Volume 8, Number 1
2014

ISSN 1807-9792

# abstracta

## Linguagem, Mente & Ação

http://abstracta.oa.hhu.de

d|u|p

abstracta
*Linguagem, Mente & Ação*

ISSN 1807-9792

Volume 8, Number 1
2014

abstracta
2004 – 2014

# Contents

**Discussion**

abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 8, Number 1, pp. 3–14, 2014

# The Regulative and the Theoretical in Epistemology[1]

Robert Lockie

Psychology Dept., University of West London,
Paragon House, Boston Manor Rd. Brentford, Middx TW8 9GA
bob.lockie@uwl.ac.uk

**Abstract**

The distinction between the regulative ('practical', 'subjective', 'decision-procedural') and the theoretical ('objective', 'absolute') pertains to the aims (the desiderata) of an account of justification. This distinction began in ethics and spread to epistemology. Each of internalism, externalism, is separately forced to draw this distinction to avoid a stock, otherwise fatal, argument levelled against them by the other. Given this situation however, we may finesse much partisan conflict in epistemology by simply seeing differing accounts of justification as answering to radically distinct desiderata of adequacy. We should see knowledge as answering to the theoretical desideratum of adequacy alone; and rationality as answering to the regulative desideratum of adequacy alone. Objections to this 'Gordian' [knot] approach to epistemology (from virtues theorists and others) are rejected. Such an approach may make for accounts that violate our ordinary language intuitions; but in developing an epistemological axiology, any such intuitions are not to the point.

## 1 Three Distinctions

This paper concerns a distinction between different desiderata for an account in normative epistemology. Before proceeding to this, our eponymous distinction, we need to flag two prior distinctions. One distinction is between the things we are seeking to account for. The things we are most concerned to account for are rationality and knowledge. A second distinction concerns the different theories (better, theory-families) that are candidates to account for these: internalist accounts, externalist accounts and (arguably) virtue epistemic accounts – though the status of the latter as a 'third force', genuinely in competition with accounts grouped under the internalism-externalism distinction remains controversial (Lockie 2008).

The internalism-externalism distinction is of course the subject of major, extended, debate. In this paper, stipulatively, I take internalist theories to be epistemically deontological theories: theories which interpret epistemic normativity on the model of reasoning *dutifully*, as one *ought* – as discharging one's intellectual *responsibilities*. This understanding of internalism (and hence of the internalism-externalism distinction) needs to be distinguished from a more purely 'accessibilist', and again a more 'mentalist' conception of this distinction[2].

[1] An earlier version of this paper was presented at the Joint Session of the Mind Association & Aristotelian Society, Stirling, and at the Spanish Society for Analytic Philosophy (SEFA), Tenerife. My thanks to audiences at these forums and to Tony Booth for comments on a previous, written, version of this paper.

[2] This understanding of internalism (hence the internalist-externalist distinction) corresponds roughly to what Bergmann (2006) calls 'subjective deontological justification' or 'epistemic blamelessness' – a notion Bergmann correctly identifies in Plantinga, Foley and Alston. Alston (1985) abbreviates this notion as '$J_{di}$' – which stands for *deontic, internalist, justification*. Plantinga just calls this same notion 'internalism', but when pushed, "classical deon-

This paper concerns a third distinction found in normative epistemology – between the regulative and the theoretical. This distinction has been well articulated over the years; yet a contention of this paper is that its significance is still not fully appreciated. It is powerfully similar – arguably identical – to a distinction drawn in ethics. We uncover this third distinction via a consideration of the arguments levelled by internalists (deontologists) against externalists (consequentialists) and vice versa. Of course the protagonists in these debates see them as being directed towards establishing which of internalism, externalism, is the true theory of knowledge or rationality; however, for us, uncovering these arguments will be in the service of another end.

