the phenomena of today and tomorrow, i.e. a differential equation,'[42] which suggests that on his view, a law is nothing over and above relations between observable phenomena. If taken seriously, this suggests that the true form of the laws of nature cannot be different to the phenomenological laws that would be formulated on the basis of observations within any given frame: there is no distinction to be made between 'the laws for an observer' and the actual laws of nature.

But if Poincaré was consistent about this view of lawhood, his acceptance of the principle of relativity would surely force him to abandon the notion of a privileged frame of reference: if there are no laws of nature above and beyond relations between what is observable, and if observation can never disclose a privileged frame of reference to us, it follows that there is no such frame. Yet Poincaré's theories remain tied to the notion of a privileged frame of reference; for instance, in his 1906 paper 'On the Dynamics of an Electron,' he gives an analysis of the motion of an electron in which he continues to refer to the 'real electron,'[43] meaning the electron as it appears to observers in the ether rest frame. Whatever Poincaré might believe in the context of philosophy, in the context of his scientific work he assumes that at least some of the actual laws of nature are distinct from the laws formulated by observers. In this case, the actual laws of nature, which single out a privileged rest frame, conspire to produce the same observable effects in all inertial reference frames so that all observers in such reference frames will formulate the same laws on the basis of their observations. Further evidence of this approach can

[42] Poincaré, *L'etat et lavenir de la Physique mathematique*, 302-324.
[43] Poincaré, *Sur la dynamique de l'electron*, 129-175.

be found in the status he accords to the relativity principle, which he takes to be an empirical result, an inductive generalisation from the null results of ether drag experiments such as the Fizeau experiment, the Trouton-Noble experiment, and the Michelson-Morley experiment—thus, for example, he writes that the principle is 'so far in agreement with experiment ... (but may) be later confirmed or disproved by more accurate tests.'[44]. Hence it is reasonable to suppose that, despite Poincaré's philosophical views about to lawhood, in practice he would recognise a distinction between 'the laws for an observer', which are the object of his own relativity principle, and the laws of nature simpliciter, which are the object of Einstein's.

Nor is this distinction a trivial one. An important motivation for Zahar's claim[45] that Poincaré deserves a major share of the credit for the discovery of special relativity is that he 'obtained, as a first heuristic component of his programme, the Principle of Lorentz covariance which is in effect a symmetry requirement,' and this heuristic has been crucial to the development of special relativity. However, because Poincaré's relativity principle is subtly different to Einstein's, he also had a different understanding of the principle of Lorentz covariance. Like Einstein he imposes the requirement of Lorentz covariance on all the fundamental equations of nature, but the equations in question are still understood relative to the ether rest frame and are never referred to any other frame of reference, and therefore for Poincaré, Lorentz covariance is a condition on the solutions to some set of equations relative to a single reference frame. For Einstein on the other hand, the motivation behind Lorentz covariance is that the equations should be unaffected by the coordinate transformations because the principle of relativity demands that they

[44] Poincaré, *Sur la dynamique de l'electron*, 129-175.
[45] Zahar, *Poincaré's Philosophy*.

should actually take the same form in all frames, not merely that they should appear to do so. Thus although Poincaré's Lorentz covariance condition is mathematically equivalent to Einstein's, it is very differently motivated and therefore as a heuristic it offers a different sort of guidance. It is therefore not entirely accurate to credit Poincaré with the invention of the methodology of later research in special relativity, because although he provided the mathematical background for this heuristic, he did not fully appreciate its physical interpretation.

1.2 The Lorentz Transformations

Since the Lorentz transformations are at the heart of the theory of special relativity, Poincaré cannot fairly be said to have been one of the originators of the theory unless his understanding of the transformations was sufficiently close to our modern understanding of them. It is often assumed that Poincaré, like Einstein, thought of the Lorentz transformations as a procedure for changing between different coordinate systems—so for instance, Brown claims that Poincaré was the first to use the generalized relativity principle as a constraint on the form of the coordinate transformations,'[46] and Zahar writes that 'Poincaré eliminated (the Galilean coordinates) in favour of the Lorentz-transformation, a transformation which goes directly from the rest frame to the effective coordinates.'[47] However, I submit that Poincaré never regarded the Lorentz transformations as coordinate transformations: he saw them as a convenient mathematical tool rather than a physical relationship between actual frames of reference.

[46] Brown, *Physical Relativity*.
[47] Zahar, *Einstein's Revolution*.

A preliminary reason to doubt that Poincaré understood the transformations in a physical way is that he never gave a derivation for them, which suggests that until Einstein's 1905 paper he was not aware that they can be obtained directly from elementary assumptions about the procedure for changing between coordinate systems of different reference frames. Therefore it is unlikely that in his original work on the transformations he intended them to be understood as a procedure for changing between coordinate systems – yet that conceptual understanding is crucial to the role of the transformations in the fully developed theory of special relativity. Moreover, the reasoning Poincaré uses when working with the transformations is consistently abstract and mathematical rather than physical: for example, although he did not provide a derivation of the general form of the transformations, in his 'Sur la dynamique de l'electron' he did offer a derivation of the factor L, which is common to all the transformed coordinates and is left undetermined by the requirement that the transformations should preserve the form of the Maxwell equations. But his derivation involves no physical considerations: he reaches his conclusion simply by applying certain mathematical constraints[48] (see Appendix). For all the elegance of this method, it seems an unusual approach to take to prove something which, if the Lorentz transformations are interpreted physically, determines the extent to which a moving object contracts in the transverse direction, and is therefore an important empirical feature of the world. The strategy would not be unjustifiable, since there are intuitively plausible symmetry principles which offer physical reasons to believe that the Lorentz transformations should form a group, but Poincaré never invokes them; he merely asserts that 'We must regard L as being a function of β, the function being chosen so that this partial

[48] Poincaré, *Sur la dynamique de l'electron*, 129-175.

group, which will be denoted by P, is itself a group.'[49]. This approach seems more in harmony with the idea that the Lorentz transformations are a convenient mathematical tool, since then we are free to stipulate that they should have any mathematical property which we find convenient and which does not interfere with their role in calculation.

It seems that Poincaré's interpretation of the Lorentz transformations was rather far removed from Einstein's physical view of them. Nonetheless, if the ways in which he used the transformations are sufficiently similar to their applications in special relativity, perhaps he can be said to have had a partial understanding of the transformations in virtue of his appreciation of their practical role. It is therefore important to examine Poincaré's ideas about the function of the Lorentz transformations, as distinct from their theoretical origins. Clearly the transformations express a relationship between two sets of coordinates: but what do these coordinates signify? For Einstein, the coordinates x, y, z, t describe the spatiotemporal location of some event with respect to an inertial frame S, and the transformed coordinates x', y', z', t' describe the spatiotemporal location of the same event as it would be measured by an observer at rest in a frame moving at speed v with respect to S (provided that the Einstein synchrony convention is used). But for Lorentz and Poincaré, the transformation is applied very differently. We consider a physical system which is in motion with respect to the ether rest frame and suppose that the coordinates x, y, z, t describe configurations of the parts of that system with respect to the ether frame. When we apply the Lorentz transformations we obtain transformed coordinates x', y', z' and t' which describe what Poincaré calls the 'ideal' system, in contrast with the 'real' system.

[49] Poincaré, *Sur la dynamique de l'electron*, 129-175.

For example, in discussion of a hypothesis put forward by which implies that the electron does not undergo length contraction, Poincaré writes that 'upon applying the Lorentz transformation, since the real (moving) electron is spherical, the ideal electron will become an ellipsoid.'[50] Here the Lorentz transformations are applied to a real, moving, spherical electron to give an alternative coördinatization which makes it an ellipsoid. Crucially, this alternative description is not referred to any physical reference frame, nor given any physical interpretation. Indeed, Poincaré never associates the transformations with the process of changing reference frames, nor does he demonstrate any awareness that there could be interesting physics related to changes in reference frames—his analyses are always carried out in the ether rest frame and the motion involved is always absolute.

But if the Lorentz transformations do not express relationships between two physical reference frames, what is their purpose? In the theory advocated by Poincaré, their main function is to permit the formulation of the hypothesis that when a system is set in motion with respect to the ether, it undergoes certain changes in its configuration such that when we apply the Lorentz transformation, we obtain an 'ideal' system which is at rest in a corresponding 'ideal' coordinate system and has the same configuration as the real system has when it is at rest in the ether rest frame. Clearly the real system can be recovered by applying the inverse transformations, and since the Lorentz transformations form a group, it follows that the real moving system is related to the ideal system (i.e. the system at rest) by a Lorentz transformation—that is, the hypothesis amounts to the requirement that when a system is set in motion it turns into the corresponding Lorentz transformed system. Unfortunately neither Lorentz nor Poincaré ever gave a complete and explicit statement of

[50] Poincaré, *Sur la dynamique de l'electron*, 129-175.

this hypothesis, which Janssen calls the 'generalised contraction hypothesis,' but it is undoubtedly necessary if the use of the Lorentz transformations is to achieve the avowed aim of preserving the relativity principle: as Janssen points out, 'the configuration of a material system at rest in the ether will have to change upon setting the system in motion if it is to generate the electromagnetic field configuration in the moving frame that is the corresponding state of the electromagnetic field configuration generated by the system at rest in the ether.'[51] We know that the system at rest obeys the Maxwell equations, and the Lorentz transformations were chosen specifically to preserve Maxwell's equations, so if the moving system changes in the way hypothesized, it will continue to obey the Maxwell equations. If it is the case that not only the Maxwell equations but all laws of nature are left unchanged by Lorentz transformations, we can prove that the relativity principle will never be violated, because the same laws of nature that are obeyed by the system at rest will also be obeyed by the moving system: as Poincaré puts it, 'two systems, of which one is fixed and the other is in translatory motion, become exact images of each other.'[52] Thus the purpose of the Lorentz transformations, in Poincaré's theory, is simply to enable us to determine what changes the system must undergo when it moves with respect to the ether in order that it will still satisfy the Maxwell equations and any other Lorentz covariant equations.