## 2 Why Internalists and Deontologists draw this Third Distinction: The Epistemic Poverty Objection[3]

The deontic conception of internalism (henceforth and throughout: just 'internalism') involves the idea of *cognitive accessibility*, of *epistemic deontology*, and of these being connected via an *'ought' implies 'can'* principle[4]. These may be used to create a problem for internalism. I may have done all I can epistemically, dutifully discharging my intellectual obligations to the limit of my abilities. Still, I may be desperately far removed from either the truth or an objectively truth-conducive basis for my belief. A classic source for this objection would be Alston (1985). Alston gives examples of dutiful but helplessly flawed cognizers, examples typical of many in the externalist literature: A tribesman may have been brought up to accept the traditions of his tribe as authoritative, and never have seen anything to call these traditions (of inquiry, etc.) into question. A person may be intellectually honest and diligent, but just rather dim; or not dim especially, but highly impressionable. A person may lack an education, being vulnerable to all kinds of unfounded hearsay and superstition as a result. In general, one may have discharged one's epistemic duties as responsibly as one is able, but still (blamelessly) be holding one's beliefs on profoundly inadequate grounds. Thus, it is argued, the deontic conception of internalism is an inadequate basis for epistemology.

This 'epistemic poverty' objection originated as an argument against ethical deontology. In both ethics and epistemology it has a stock response. This is to draw a distinction between *objective* and *subjective* duty[5]. One is culpable, blameworthy, irresponsible, should one fail to discharge one's subjective duty (doing what one has reason to believe will bring about the Right); one is not blameworthy, irresponsible, merely in virtue of failing to discharge one's objective duty (actually maximising the Good) – which failure may be quite out of one's hands. Owens (2000) notes this distinction in Sidgwick (*c.f.* 1907: 413). Plantinga

tological internalism" – and deprecates those pure accessibilist internalists who depart from what he (an externalist) nevertheless identifies as its "deep integrity" (Plantinga 1993: 28). I would endorse the account found in the first chapter of Plantinga (1993) as still the best single overview of this conception of the internalism-externalism distinction, and as glossing the understanding of these terms that I am operating with. For the understanding of the virtues position that is operative in this paper see Lockie (2008).

[3] The phrase 'epistemic poverty' is owed to Bonjour (2003: 176).

[4] The principle that 'ought' implies 'can' is in no sense proprietary to internalism; at least as great reliance is placed on it by most externalists, who argue by contraposition that since our levels of control, access, and freedom are greatly less than would be required for us to be held responsible for our beliefs, we must therefore abandon any conception of epistemic value which sees this as the discharge of intellectual duties (e. g. the 'doxastic voluntarism' debates). There are those who reject 'ought' implies 'can' in epistemology (for example, Ryan 2003, Bergmann 2006, Hieronymi 2008 and Owens 2000) but these exceptions are not common.

[5] Objective and subjective *duty* because we are here dealing with a response to an objection to ethical/epistemic deontology; shortly we will generalise this response to a distinction between the subjective and objective without any restriction to duties. I note this because there are clear problems with a (strong) notion of *objective* duty – at least for one who embraces 'ought' implies 'can' and seeks after more than just an account of the Good.

(1993) credits his version of this distinction to Aquinas; though a more immediate source might well be Chrisholm (1957) who draws the same distinction using the terms 'practical' for *subjective* and 'absolute' for *objective* – himself crediting Richard Price. To escape the epistemic poverty objection, deontic, oughts-based justification must be restricted in its application to the subjective, practical realm. There is another, objective, absolute, sense of being justified for which the discharge of duty, the fulfilment of obligations, be we ever so diligent, is not guaranteed to satisfy. Consider, in light of this, a claim such as the following:

> I shall assume that only *right* epistemic rules make a difference to genuine justifiedness. This point should be equally acceptable to both internalism and externalism (Goldman 2009: 5–6).

This point will be "equally acceptable to both internalism and externalism" only should it be read by each under a different interpretation of "right epistemic rule". For the internalist, this means *subjectively* right; for the externalist, *objectively* right[6]. Argument at cross purposes beckons if we do not keep this in mind.

## 3 Why Externalists and Consequentialists draw this Third Distinction: The DDP Objection

In ethics there are a set of stock objections to consequentialist theories – e. g. Act Utilitarianism – and in turn a stock response to this set of objections (to draw the regulative-theoretical distinction). The objection and associated response carried over into epistemology. Brink (1986) called this family of objections in ethics generally, the objection from the 'personal point of view'. There are several such objections that do not concern us; but a version which does – found in each of Bales (1971), Brink (1986), Smith (1988) and *passim* in the literature – is as follows: Working through the act-utilitarian (or other) consequences of even a simple choice of action is likely to be a highly involved matter. Bring other choices into the equation, factor in a diachronic time-scale, incorporate a need to take on new information in real time, and the matter becomes massively more involved. The process of calculating these consequences – to even a modest level of surety – takes time. Frequently, for the agent to embark on the process of calculating the consequences of a course of action, will itself be to choose one way or other *how* to act (and often, to choose wrongly)[7].

This objection to consequentialism in ethics carried over directly into epistemology, where it was levelled by internalists against externalism. One traditional ambition of epistemology

[6] Goldman subsequently considers the application of his interpretation of this principle to a specimen "rightness criterion" from the internalist camp – and he chooses as his specimen Richard Foley. He objects that the subjectivism of a 'Foley Rationality' approach "makes Foley's approach ill suited to the objectivist, nonrelativistic spirit of our entire framework" (Goldman 2009: 28). But this section has argued such a subjectivism is and must be a feature of any deontically internalist account. What then becomes of Goldman's claim that "this point should be equally acceptable to both internalism and externalism"? Against an internalist of Foley's stamp, I'd suggest this comes worryingly close to begging the question.

[7] Pettit & Brennan (1986) refer to this state of affairs in ethics as one in which the consequentialist conception of the Good is 'calculatingly elusive' or (more strongly) 'calculatingly vulnerable'. A common response to this objection (in its many forms) is to look for auxiliary rules – 'rules of thumb' – perhaps (Smith 1988) the rule to work from expected utility, or perhaps obedience to the rules of common morality. We will not discuss these responses here; noting however, that whatever may be said in favour of such approaches on their own terms, there is reason to doubt they will be adequate as responses to this objection. Two objections that are well discussed in the literature are firstly, that an expected consequentialism is precisely not a form of consequentialism. A 'bounded' restriction of the agent's justificatory status to *expected* consequences is motivated by an 'ought' implies 'can' *deontic* limit on the grounds for normative appraisal – *c.f.* Clifford's (1947: #1) judgement on his ship owner: whether a given decision is justified is decided *before* the consequences are in. Secondly, as regards auxiliary rules / rules of thumb: the objection is made that these lead to regress (Bales 1971, Brink 1986, Smith 1988, Goldman 1980).

is to offer the agent "rules for the direction of the mind"; that is (whether actually rules-based or not) an epistemology that can offer *guidance* in cases where the agent is undecided and facing a doxastic choice. But for an account to be able to offer me practical help in cases of judgement under uncertainty, it is necessary for it to restrict itself to resources – a justificatory ground – which may be available to me in that epistemic dilemma, with those limited resources (processing/capacity limitations, time constraints to reach decision, restricted knowledge-base, etc.) Decision-making requires me to have *access* to my justifiers. Goldman (1980) referred to this as the aspiration that epistemology should offer the agent, facing a decision, a 'doxastic decision procedure' (DDP) – where this latter is a dummy for whatever set of rules for guidance (whether actually rules-based or not) the epistemologist's theory finally divulges. But the objective, truth-directed nature of an externalist theory in epistemology is not guaranteed to give the undecided agent access to any such DDP. In the language of the psychologists, such theories may yield accounts of justified cognition that are *computationally intractable* – hence unusable for the purpose of guidance under uncertainty. So, relative to this ambition, externalism is a failure.

*A digression: DDP objections to internalism*

Goldman (1980) made a *tu quoque* response to the DDP objection – namely, that it applies no less to many varieties of *internalism*. His point is well taken as far as it goes: many supposedly internalist theories are indeed too complex and defeasible for an agent to have access to their criterion of epistemic success. Indeed, this may serve as a criticism of certain (e. g. 'mentalist'[8], or highly complex introspectionist-foundationalist[9]) conceptions of internalism. But not all species of internalism are like this. On a deontic conception of internalism together with a commitment to 'ought' implies 'can', one may start with the powers of the agent and delimit one's account of their epistemic requirements accordingly. This 'bounded' (accessibilist) notion of epistemic justification leads to accounts of epistemic justification[10] that are *not* vulnerable to the DDP objection.

*The main response: regulative vs theoretical*

The classic response by externalists and consequentialists to the DDP objection again involves making a dichotomous distinction as to the aims of an epistemic (/ethical) theory. The terminology of this distinction varies, though in ways which carry some useful semantic pointers as to the underlying differences between its two terms. In what follows I use the terms *regulative* and *theoretical*. A list of cognate terms, with sources, is provided in Table 1.