It is important to be aware that although Poincaré used the Lorentz transformations to ascertain the changes necessary to preserve the relativity principle, he did not take the transformations to be a cause or an explanation of these changes—he thought it necessary to postulate distinct explanations for length contraction, local time, and

[51] Janssen, *A Comparison*.

[52] Poincaré, *Sur la dynamique de l'electron*, 129-175.

other such phenomena. He dealt with the contraction dynamically, writing that 'a special force must be invoked to account for both the contraction and the constancy of two of the axes,'. [53] Accounting for the phenomenon of local time is less straightforward: Lorentz avoided the issue by asserting that the difference between local time and actual time is too small to be significant, but Poincaré realised that without some way of accounting for the required change in the temporal coordinate, the theory would conform to the relativity principle only approximately, which he found unacceptable. He therefore sought to show that if clocks were synchronised according to a particular synchrony convention, equivalent to the one which Einstein later adopted, clocks in moving systems would read off local time rather than real time, thus effecting the required change.[54]

In summary, Poincaré's use of the Lorentz transformations differed from Einstein's in two key ways. First, for Poincaré, the transformations are not used to give relations between any two inertial frames of reference—they are defined only relative to the ether rest frame, so that the velocity v appearing in the formula must always refer to a velocity with respect to the ether rest frame, i.e. an absolute

[53] Poincaré, *Sur la dynamique de l'electron*, 129-175.

[54] However, Poincaré's derivation only gives the equation for local time which appeared in the original, first-order version of the transformations, not the exact transformations which Lorentz published in his paper 'Simplified Theory of Electrical and Optical Phenomena in Moving Systems' (Lorentz, *Simplified Theory*, 427-442): Poincaré's local time is given by the formula $t' = t - vx/c^2$, whereas the exact time transformation is given by $t' = \gamma(t - vx/c^2)$. The derivation was therefore adequate in 1898 when Poincaré first provided it, but had gone out of date by the time of his 1906 paper. However, no steps were taken by either Lorentz or Poincaré to show that the time coordinate would change in a way consistent with the exact version of the transformations; Poincaré must surely have been dissatisfied with this result, but perhaps he hoped to resort to Lorentz's earlier strategy and claim that the new temporal coordinate required by the exact transformation would be so close to the local time that the difference would not matter -an assumption that would be justified in most cases, since γ is close to 1 unless the velocities involved are very large.

velocity. Moreover, even if we restrict ourselves to transformations involving the ether rest frame, Einstein's usage does not coincide with Poincaré's, since for Einstein the transformations express a relationship between coordinate systems, whereas for Poincaré they are merely a means of predicting the physical changes that a system undergoes when set in motion relative to the ether.

1.3 The direction of explanation

There is some controversy over the true nature of the disagreement between the theories of Poincaré and Einstein. The difference is not empirical—Janssen[55] shows that although the Lorentz-Poincaré theory before 1905 is not exactly empirically equivalent to special relativity, it can be made so with minimal alterations. It might seem that the main point of difference is ontological: Poincaré's theory makes essential reference to the ether and thus to a privileged rest frame, while Einstein's theory does not. But we should not attach too much importance to this fact, for Einstein was careful to point out that his theory does not actually rule out the existence of the ether. Another popular view, supported by Goldberg[56], Miller[57] and Hirosige[58], is that the differences stem from Poincaré's commitment to the electromagnetic world-picture, upon which the only basic constituents of the world are charged particles and electromagnetic fields. But this does not seem to account completely for the distinctions between the theories; after all, the principle of relativity and the light postulate could certainly be true even in a wholly electromagnetic world and Lorentz covariance could still be derived from them,

[55] Janssen, *A Comparison*.
[56] Goldberg, *Henri Poincaré and Einstein's Theory of Relativity*, 934-944.
[57] Miller, *A Study of Henri Poincaré*, 320.
[58] Hirosige, *The Ether Problem*, 3-82.

so such a commitment would not suffice to prevent Poincaré from taking Einstein's approach. Moreover, Katzir[59] points out that in the 1906 paper, Poincaré accepts Lorentz's model of the electron, which is compatible with the relativity principle but not the electromagnetic worldview, over Abraham's model, which is compatible with the electromagnetic world view but not the relativity principle, and goes on to invoke the relativity principle as the reason for his choice. This demonstrates that Poincaré was willing to put the relativity principle above any attachment he may have had to the electromagnetic worldview, and therefore that view should not have interfered with his work on the relativity principle.

Nonetheless I think the intuition that the electromagnetic worldview plays some role here does contain an element of truth: the differentiating factor was not Poincaré's commitment to the electromagnetic world-picture itself, but to the explanatory strategy associated with it. He was willing to accept the existence of particles which are not charged and forces which are not electromagnetic, but he remained faithful to the underlying motivation for the electromagnetic picture, believing that all observable phenomena should be accounted for by appeal to the nature of the fundamental particles and forces. This idea was certainly not unique to proponents of the electromagnetic worldview; a similar motivation lies behind the mechanical world-picture, upon which all macroscopic phenomena are produced by Newtonian interactions between moving microscopic particles. Indeed, a commitment to the explanation of the macroscopic in terms of the microscopic seems to have been a general characteristic of physics in the era leading up to Einstein—for example, beginning with Boltzmann's 1872 H-theorem, it was an on-going pro-

[59] Katzir, *Poincaré's Relativistic Physics*, 268-292.

ject to show that the macroscopic laws of thermodynamics could be derived from assumptions about microscopic features of systems.

This explanatory strategy is a consistent feature of Poincaré's work, as exemplified by his assumption that the physical changes predicted by using the Lorentz transformation must be explained by piecewise derivation from force laws and microscopic phenomena. It is therefore not surprising that Poincaré never thought to view the relativity principle as explanatory in and of itself; as Katzir puts it: 'instead of deducing consequences from (the relativity principle), he used it mainly to confirm or refute various hypotheses.'[60] Poincaré regarded the principle rather like a general summary of the empirical evidence, such that those theories which violated it could be taken to have been indirectly disconfirmed. For instance, in his 1905 paper, he offers a proof that 'Lorentz's hypothesis (about the contraction of the moving electron) is the only one which is compatible with the impossibility of manifesting absolute motion,' and claims that 'Lorentz's analysis is thus fully confirmed.'[61] The principle of relativity thus functions as supporting evidence for the contraction hypothesis, but not as an explanation for the contraction, since Poincaré goes on to offer an entirely separate explanation in terms of a force law, insisting that if one believes the electron contracts, 'one must admit ... the existence of a supplementary potential proportional to the volume of the electron.' Moreover, even after the publication of Einstein's paper, Poincaré did not accept the use of the relativity principle to explain such phenomena: Pais[62] emphasizes the fact that even in 1908, Poincaré was unwilling to take length contraction as a consequence of the relativity principle together with the light postulate. This makes sense in light of the fact that according to Poincaré

[60] Katzir, *Poincaré's Relativistic Physics*, 268-292.
[61] Poincaré, *Sur la dynamique de l'electron*, 129-175.
[62] Pais, *Subtle Is the Lord*, 224-234.

and his peers, all legitimate scientific explanations ought to involve an appeal to microscopic laws, so in their view relativity principle was simply not the right kind of thing to function as an explanans.

Einstein's 1905 paper was revolutionary precisely because he broke with the long-standing tradition of explaining the macroscopic in terms of the microscopic. Rather than taking force laws as fundamental, he made the relativity principle the basic axiom of his theory and used it to derive constraints on the form of the laws governing phenomena at both the microscopic and macroscopic levels. This means that the scientific community's adoption of special relativity involved not only a theory change, but an alteration in general ideas about what is to be expected and what requires explanation. For an ether theorist, or anyone committed to the existence of a privileged rest frame, both absolute velocity and absolute acceleration are real features of the world and it is therefore prima facie to be expected that we should be able to detect them. The fact that we cannot detect absolute velocity then seems to stand in need of explanation, hence the intuition on the part of Poincaré and his peers that it is important to present a derivation of the relativity principle. But suppose we discard the notion of an absolute rest frame; it follows that there is no such thing as absolute velocity. Our inability to detect absolute velocity is therefore entirely to be expected, and the relativity principle requires no derivation.[63]

[63] Notice that we can rid ourselves of absolute velocity without also discarding absolute acceleration, since we can claim that all inertial reference frames are equivalent without getting rid of the notion of inertial reference frames altogether. If we choose to justify the claim that there is no privileged rest frame by the relationalist strategy of denying the existence or causal efficacy of absolute spatial structure, then the fact that we can apparently detect absolute explanation will stand in need of explanation, which might for instance be offered by taking acceleration to be relative to the total distribution of mass in the universe (Barbour, *The Discovery of Dynamics*), but the explanatory strategy offered here is perfectly coherent without any such extension.