Externalists/consequentialists are criticised by their opponents for offering epistemic (/ethical) theories that may not be usable by an agent, facing a decision, to regulate thought. Their response is simply to note that this regulative (decision-making, action-guiding) ambition is not a desideratum of their kind of account. Rather, in the case of epistemology, the externalist seeks only to specify when a belief, or a belief-making-process, or a course of cognitive conduct is justified objectively (say, in terms of truth-maximisation or error-avoidance)[11].

[8] This is a familiar objection to 'mentalist' conceptions of internalism (e. g. Conee and Feldman 2001). Simply put: even exclusively mental justifiers may still be massively inaccessible.

[9] If you doubt this, just re-read Chisholm's (1989) across various editions.

[10] A paradigm of such an account in normative epistemology would be Richard Foley – e. g. (1993). In the economics and psychology literature this becomes the notion of 'bounded rationality'.

[11] I shall use this family of contrast terms strictly as explicated here, and have myself been confused by other authors' unexplicated usages. Note that this distinction is not to be assimilated to Sosa's *animal vs reflective* knowledge – an orthogonal distinction that it precedes by many decades. Goldman's 'strong vs weak' justification is clearly somewhere in the vicinity of our distinction (though Goldman (1988: fn. 1) distinguishes that distinction from this). However, I consider the 'strong vs weak' distinction to be confusing and ill formed, accompanied as it is by stipulative and

**Table 1:** Terms used to draw this distinction

| Regulative | Non-Regulative | Goldman (1985) after H.M. Smith |
|---|---|---|
| Doxastic Decision Procedure | Theoretical | Goldman (1980) |
| Regulative | Theoretical | Goldman (1980) |
| Practical Right<br>[Practical Virtue] | Absolute Right<br>[Absolute Virtue] | Chisholm (1957)<br>[after Richard Price] |
| Practical | Theoretical | H.M. Smith (1988) |
| Decision Procedure | Right-Maker, | Bales (1971) |
| Decision Procedure | Criterion of rightness (Sidgwick 1907) | Brink (1986) |
| Motive | Standard | Sidgwick (1907) |
| Subjective | Objective | Alston (1985), Plantinga (1993) after Aquinas |

## 4 Two Motivations but one Distinction

Each of internalism, externalism, levels a standard objection against the other. Each must draw a distinction to escape the objection to their position. Each draws this distinction in terms of the desiderata of their theory, what their theory aims to account for – and what instead it surrenders, what it acknowledges to be no part of its aim. And each does this by explicitly borrowing both objection and distinction from an already developed body of argument in ethics.

Externalist theories identify a very central aspect of what one expects of an epistemic theory: their different, proprietary, ways of marking a connection with the truth, their *objectivity*. Minimally, this connection with the truth is represented by the idea of truth as a necessary condition on knowledge; though, of course, conventional externalist theories tend to go far beyond this in their different (reliable, counterfactual, etc.) accounts of warrant. These wide variations of detail do not affect the point: that one cannot mark the objectivity, factuality, that such accounts base their epistemic success term upon – their emphasis on what is actually truth conducive (or error avoiding) in cognition – without losing any necessary connection with accessibility. Making epistemic normativity *essentially* objective means that it can be at best only *contingently* accessible. For the world is as it is, and we are as we are, and as fallible beings with widely differing epistemic resources, our ability to achieve a given objective epistemic status will be tenuous and uncertain. The potential inaccessibility characteristic of externalist theories is then not a definitional primitive, but a derived consequence of their objectivity – however this latter be construed, even if it be no more than the attainment of truth, much less if it be something more.