2. Poincaré and Conventionalism

Arguments to the conclusion that Poincaré anticipated the theory of special relativity often invoke his philosophical beliefs as reasons to think he was tending towards a special relativistic interpretation of his theory. For example, Zahar claims that 'Poincaré looks upon the effective coordinates as the only physically significant parameters (because) he was not wedded to the classical ontology according to which absolute time, the ether frame and the Galilean coordinates are the only intelligible entities.'[64] But although Poincaré found the effective coordinates useful, he continued to see the Galilean coordinates as the true description of reality: the effective coordinates are taken to be a mistake that we make in consequence of the behaviour of objects in motion. Zahar is describing not what Poincaré actually did, but what it seems that he logically should have done in light of his philosophical commitments. This highlights an important feature of Poincaré's practice: he is often inconsistent about applying his philosophical principles to his scientific work. In this section I will examine some apparent contradictions between his philosophy and his science, then examine a resolution of this conflict and consider the impact of this analysis on Poincaré's role in the development of special relativity.

2.1 Poincaré's Science and Poincaré's Philosophy

A comparison between Poincaré's philosophical views and his scientific theories can easily create the impression of a contradiction. In particular, his theories are often grounded on ontological assumptions which he has explicitly denounced as 'purely conventional' in

[64] Zahar, *Poincaré's Philosophy*.

his philosophical papers. For instance, in his 1898 paper on time, Poincaré makes it clear that the constancy of the one-way speed of light is a matter of convention: he writes that we begin 'by supposing that light has a constant velocity, and in particular that its velocity is the same in all directions. That is a postulate without which no measurement of this velocity could be attempted.' Yet in his 1900 paper on Lorentz's theory, he describes two observers in motion relative to the ether but stationary relative to each other, and claims that: 'they are not aware of their common motion, and consequently believe that the signals travel equally fast in both directions.'[65] If the constancy of the speed of light relative to the ether were a factual matter, then the signal travelling in the opposite direction to the observers' own direction of motion would be travelling more quickly than the other in their frame of reference, and therefore they would indeed be mistaken in their belief that the signals are travelling equally fast in their own reference frame. But if the constancy of the speed of light is merely a convention, as Poincaré has claimed, it is not possible to make a mistake about the one-way speed of light, since the assignment of some value to the one-way speed in any frame of reference merely a choice of convention. Thus there seems to be a contradiction between his views on the conventional status of the one-way speed and his suggestion that the observers are in error.

Darrigol[66] suggests that this contradiction can be resolved if we assume that Poincaré's meaning in 1898 was that the constancy of the one-way speed of light is a convention only in the ether frame—this is certainly implied by his comment that the convention is 'accepted by everyone,'[67] because clearly, in virtue of the prevalence of the ether theory, the constancy of the speed of light in all other

[65] Poincaré, *The Theory of Lorentz*, 252-278.
[66] Darrigol, *The Mystery of the Einstein-Poincaré Connection*, 614-626.
[67] Poincaré, *The Measure of Time*, 224-234.

frames was not accepted by everyone at the time. However, I find this interpretation problematic. If the constancy of the one-way speed of light is only a convention in the ether rest frame, what meaning does it have to say that the two observers are wrong to judge that it is constant in their own frame? Presumably they are wrong with respect to the convention that it is constant in the ether rest frame—but if this is only a convention, surely there is nothing to stop them from establishing whatever convention they like in their own frame. Of course, this would disagree with the convention in the ether rest frame, but the disagreement would never lead to any explicit contradictions, because we cannot establish empirically which frame is in fact the ether rest frame. Indeed, because of our inability to identify the ether rest frame, this convention would amount to nothing more than an agreement about how we should talk about motion with respect to different frames if, per impossible, we could distinguish the ether rest frame; it would tell us nothing about how to refer to motion in any actual reference frame. Given that the role of a convention is to determine how we ought to behave and speak in certain situations, it is somewhat implausible that we should have any convention which applies only to the ether rest frame; it seems much more likely that Poincaré intended to claim that the constancy of the speed of light is a convention in all frames of reference, in which case the contradiction stands.

A similar conflict exists between Poincaré's philosophical views on absolute space and his use of the notion of space in his scientific work. He could not be clearer about his opinion of the notion of absolute space: 'Whoever speaks of absolute space uses a word devoid of meaning.'[68] Yet his scientific writings consistently presuppose the existence of a privileged frame of reference and suggest that only ob-

[68] Poincaré, *The Measure of Time*, 224-234.

servers who are at rest in this frame of reference see phenomena as they really are. This is not an explicit contradiction, because for Poincaré the privileged frame of reference is merely the ether rest frame––for instance, he defines the 'absolute motion of the earth' as 'its motion relative to the ether instead of relative to other celestial bodies.'[69] He is therefore able to talk about absolute motion without presupposing the existence of absolute space. Nonetheless, the use of the ether rest frame provides a means of avoiding the consequences of the denial of absolute space—it is essentially a way of reconciling the conventional view that there is such a thing as the true velocity and configuration of an object with the philosophical view that talk of absolute space is meaningless. This strategy therefore permits Poincaré to go on doing physics relative to a single frame, ignoring the interesting consequences which arise from the consideration that all frames of reference are in fact equivalent. The treatment of space in Poincaré's scientific work may not directly contradict his conventionalism, but it certainly seems to be in tension with the spirit of his philosophy.

Such tensions even extend to Poincaré's view of the relativity principle. In his scientific papers it is consistently regarded as an empirical fact, an inductive generalisation from experiment which requires explanation in terms of more basic theories. Yet in his philosophy he claims that there are two reasons to believe in it: the first is its confirmation by experiment, but the second is that 'the contrary hypothesis is repugnant to the mind,'[70] which is presumably intended to imply that there is an a priori element to our acceptance of it. But if we have a priori reasons for believing in the principle, the demand for its explanation in terms of more fundamental theories loses

[69] Poincaré, *Sur la dynamique de l'electron*, 129-175.
[70] Poincaré, *Science and Method*.

much of its urgency, so if this was really Poincaré's view is it puzzling that he devotes so much effort to deriving it from more fundamental hypotheses—indeed, it is surprising that he never thought to do as Einstein did and take it as an axiom from which other hypotheses can be derived.

2.2 Resolving the contradiction

The apparent contradictions between Poincaré's philosophical beliefs and his scientific practice can be understood as a consequence of the balance he was required to achieve between his philosophy and the practical demands of science. Clearly, Poincaré's conventionalism was so extreme that science could not possibly produce coherent and developed scientific theories if it were to abandon everything that he believed to be conventional. Thus Poincaré always makes it clear that the assertion that some rule or principle is conventional is not at all equivalent to the claim that we should cease to use that rule or principle; his approach to the circumstances in which it is appropriate to reject a convention is much more nuanced. In his 1904 essay, 'The Future of Mathematical Physics,' he discusses this issue, observing that although experimental results can never contradict a convention, the convention will nevertheless be threatened if in order to retain it we are forced to add ad hoc hypotheses such that it ceases to be predictively useful—for instance, the attempt to explain the apparent violation of conservation of energy in the case of radiation from radium by the hypothesis that unobservable quantities of energy are constantly travelling through space in all directions, and some of this energy is converted into an observable form inside radium. Poincaré is of the opinion that in such circumstances the relevant convention should be abandoned, but we should 'abandon the conventions only after having made a loyal effort to save them.'[71] Indeed, in papers such as his 1900 contribution to the Lorentz Festschrifte, he makes a valiant attempt to preserve the principle of action-reaction in the face of its apparent violation by electrodynamic phenomena.

Poincaré often refers to the network of conventions used in science as a framework; so for instance he claims that 'space is another framework which we impose on the world.' The fact that all his scientific work presupposes the existence and reality of space therefore reflects a deliberate choice to work within the traditional framework of science. Indeed, his commitment to retaining conventions whenever possible makes it clear that he does not see it as the place of the scientist to question this framework; as Stein observes, 'the basic mathematical presuppositions of physics were seen by Poincaré as defining a framework within which it is the task of the theoretical physicist to fit all phenomena.'[72] With this in mind, we can understand why Poincaré insisted on presupposing long-established scientific conventions in his own scientific work, for example, by producing explanations which conform to the traditional explanatory strategy of explaining macroscopic phenomena by appeal to the motions of microscopic particles in absolute space. He recognised that this explanatory framework is a freely chosen convention, but he believed that the role of science is to construct theories within this framework, not to investigate the nature of the framework itself.

Einstein, on the other hand, is much more willing to discard conventions—perhaps the clearest example is his acceptance of the relativity of simultaneity. Poincaré would theoretically have agreed with Einstein that absolute simultaneity is merely a convention, but, unlike Einstein, he always retained the convention in his scientific

[71] Poincaré, *L'etat et lavenir de la Physique mathematique*, 302-324.
[72] Stein, *Physics and Philosophy Meet*.

work. However, we should be aware that Einstein certainly did not get rid of everything that he believed conventional. For example, he retains the fiction that distances and coordinates within a single frame have objective, determinate values, even though he acknowledges that positions can only be defined 'by the employment of rigid standards of measurement,'[73] which suggests that he had an awareness of the conventional nature of such standards. Thus Poincaré and Einstein were in agreement that science can progress only with the support of a network of descriptive conventions; the difference is merely that Einstein was willing to be more flexible about which conventions are retained.

2.3 The Convenience Thesis

Given that Poincaré's approach to the relativity principle and the Lorentz transformations was derived in part from his insistence on adhering to standard scientific conventions, we must consider why he believed this to be the right methodology for science if we are to understand the fundamental source of his divergence from Einstein. I suggest that the answer can be found in his philosophical views regarding the origin of our scientific conventions. One feature of Poincaré's conventionalism that makes it particularly nuanced and interesting is his insistence that experience guides us in choosing conventions: for instance, after arguing that the geometry of space is a matter of convention, he writes that 'experiment ... tells us not which is the truest, but which is the most convenient geometry.'[74] Similarly, after arguing that the laws of acceleration and composition of forces are conventions, he writes that 'they would seem arbitrary if we forgot the experiences which guided the founders of science to their adoption and which are, although imperfect, sufficient to justify them.'[75] As Ben-Menahem puts it, Poincaré 'critiques both an oversimplified conception of fact and an equally oversimplified conception of convention,' and this gives his conventionalism much greater plausibility than versions upon which conventions are taken to be arbitrary[76]. However, there are two related theses in this vicinity, which neither Poincaré nor later commentators have adequately distinguished. The first is the claim that not all conventions are equally good: as Ben-Menahem writes, 'the choice of a coordinate system or measurement unit is intricately linked to objective features of the situation.'[77] Thus we can accept that certain conventions cannot be judged true or false, but still argue that it is possible to make rational choices between conventions for practical reasons. The second is what I will call the convenience thesis: the actual process by which conventions come to be selected is such that the conventions we ultimately choose are the most convenient, so in the sciences we somehow end up selecting the conventions which allow us to express the laws of nature in the simplest possible way. The two theses are frequently treated together, as if the very fact that we can make a reasoned choice between conventions implies that the historical process by which conventions are determined will always select the most appropriate conventions.

But in fact, the convenience thesis by no means follows immediately from the claim that not all conventions are equal. Indeed, there is good reason to think that it is false, because it overlooks the fact that our conventions are not easily altered, and certainly do not change in step with our understanding of the laws of nature. The

[73] Einstein, *On the Electrodynamics of Moving Bodies*, 891-921.
[74] Poincaré, *Science and Method*.

[75] Poincaré, *Science and Method*.
[76] Ben-Menahem, *Conventionalism*.
[77] Ben-Menahem, *Conventionalism*.

conventions regarding the measurement of distance and time used by physicists like Lorentz and Poincaré were essentially the same conventions that had been used for thousands of years—yet there seems no reason to suppose that the conventions which permitted the simplest expression of the laws of nature as our ancestors understood them would still permit the simplest expression of the laws of nature as understood in 1900. The early history of special relativity provides a clear illustration of this problem: in retrospect, we can see that discarding certain conventions about the absoluteness of simultaneity allows us to attain greater simplicity both in the form of our fundamental equations and also in the structure of the explanations we are able to offer, but there is no way this could have been taken into account in the original formation of our conventions regarding space and time, because it is a consequence of features of electrodynamics such as the invariance of the speed of light, and this theory was completely unknown at the time the conventions were formed. Thus one conclusion to be drawn from this episode is that even if conventions are somehow selected in such a way as to be most convenient for the purposes of science at the time of their creation, it does not follow that the same conventions will still be the most convenient after science has had a chance to develop.

Moreover, Poincaré's adherence to the simplicity thesis was certainly one of the major factors that prevented him from seeing the possibility of using the relativity principle as *explanans* rather than *explanandum*. The convenience thesis is in itself no more than a historical claim, but it naturally leads to a number of normative claims about how scientists should behave with respect to matters which they believe to be conventional. If it is accepted that we naturally come to adopt the conventions which permit the simplest expression of the laws of nature, this provides a powerful motivation to retain the conventions that we currently have, and consequently, Poincaré

held that it is the role of science to work within the established framework of conventions rather than question that framework. In light of this view, it is entirely comprehensible that Einstein's approach did not occur to Poincaré. Einstein's success arose from the insight that by abandoning certain conventions we may achieve a great simplification in the structure of our explanations; but Poincaré believed such that such conventions had come to be selected precisely because they enabled us to give the simplest formulation of the laws of nature, so it is not surprising that he did not anticipate that rejecting a convention could lead to further simplifications.

3. Conclusions

Poincaré's contributions to the field of special relativity were undoubtedly invaluable, but nonetheless those contributions do not constitute an independent discovery of the theory. His conceptual grasp of certain elements, particularly the Lorentz transformations, was very different to Einstein's—and this is not merely a small difference in interpretation, but a substantive disagreement about the possible applications of the transformations. I have argued that Poincaré's choice to restrict his scientific explanations to traditional forms was related to certain conventionalist doctrines that he held, particularly the convenience thesis. This thesis led Poincaré to maintain that science should not question the conventional framework within which it is mandated to work, and as a consequence, he was unwilling to give up the conventional framework of explanation. Therefore an explanatory approach like Einstein's was unavailable to him, and he never appreciated the interesting possibilities that arise from a willingness to explain in non-traditional ways.

Emily Adlam
Oxford University

References:

Aristotle. *Physics*. English translation in Hardie, R. and Gaye, R. *Physica*. Oxford: The Clarendon Press,1930.

Barbour, J. *The Discovery of Dynamics*. Oxford: Oxford University Press, 2001.

Ben-Menahem, Y. *Conventionalism*. Cambridge: Cambridge University Press, 2006.

Brown, H. *Physical Relativity*. Oxford: Oxford University Press, 2005.

Darrigol. O. "The Mystery of the Einstein -Poincaré Connection." *Isis* 95 (2004): 614 -26.

Einstein. A. "On the Electrodynamics of Moving Bodies." *Annalen der Physik* 17 (1905): 891-921.

Goldberg, S. "Henri Poincare´and Einstein's Theory of Relativity." *American Journal of Physics* 35 (1967): 934-44

Hirosige. T. "The Ether Problem, the Mechanistic Worldview, and the Origins of the Theory of Relativity." *Historical Studies in the Physical Sciences* 7 (1976): 3-82.

Janssen, M. "A Comparison between Lorentz's Ether Theory and Special Relativity in the Light of the Experiments of Trouton and Noble." PhD diss., University of Pittsburgh, 1995.

Janssen, M. and Stachel, J. "The Optics and Electrodynamics of Moving Bodies." (2008).

Katzir, S. "Poincaré's Relativistic Physics: Its Origins and Nature." *Physics in Perspective* 7 (2005): 268-92

Lorentz, H. "Simplified Theory of Electrical and Optical Phenomena in Moving Systems." *Proc. Acad. Science Amsterdam* 1 (1899): 427-42.

---The Theory of Electrons. Leipzig and Berlin (1916).

Miller, A. "A Study of Henri Poincare's 'Sur la Dynamique de l'e lectron's'." *Archive for History of the Exact Sciences* 10 (1973): 320

Pais, A. *Subtle Is the Lord: The Science and the Life of Albert Einstein*. New York: Oxford University Press, 1982.

Poincaré, Henri. "Sur la dynamique de l'electron." *Rendiconti del Circolo matematico di Palermo* 21 (1906): 129-75. English translation in Logunov, A. (2001). On the Articles by Henri Poincaré 'On The Dynamics of the Electron'. *Joint Institute for Nuclear Research*, Dubna.

---"The Measure of Time." English translation in *The Foundations of Science (The Value of Science),* New York: Science Press (1913)

---"The Theory of Lorentz and the Principle of Reaction." *Archives neérlandaises des Sciences exactes et naturelles* 5 (1900): 252-78.

--- "L'electricite et optique: La lumiere et les theories electrodynamiques." Paris : Gauthier-Villars, 1901.

--- (1902). *Science and Method*, trans. Science and Method. London : Dover Publications Inc, 1952.

--- "L'etat et l'avenir de la Physique mathematique." *Bulletin des Sciences Mathematiques* 28 (1904): 302-24.

Stein, H. "Physics and Philosophy Meet: The Strange Case of Poincaré" University of Chicago, unpublished.

Zahar, E. *Einstein's Revolution*. Chicago: Open Court Company, 1989.

--- *Poincaré's Philosophy*. Chicago: Carus Publishing Company, 2001.

The Knowability Paradox: Solutions and Solution

Walker Page
Wheaton College

IN 1963 FREDERIC FITCH published a paper in which he analyzes various concepts such as desiring, believing, knowing, and others.[78] Although his discussion has been relatively unknown for much of its existence, it has recently become a matter of fairly vibrant debate, and is considered to be important for a variety of topics. In his analysis of knowing Fitch provides a reductio argument, which concludes that if all truths are knowable and there are some truths that are not known, then it follows that all truths are in fact known. In light of this conclusion, the argument has become particularly relevant in discussions of realism and anti-realism, and has been offered as a reason to reject verificationism. The argument is surprisingly simple, which is part of why it has been so controversial. As Timothy Williamson says at the start of his discussion of the paradox, "Perennial philosophers' hopes are unlikely victims of swift, natural deduction. Yet anti-realism has been thought one."[79]

The paradox is based on a proof of the following theorem–what Fitch calls 'Theorem 1', "If ⟨ is a truth class which is closed with respect to conjunction elimination[80], then the proposition, [p & ¬(αp)], which asserts that p is true but not a member of ⟨ (where p is any

[78] F. B. Fitch, "A Logical Analysis of Some Value Concepts", *Journal of Symbolic Logic* 28 (1963), pp. 135-142.
[79] Timothy Williamson, "Intuitionism Disproved?", *Analysis*, 42.4 (Oct. 1982), p. 203.
[80] I provide a description of what the term 'closed with respect to conjunction elimination' means below.

proposition), is itself necessarily not a member of ⟨ ."[81] In symbolic form the theorem is,

Theorem 1: $\vdash \neg \alpha\ [p\ \&\ \neg\ (\alpha p)]$

In this paper, I discuss what has come to be known as the Paradox of Knowability and the implications it has for various theories of truth. Most of the paper consists in a presentation of the paradox and an analysis of some ways in which individuals have tried to respond to it. At the end, I put forward my own possible solution that appeals to a pragmatic view of truth.

The paradox rests on six basic principles:[82]

Knowability Principle: (KP) $\forall p(p \supset \Diamond Kp)$[83]

This is simply the claim that all truths are knowable or possibly known, and they are only true if they are knowable.

Non-omniscient Principle: (NonO) $\exists p(p\ \&\ \neg Kp)$

This is the principle that not all truths are known, or more precisely, that there is at least one truth that is not known. Since neither any one person is omniscient, nor is the collection of all human knowers omniscient, it follows that there are at least some truths that are not known. This is a fairly intuitive and uncontroversial claim.

[81] Ibid., p. 138.

[82] My presentation of the paradox closely follows the presentation in the Stanford Encyclopedia of Philosophy article on Fitch's Paradox of Knowability. The names of the first two principles are those described in the Stanford Encyclopedia article. The names of the last four are my own.
Brogaard, Berit and Salerno, Joe, "Fitch's Paradox of Knowability", *The Stanford Encyclopedia of Philosophy (Fall 2012 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2012/entries/fitch-paradox/>.

[83] Let *K* be the epistemic operator, 'it is known by S at some time that.' And ◊ means 'it is possible that'. It is also important to note that the *K* operator must either refer to the same knower in all its uses throughout the proof, or to the same *closed* class of knowers. This prevents ambiguity from arising and is necessary for *K* to actually substitue for α.

Even if the collection of all human knowers could possibly be "omniscient," it remains obvious that this has not obtained. For example, I am fairly confident that no one does or ever will know the exact number of atoms that constitute this particular table that I'm using. Or again, it seems reasonable to think that no one knows the exact number of leaves on the tree outside, or the number of steps I have taken today, or the exact number of hairs on my head at the present moment.

The next two principles are fairly modest epistemic principles.

Conjunction Elimination Principle: (A) $K(p\ \&\ q) \Rightarrow Kp\ \&\ Kq$ where ⇒ means strict or necessary implication (*i.e.* $\Box\ (p \supset q)$).

This principle affirms that necessarily, if a conjunction of two propositions is known then each of the conjuncts is known, and the class of knowing is thereby, to use Fitch's terminology, closed with respect to conjunction elimination. This term, 'closed with respect to conjunction elimination,' refers to the distribution of the K operator over conjunctive statements.

Knowledge–Truth Principle: (B) $Kp \Rightarrow p$

This is simply the principle that necessarily, what is known is true. That is, knowledge entails truth. The next two principles are fairly modest modal principles.

Necessity Principle: (C) If $\vdash p$, then $\vdash \Box p$

This principle claims that all theorems are necessary, meaning that if a theorem is proven to be true in a certain system, then it is

necessarily true in that system.[84] For any proven theorem, it is proven that that theorem is necessary.

Impossibility Principle: (D) $\Box\neg p \Rightarrow \neg\Diamond p$

This is simply a translation principle in modal logic that claims that necessarily, if something is necessarily not true, then it is not possibly true.

From these six principles the *reductio* follows readily, and a problem arises for the Knowability and Non-omniscient principles. First, let us assume that (KP) and (NonO) are true. If (NonO) is true, then so is an instance of it, and if (KP) is true, then the instance of (NonO) can be plugged into (KP). Thus, the proof goes as follows:

(1) $p \,\&\, \neg Kp$ (instance of (NonO))

(2) $(p \,\&\, \neg Kp) \supset \Diamond K(p \,\&\, \neg Kp)$ (substitution of (1) for p in (KP))

By modus ponens it follows from (NonO) and (KP) that,

(3) $\Diamond K(p \,\&\, \neg Kp)$ (M.P., (1)–(2)).

(3) claims that it is possibly known that there is an unknown truth, and it clearly follows from (1)-(3) that this modal statement is true. If this is the case, then an instance of (3) must also be true. So it can be supposed that it is known that there is an unknown truth.

(4) $K(p \,\&\, \neg Kp)$ (instance of (3))

From the Conjunction Elimination Principle it follows that,

(5) $Kp \,\&\, K\neg Kp$ (from (4), by (A))

[84] "Add to propositional calculus a weak modal logic: necessary truths are true; if a conditional and its antecedent are both necessarily true, then so is its consequent; and permit inference of the necessitation of anything proved in this system." (Hart, 156)

Now, applying the Knowledge–Truth Principle to the right conjunct of (5) (which can be done by the simplification of (5)) it follows that,

(6) $Kp \,\&\, \neg Kp$ (from (5), by (B) applied to right conjunct)

Since (6) is a contradiction and it logically follows from (4), (4) must be rejected,

(7) $\neg K(p \,\&\, \neg Kp)$ (from (4)-(6), by *reductio*)

Having arrived at this conclusion, the Necessity Principle can be applied, and (7) can be considered to be necessarily true (in the system) since it has been proven in the system.

(8) $\Box\neg K(p \,\&\, \neg Kp)$ (from (7), by (C))

Finally, from this, along with the Impossibility principle, it is shown that,

(9) $\neg\Diamond K(p \,\&\, \neg Kp)$ (from (8), by (D))

Thus, it cannot be known that there is an unknown truth. That is, it cannot be known that there is a proposition that is both true and unknown. It might be possible that there is an unknown truth, but this cannot be known. This, however, contradicts (3), so a contradiction follows from an instance of (NonO) when it is plugged into (KP), and the advocate of the view that all truths are knowable must deny (NonO),

(10) $\neg\exists p(p \,\&\, \neg Kp)$,

from which it follows that all truths are known by someone at some time,

(11) $\forall p(p \supset Kp)$

This argument has been most troubling for truth theorists of an anti-realist stripe who accept the claim that all truths are knowable. For as the proof shows, accepting this claim, along with some fairly basic principles, leads one to accept that all truths are known by someone, which seems to be an unacceptable position since, as noted above, examples of unknown truths are many. It has been recognized as particularly problematic for the proponent of a strong form of verificationism. Hart takes this "strong form" to be verificationism proper and consequently considers it to be erroneous.[85] According to him, verificationism holds the following four theses: 'What is true is meaningful,' 'What is meaningful is verifiable,' 'What is verifiable can be known,' and from this it clearly follows by transitivity that 'What is true can be known.' Hart takes the knowability paradox at face value, and since he accepts (NonO), he thereby rejects the claim that all truths are knowable. Since the paradox follows from the claim that what is true can be known, and this claim, according to Hart follows inherently from verificationism, he thinks that verificationism should be rejected.

Many individuals have tried to defend this form of verificationism in their responses to the knowability paradox. They have provided a great variety of different methods for responding to it. In what follows I discuss some of those methods[86] and suggest my own possible response to the paradox.

[85] W. D. Hart, "The Epistemology of Abstract Objects", *Proceedings of the Aristotelian Society Supplementary Volume* (1979), p. 156. Note that it is Mackie (p. 91) who considers this to be a "strong form of verificationism." Ironically, Williamson (p. 203) calls this "weak anti-realism", though they probably mean these things in slightly different ways.

[86] The Stanford Encyclopedia of Philosophy article on Fitch's Paradox of Knowability is where most of these methods were made known to me. Most of the responses to them, however, are my own.

It is evident that the proof is logically valid according to the principles of classical logic. Some, however, have suggested that it is not valid according to intuitionistic logic.[87] One of the primary differences between classical and intuitionistic logic is that the latter rejects the law of excluded middle (LEM): $(A \vee \neg A)$. Consequently, intuitionists do not accept the principle that double negation can be eliminated (i.e. they do not accept that $(\neg \neg A \supset A)$), which is necessary to make the move from (10) to (11). As Williamson says, "The elimination of double negation is intuitionistically, a notorious failure" (p. 205). The reason for this is that if they did accept the principle of double negation, the law of excluded middle would follow readily from the intuitionistically provable $\neg \neg (A \vee \neg A)$.[88] Because of this rejection of double negation, the proof must end at (10) for the intuitionist, but she must still be committed to (10).[89]

Some intuitionists have argued that given this alternative conclusion the anti-realist need not be reduced to the absurdity of (11), but rather has reason for accepting intuitionistic logic. The fact that intuitionistic logic seems to (in some sense) avoid the paradox has been put forward as a reason for accepting intutionistic logic rather than classical. The problem with this is that the debate between the classical and the intuitionistic logician is based on things that are independent of the paradox (e.g. the acceptance or rejection of LEM). The conclusion might not be paradoxical when intuitionistic logic is used, but it is paradoxical when classical logic is used, and so the

[87] For example, see Williamson, "Intuitionism Disproved?"

[88] This is explained in the Stanford Encyclopedia of Philosophy article on Intuitionistic Logic. See Moschovakis, Joan, "Intuitionistic Logic", *The Stanford Encyclopedia of Philosophy (Summer 2010 Edition)*, Edward N. Zalta (ed.)

[89] In light of this, some have objected that the weaker claim made in (10) that there are no unknown truths, still seems nearly as troublesome as the claim made in (11) that all truths are known, and so the intuitionist does not provide much solace. See Brogard for more on this.

paradox cannot be used as a reason for accepting intutionistic logic. The fact that the paradox results when using classical logic might provide an incentive for accepting intuitionistic logic, but it cannot be considered a sufficient reason or justification for accepting it. The classical logician will not be convinced by the intuitionist; for according to her (11) is simply entailed by (10), and the paradox still follows. So it seems that there must be independent reasons provided for accepting intuitionistic logic rather than classical. The paradox itself cannot be used as a reason for accepting one over the other.

Another attempt to show that Fitch's proof is inadequate has been to claim that (6) is not problematic as the paradox requires. (6) is supposed to be the contradiction that results, from which one can infer (7). But some say that a paraconsistent logic should be used when talking about knowledge. Paraconsistent logics challenge the inference relation that anything can be considered a logical consequence of a proven contradiction (i.e. they challenge this inference relation: $\{A, \neg A\} \vDash B$ for every A and B). These have been used to support dialetheism, which claims that there are propositions that are both true and false; there is a proposition such that both it and its negation are true.[90] In response to the knowability paradox these concepts have been used to support the claim that there are in fact cases when something is both known and unknown.[91] (6) is not impossible because there are cases in which knowledge is inconsistent, and if (6) is not impossible, then the absurd conclusion of the paradox does not follow.

[90] For more on Paraconsistent logic and Dialetheism, see the articles on each of these topics in the *Stanford Encyclopedia of Philosophy*. This is where my brief discussion comes from.

[91] As stated above, paraconsistent logics do not accept the claim that just any inference can be made from a derived contradiction. And dialetheists accept that there are contradictory propositions.

Beall applies this method and appeals to what has been called the knower paradox,[92]

(k): The proposition *k* is unknown.

He thinks that this is a case in which knowledge is inconsistent and that (6) turns out to be true for (*k*). He tries to show this by arguing that whether (*k*) is known or not known, it turns out that it is both known and not known, and thereby provides an instance where (6) should be accepted. His reasoning goes as follows. Let us first suppose that (*k*) is known. If (*k*) is known, then by the Knowledge-Truth principle, it follows that (*k*) is true. But if (*k*) is true, then it is true that (*k*) is unknown since this is precisely the claim that is made in (*k*). Thus, one must conclude that (*k*) is both known and unknown.

Now let us suppose that (*k*) is unknown. If (*k*) is unknown, then (k) is true, in which case we can know that (*k*) is true. But if this is so, then one must once again acknowledge that (*k*) can be both known and unknown. Thus, it seems that whether (k) is known or unknown it turns out that we have found an instance in which (6) is true.[93]

The main problem with Beall's argument is that it does not even get off the ground without first presuming that (*k*) actually refers to a proposition, and can thereby be an object of knowledge. Without this, the knower paradox can hardly count as an instance of (6). For if (*k*) does not correspond to a proposition, then it clearly cannot be an object of knowledge, in which case it does not count as an example of (6). But in order to make the assumption that (*k*) corresponds to a proposition, one must first assume the possibility of (6), which is precisely what is denied in the paradox. As Beall himself says, "Fitch's proof… assumes that for no claim *p* or world w do we have

[92] Beall, JC, 2000. "Fitch's Proof, Verificationism, and the Knower Paradox." *Australasian Journal of Philosophy* 78, pp. 242-247.

[93] This closely mirrors Beall's own presentation of the argument, see p. 243.

Kp and ¬*Kp*."[94] Now if Fitch's proof assumes the impossibility of (6), and Beall's argument concerning the knower paradox assumes its possibility, then Beall's argument clearly cannot be used against the knowability paradox. In order for this response to work, Beall must provide independent evidence for the possibility of (6), which he does not do. Until this is provided, there is no reason to think that this is a legitimate response to the paradox.

Another possible response to the paradox might be found in the affirmation of theism, or more specifically, in the affirmation of the existence of an omniscient being. If one holds this position, it seems that one can–and must–reject (NonO), without which the paradox does not work. For it is clear that if one accepts the existence of someone that knows all truths, then it cannot be the case that there is a truth that nobody knows.

It is somewhat surprising that this solution has not been put forward before (at least not to my knowledge). But perhaps the reason for this is that there is a conspicuous flaw for this response. Obviously it will not be very helpful for individuals who reject the existence of an omniscient being, but the problem goes beyond this. The problem is that this does not even seem to be a helpful solution for the theist who does accept the existence of such a being. The epistemic operator K can be restricted in a way that excludes an omniscient being. The solution only works if the epistemic operator K means 'it is known by S at some time that,' and 'S' is construed broadly enough to allow for non-human rational beings. But K can be restricted to a much narrower class–for example, to mean 'it is known by S_h at some time that' where 'S_h' means 'some human'– and the paradox still follows. This becomes clear if one uses this phrase in (4),

[94] Beall, p. 244

(4*) It is known by S_h that both *p* and it is not known by S_h that *p*.

This is clearly contradictory, and it can be seen that the conclusion of the paradox still follows even with this minor alteration to the epistemic operator. So one can accept the existence of an omniscient being but still be faced with the paradox of knowability. The only thing that the paradox requires, at least in order for it to be relevant for us, is that there be no omniscient humans. And this is clearly true.

This points us in the direction of another solution that some have put forward.[95] It has been supposed that the real problem with the paradox is that it requires an epistemic attitude that is too strong. The truth-entailing property that is included in the Knowledge-Truth Principle is too strong, and a weaker view that does not depend on facticity should be considered. The proponent of this solution thinks that the epistemic operator should be changed to something like 'S has evidence at some time that,' or 'S justifiably believes at some time that,' or 'it is verified by S at some time that,' or even something as weak as 'it is believed by S at some time that.' Each of these is much weaker than the original form of K, 'it is known by S at some time that,' and it is thought that this might prevent the argument from going through.

The problem is that when these other options are plugged into the argument, even these weaker versions of K result in a paradoxical conclusion very similar to the original.[96] This can be seen at the very first premise of the argument. If we plug the first option into (4) we get,

[95] This method is discussed in Mackie (pp. 91-92) and in Edgington (pp. 558-559).

[96] Mackie points out that the argument only goes through if the epistemic operator is indexed to some specific time t_1, instead of 'at some time.' This is a legitimate point, and seems to be necessary for any formulation of the argument. This, however, doesn't seem to make the paradox any less problematic. In most of my discussion I am simply ignoring the dimension of time because it seems that the paradox can be formulated in terms of a specific time t_1 and still be just as troublesome.

(4**) It is justifiably believed by S that both *p* and it is not justifiably believed by S that *p*,

which is clearly inconsistent. And the same problem results when plugging any of the other options into (4). Thus, it does not seem that an epistemic principle that does not entail truth makes much difference.

At this juncture it appears that most of the arguments that try to defend the verificationist against the paradox of knowability are at best up for debate and at worst completely unsuccessful.[97] But does this mean that anti-realism should be rejected? It seems not; for there are anti-realist positions that are not verificationist, and at this point I would like to suggest a solution to the paradox that appeals to a pragmatic view of truth. Suppose that one views the K operator in terms of consensus, and truth is whatever proposition there is consensus about at the end of all inquiry.[98] Is such a view susceptible to Fitch's paradox? It appears not for the following reason.

On such a pragmatic view the epistemic operator–Let us call it K*– would be construed as, 'There is consensus among G at some time', where 'G' means 'some group'. Given this, it is clear that the Knowledge-Truth Principle does not always work since there are indeed times when there is consensus that p, but it is not the case that p. Not only this, but if the epistemic operator is construed in this way the contradiction in (6), which is necessary for the argument to work, cannot actually be derived. This is clear because when one plugs K* in at (4), there is no contradiction,

[97] There are some arguments that I have not discussed here that might still be able to salvage the verificationist view from the apparent death blow of the paradox. For example, some have wanted to try and restrict the universal operator in the Knowability Principle in such a way that the paradoxical conclusion does not follow. For example, see Dummett, 2001.

[98] Something like this can be found in Charles Peirce, "The Fixation of Belief", and "How to Make Our Ideas Clear".

(4***) There is consensus among G that both *p* and there is not consensus among G that *p*.

The reason there is no contradiction is because it is possible that there be consensus among some group that *p*, and that there be consensus that there is not consensus that *p*.

Let us consider a concrete example to make this clearer. Suppose that there is a group of four very distrusting individuals who have been brought together to vote on some particular issue, say whether or not skim milk is healthy. It turns out that they all agree that it is healthy, and each person expresses this to the other individuals. It is clear that there is consensus that skim milk is healthy since they all agree that it is. But suppose that each person thinks the others are lying when they claim to think skim milk is healthy. Because of this, each person thinks that he or she is the only one that really thinks skim milk is healthy. It follows from this that each person agrees that the others do not think that skim milk is healthy. That is, there is consensus that there is not consensus that skim milk is healthy. Thus, there is consensus among the group concerning two things: both that skim milk is healthy and that there is not consensus among the group that skim milk is healthy.

This solution only works if 'consensus' is understood in a particular way. Consensus must be viewed only to mean that each member of a group believes that something is true.[99] If one views 'consensus' in a way that requires both that each member of the group believes something to be true and that they are each aware that each person believes it to be true, then the paradox still results. But this latter view of consensus does not seem very plausible. There

[99] There might be a weaker view of consensus that is acceptable, namely, that only the majority of people in the group must think something to be true, but I will not pursue this possibility here.

are clearly cases when we say there is consensus, but the individuals are not aware that everyone in the group thinks something is true. For example, we can imagine a research survey in which individuals are asked to respond to a single yes or no question, and it turns out that everyone puts yes. Even if the researchers never reveal to the individuals that everyone put yes, the researchers can clearly still say that there is consensus among the group of people who answered the question.

Thus it appears that a pragmatic approach can provide an outlet for the survival of the anti-realist. There is, however, one more solution–which might not be considered a solution at all–that I have not considered in detail. This was hinted at earlier in the discussion of Hart. One might follow Hart, simply accept the paradox at face value, reject anti-realism, and accept realism. However, the anti-realist who has fervently sought to respond to Fitch's paradox has probably done so not only because she wants to defend anti-realism, but also because she finds realism unacceptable for independent reasons. If this is the case, then I have provided another outlet, namely the pragmatic approach, by which the anti-realist can rest in her peaceful slumber. Of course, it might be objected that there are independent reasons for rejecting a pragmatic view, and this would indeed be deleterious for the anti-realist. The point of this paper, however, is to show that, ceteris paribus, there are at least two options for the individual who is faced with Fitch's paradox, one that is realist and one that is anti-realist. Whether or not one of these should be preferred is a discussion for another time.

Walker Page
Wheaton College

References:

Beall, JC, 2000. "Fitch's Proof, Verificationism, and the Knower Paradox." Australasian Journal of Philosophy 78, pp. 242-247.

Brogaard, Berit and Salerno, Joe, "Fitch's Paradox of Knowability", The Stanford Encyclopedia of Philosophy (Fall 2012 Edition), Edward N. Zalta (ed.)

Dummet, Michael, 2001. "Victor's Error." Analysis 61, pp. 1-2.

Edgington, Dorothy, 1985. "The Paradox of Knowability." Mind 94, pp. 557-568.

Fitch, Frederic, 1963. "A Logical Analysis of Some Value Concepts." Journal of Symbolic Logic 28, pp. 135-142.

Hart, W. D., 1979. "The Epistemology of Abstract Objects: Access and Inference." Proceedings of the Aristotelian Society supplementary, 53, pp. 153-165.

Mackie, John L., 1980. "Truth and Knowability." Analysis 40, pp. 90-92.

Moschovakis, Joan, "Intuitionistic Logic", The Stanford Encyclopedia of Philosophy (Summer 2010 Edition), Edward N. Zalta (ed.)

Priest, Graham and Berto, Francesco, "Dialetheism", The Stanford Encyclopedia of Philosophy (Summer 2010 Edition), Edward N. Zalta (ed.)

Priest, Graham and Tanaka, Koji, "Paraconsistent Logic", The Stanford Encyclopedia of Philosophy (Summer 2009 Edition), Edward N. Zalta (ed.)

Williamson, Timothy, 1982. "Intuitionism Disproved?" Analysis 42, pp. 203-207.

What Functional Reductionism Means for Normative Epistemology

Alexander Agnello
Concordia University

IN HIS BOOK TITLED *Physicalism, or Something Near Enough*, Jaegwon Kim takes a position that compliments David Chalmers' position by demonstrating how "psychological states" (intentional mental states) are functionally reducible to a physical base (2005, 161). By contrast, Kim characterizes what Chalmers calls "phenomenal states" (qualia) as the subjective properties of experience that resist this reduction (2005, 161). In order to illustrate what it means to functionally reduce a property, Kim (2005) refers to how a gene is reduced in molecular biology to the DNA molecules that carry out its causal task of transmitting heritable traits. He argues that "when all this is in, we can say that the gene has been physically reduced, and that we now have a reductive explanation of how the process of heredity works at the molecular level" (Kim 2005, p. 163). The objective of this paper is two-fold: first, I will provide reasons to doubt the plausibility of this process, and second, I will argue that this process is incompatible with Kim's defense of normative epistemology. The first part of the essay provides reasons to suggest, counter to Kim's arguments, that intentional mental properties are functionally irreducible. The argument I present hinges on the noted problems with reducing qualia to functional or physical properties. The second part considers the difficulties with reconciling functional reductionism with Kim's defense of normative epistemology. I argue that functional reductionism disregards the primacy of normative standards for belief formation and is more in keeping with the concept of naturalized epistemology W.V.O. Quine proposes.

Kim (2005) explains the process of functional reductionism in three steps. The first is the conceptual step of interpreting a property, functionally defining it through an analysis of the causal work that it is intended to perform (Kim 2005). Once this is determined, the scientific method is used to discover the physical realizers that are responsible for performing the specific causal work (Kim 2005). The final step involves developing an explanation at the physical level of how these mechanisms execute their intended functions (Kim 2005). Kim (2005) claims that if the first step can be successfully completed, then it is shown that the property in question is theoretically reducible.

In determining which mental properties can be reduced to a physical base, Kim appeals to various considerations made to discuss how intentional properties can be characterized by the causal work they perform in the "overall economy of human behavior" (2005, 165). He provides the example of an indigenous tribe who communicates and demonstrates behavioural patterns that we, as beings with beliefs and intentions, can identify with (Kim 2005). We infer through perceived patterns of behaviour that these beings are also capable of formulating beliefs and intentions since they employ methods for communication that conform to the standards of rationality (Kim 2005). In arriving at this conclusion, Kim is convinced that one cannot avoid thinking of intentional mental states like thought, belief, and desire as supervening on observable facts (2005, 166). He believes that such an example provides a powerful reason for thinking that these properties can be defined and interpreted through their roles in behavioural causation (Kim 2005).

By contrast, qualitative states of consciousness are metaphysically irreducible to physical property because their role in the causal network cannot be identified through observable behaviour. Kim (2005) points to the inverted spectrum thought experiment as evi-

dence that the causal work of qualia does not define, functionalize, or relate to the phenomenal state. The experiment implies that an inverted spectrum would go undetected because a person would react the same way to stimuli (describe the car as blue) even if her experiences have qualia that are normally produced in others by yellow objects (Byrne 2004). This insight leads Kim (2005) to conclude that qualia cannot be functionally defined because we encounter an "explanatory gap" in analyzing how physical properties give rise to qualitative states (Levine 1983, p. 357).

The second, and perhaps more controversial claim Kim makes, is that there is nothing beyond the causal work related to the observable behaviour that is involved in the generation of a belief. Kim writes:

beliefs can generate further beliefs; in conjunction with desires, they can cause further desires; and so on. However, these further mental states, too must ultimately be anchored, conceptually and epistemologically, in observable behavior … [a]s far as intentional states are concerned, we are within the domain of behavior and the physical mechanisms involved in their production; they do not take us outside this domain (2005, 167).

In considering the causal work that is performed by underlying realizers, or what he calls "the causal mechanisms that ground belief", Kim argues that an interpretation of these components does not take our analysis beyond the physical domain (2005, 167). This idea can be captured by the strong causal closure principle[100]: "no causal chain will ever cross the boundary between the physical and the

[100] The weak formulation of causal closure states that any event occurring at *t* has a physical cause occurring at *t* (Kim, 2005, p.43).

non-physical" (Kim 1998, 40). Functional reductionism's adherence to this principle allows Kim to respond to the causal exclusion problem purportedly engendered by non- reductionist forms of physicalism. This problem arises when mental property M and physical property P are both regarded as distinct and sufficient causes for event P* at time t (Kim 1998). This is considered a causally overdetermined system since either of these causes could generate

event P*, and thus, only one of them is required in order to explain how P* occurs (Kim 2005). Prima facie, overdetermination seems improbable in that it implies that these two distinct causes consistently conspire to bring about events. Furthermore, Kim argues that in order for overdetermination to be considered a viable option, we need to be able to refer to cases where the mental cause occurs without the physical cause (2005, 46-9). Given that science is adept at identifying the physical laws governing cause and effect, Kim eliminates the causal exclusion problem along with the possibility for overdetermination by embracing reductive physicalism. The supposed issue with non-reductionist physicalist theories like Donald Davidson's (1970) anomalous monism, which sees mental events as being identical with, but not nomologically reducible to physical events, is that it renders mental properties causally irrelevant (Phillips 1995).

Even if we accede to Kim that functional reductionism is capable of rescuing intentional mental states from the threat of epiphenomenalism, there are still reasons to suggest that one cannot neatly interpret all the components of belief through observable behaviour and its physical realizers. One might argue that functional reductionism faces a problem that is analogous to the explanatory gap when it attempts to functionally define intentional mental properties. While identified patterns of behaviour can allow us to deduce that an individual formulates beliefs, there is doubt as to whether the causal

work can provide us with the information that is needed to attribute any particular belief to an event. When considering the generation of belief, one could conceive of several reasons for why individual X decided to do action Y in the circumstances. Two physical events that are analogous, like two students raising their hand after a professor has asked a question, can be produced by two separate beliefs: the first student has a genuine answer for her professor's question while the second absentmindedly raises his arm because he believes it will help relieve the pain in his shoulder. The issue is not whether physical realizers are foundational to belief attribution. The observed behaviour that was described in Kim's example with the indigenous people, which involved group communication through a common language, can be reasonably explained by attributing belief to these individuals. Rather, the issue is whether functional reductionism is able to account for the causal work that allows us to identify the intentional content of mental states. Observers will deduce belief X on the part of individual Y while remaining unaware (despite the evidence available) of how she is affected by the phenomena. In the same way, a functional interpretation of intentional mental states can allow us to deduce a belief X on the part of individual Z, but this may ultimately fall short if observable behaviour does not provide us with the information required to discern the thing believed. In such a case, a theory about the content of a belief is underdetermined by the evidence[101]. So while qualia cannot be functionally defined because none of its physical realizations are constitutive of the experience, beliefs are also not functionalized in the way Kim hypothesizes

[101] This argument is underscored by the undetermination of scientific theory by evidence. The idea of underdetermination was discussed by Pierre Duhem (1954) with respect to how the available evidence may be insufficient for confirming theories in physics.

if the thing believed cannot be identified with any particular behavioural state.

A possible way to account for the difficulty associated with identifying the object of belief through the causal work is to attribute the problem to an epistemic gap as opposed to an in- principle problem with functional reductionism. Kim (2005) openly admits that logical behaviourists and functionalists have come up short in producing concise functional definitions for complex intentional mental states. However, Kim (2005) also notes that a partial functional analysis of these properties can still provide a point of departure for developing a better understanding of the associated biological realizers. He concludes that it is unnecessary to know everything that belief does before we uncover the possible neural mechanisms; "a partial list will be enough to start us off" (Kim 2005, 167). It might be argued that while this epistemic gap exposes areas that are in need of further development, it does not, on its own, prove the theory false. However, issues have been raised regarding whether this interpretative project is possible on a theoretical level. If we cannot locate the intentional content of belief in observable behaviour, we then face the difficulties that are encountered and deemed irresolvable by Kim in his attempts to functionally interpret phenomenal properties. If what I propose is correct, then like quale, intentional mental states can be related to a number of physical events, none of which relate exclusively to, or are constitutive of, that content.

In a seminal paper, "Epistemology Naturalized", Quine (1969) argues that epistemology should centre on a study of the causal connection between sense experience and belief attribution. Consequently, epistemology would shift its focus away from the convention of outlining the normative concepts that regulate belief formation. In reply, Kim argues that this transformation cannot be justified because "the two disciplines do not investigate the same relation"

(1993, 227). Normative epistemology aims to define the justificatory relationship between our beliefs and evidence (Kim, 1993). Instead, Quine's naturalized concept of epistemology aims to emulate the natural sciences in its methods by providing a causal account of knowledge through an empirical study of how theory relates to the physical world (Kim 1993). On Kim's (1993) view, one cannot disregard the role that epistemic intuitions play in providing theorists with a standard to assess counterfactual claims and examples; without this, the concept of belief would be unintelligible. In "What is Naturalized Epistemology", he argues: "[i]t is not merely that belief attribution requires the umbrella assumption about the overall rationality of cognizers ... [i]t requires belief evaluation, in accordance with normative standards of evidence and justification" (Kim 1993, 229). However, it can be argued that these conditions are unacknowledged by the functional reduction of intentional mental properties. At the preliminary stage of the process, we infer that the indigenous people formulate beliefs based on their rational behaviour. But ultimately, functional reductionism's goal is to provide a causal explanation of how beliefs are produced by scientifically respectable properties. A gene is physically reduced in molecular biology to its causal function as a transmitter of heritable characteristics. The belief one attributes to the members of the tribe is essentially defined by the role it plays in behavioural causation. Thus, it is unclear why Kim argues that belief attribution cannot be understood without the normative standard of justification. Functional reduction affirms that the only processes which are causally responsible for belief attribution are those that can be identified and studied through naturalistic resources.

It might be proposed that Kim could gain consistency between functional reductionism and normative epistemology by appealing to his position on epistemic supervenience. The idea is that Kim's account can provide a natural criterion for epistemic concepts without committing the naturalistic fallacy introduced by G. E. Moore. To commit the naturalistic fallacy is to assume "that because some quality or combination of qualities invariably and necessarily accompanies the quality of goodness, or is invariably and necessarily accompanied by it, or both, this quality or combination of qualities is identical with goodness" (Prior 1949, 1). So while we may conclude that certain actions are good by virtue of certain natural qualities that those actions possess, it is not to say that the possession of those natural qualities renders the action identical to the good. The "Good" cannot be encapsulated by any instance in nature; it is, by definition, a non-natural property. What is being suggested here is that Kim's use of epistemic supervenience does the same work for normative judgements. Kim states:

> [b]eing a good car, say, cannot be a brute and ultimate fact: a car is good because it has a certain contextually indicated set of properties … The same goes for justified belief: if a belief is justified, that must be so because it has certain factual, non-epistemic properties, such as perhaps that it is "indubitable", that it is seen to be entailed by another belief that is independently justified, that it is appropriately caused by perceptual experience, or whatever (1993, 235).

An instance of justified belief, "I believe Henry owns a Ford", is not equivalent to the non- natural property of justification. There may be natural properties that make this belief justified (the fact that he has shown me his license and registration, the fact that I see him at the dealership for routine maintenance), yet these natural properties do not constitute justification itself. Through such an explanation, Kim could maintain that normative principles are separate from the physical realm in the sense that they are not based in any particu-

lar natural instantiation. However, Kim cannot revert to the doctrine of epistemic supervenience to explain how an epistemic property like justification resists physical reduction. Kim's formulation of the principle of causal closure eliminates any properties from the causal network that do not have a physical realizer as its ultimate cause. This principle excludes the possibility of having natural properties that relate, but do not equate, to non-natural properties in the causal network without them existing as epiphenomenal properties.

I argued that functional interpretations of intentional mental properties may be underdetermined by the causal work associated to those properties. This is because the intentional content of a belief cannot be identified through an analysis of observable events. Moreover, functional reductionism's reliance on the natural sciences to examine the causal relations between mental properties and their physical realizers undermines the overall importance of normative standards for belief attribution. This inconsistency between normative epistemology and functional reductionism's methods has also raised an ontological problem: the normative concept of justification cannot exist in the causal network as a non-natural property. If Kim responds to this claim by reverting to the doctrine of epistemic supervenience, I have argued that he must be willing to accept the epiphenomenalism of intentional mental properties.

Alexander Agnello
Concordia University

References:

Byrne, Alex. 2004, November 10. Inverted qualia. *Stanford Encyclopedia of Philosophy*. Retrieved November 2, 2012, from http://plato.stanford.edu/entries/qualia-inverted/

Davidson, Donald. 1970. *Actions and Events*. Oxford: Clarendon Press, 1980.

Duhem, Pierre. (1914). 1954. *The Aim and Structure of Physical Theory*, trans. from 2nd ed. by P. W. Wiener; originally published as *La Théorie Physique: Son Objet et sa Structure* (Paris: Marcel Riviera & Cie.), Princeton, NJ: Princeton University Press.

Kim, Jaegwon. 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem & Mental Causation*. Cambridge: MIT.

——. 2005. *Physicalism, or Something Near Enough*. Princeton, N.J.: Princeton University Press.

——. 1993. *Supervenience and Mind: Selected Philosophical Essays*. New York, NY, USA: Cambridge University Press.

Levine, Joseph. 1983. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64: 354-361.

Phillips, E. Hollibert. 1995. *Vicissitudes of the I: An Introduction to the Philosophy of Mind*. Englewood Cliffs, NJ: Prentice-Hall.

Prior, Arthur. 1949. *Logic and the Basis of Ethics*. Oxford: Clarendon.

Quine, W.V.O. 1969. *Ontological Relativity and Other Essays*. New York: Columbia University Press