Internalist theories also identify a very central aspect of what one expects of an epistemic theory: their different, proprietary, ways of marking a connection with the subject, their help to the subject as a guide, their directiveness – of the subject's cognitive conduct, of thought. Internalist theories' restrictions on only accepting a justificatory ground that may be accessible

unmotivated entailment claims (e. g. that 'strong' entails 'weak' – Goldman 1988: 56); claims which represent what I have (below) entitled 'halo effects'. There is a double dissociation between the internal and the external success notions in epistemology. Any terminology which obscures the radical nature of that double dissociation is to be regretted.

to the agent is also then best seen not as a definitional primitive, being rather *derived* from the desideratum of satisfying this regulative, directive, aim: of satisfying it not accidentally but essentially. To be necessarily capable of directing cognitive conduct – of providing "rules for the direction of the mind" – a theory must be accessible, it must not go beyond the resources of the epistemic agent: resources within the compass of his intellectual abilities, or at least his 'Zone of Proximal Development' (Vygotsky 1978, Plato 1992: 197c–d, the *Theaetetan* aviary example)[12].

The two families of theory, internalism and externalism, have then, as an outcome of their conflict, each been forced to draw the same distinction: a distinction between two separate desiderata of adequacy. Under severe pressure from the other's arguments, each has been forced to abandon any pretensions to satisfy one arm of that distinction – to meet one desideratum of adequacy. Between them, they account for both desiderata, separately, they each account for one alone.

# 5 Irenic Resolution or Gordian [Knot] Threat?

What is not often noticed is that this task-separation at once offers us both the promise of an irenic resolution of some tangled traditional disputes in epistemology and the threat that any such resolution may be Gordian in nature. We may see as irenic a solution that promises a simple division of labour in epistemology (and ethics). Many erstwhile disputes are then not so much solved as dissolved. Internalist accounts alone offer us the resources to satisfy the regulative desideratum; yet must disavow any claim to satisfy the theoretical desideratum. This surely leaves such accounts uniquely well-suited to offer us our position on rationality. And externalist accounts alone offer us the resources to satisfy the theoretical desideratum; yet must disavow any claim to satisfy the regulative desideratum. This surely leaves such accounts uniquely well-suited to offer us our position on knowledge. The fact remains, however, that a number of knotty problems will have been severed rather than untied. For we will end up with an account of knowledge which openly flouts paying even lip service to the regulative desideratum of adequacy; and an account of rationality which openly flouts paying even lip service to the theoretical desideratum of adequacy. We have seen that there are powerful motivations to go down this route; but were we to do so, could we abide the destination we would find ourselves?

## 5.1 Objections to the Gordian Threat: The Virtues Objection

Virtues theorists in both ethics (Aristotle 2000) and epistemology (Zagzebski 1996) are wont to claim that we have achieved our epistemic end (most commonly knowledge) or our ethical end (sometimes the Good, sometimes the Right) just in case we have maximised satisfaction of both desiderata: Theoretical and Regulative. They will deny that we have achieved our success state should only one desideratum be satisfied; thus they will deny that drawing this third distinction renders otiose (solves or dissolves) any or many of the perennial disputes found in normative epistemology.

*Response: Stipulative*

As I have argued at length in Lockie 2008, this objection is merely stipulative. We, all of us (pro or anti virtues-theory) can and do recognise these two desiderata; and may identify in any given case whether this or that desideratum has separately been satisfied. As distinct axiological projects, delineating distinct axiological properties they exist (virtues theorists do not typically

[12] Cf. Kornblith's (1983) objections to Boyd and Goldman: we cannot identify justified belief with *reliably regulated belief*, but (with suitable, e. g. time-indexed, provisos) we can assimilate it to *diligently regulated belief.*

deny this – nor can they). One may, after recognition of this truth, stipulate that ‘virtue’, now employed as a term of art, applies only when both desiderata are met (e. g. Zagzebski 1996); and if this is felt a useful (stipulative) restriction of our philosophical terminology, then fine. What cannot be done to win anything other than a terminological victory, is to specify that knowledge (say) or rationality (say) require this stipulatively so-defined conjoint state of ‘virtue’ (q.v.) to be met; and thus that a candidate account of knowledge, say, which (suppose) brilliantly meets several theoretical desiderata is to be dismissed for failing to meet certain wholly distinct regulative desiderata – thus failing to be a state of virtue, as stipulatively so-defined. It is noteworthy that non-Aristotelian conceptions of virtue, particularly the Stoic (Annas 2003) recognise this point and tend to restrict their account of the ‘virtuous’ to one satisfying the regulative desideratum alone: of which more in the next sub-section.

### 5.2 Objections to the Gordian Threat: The ‘Violated Intuitions’ Objection

Envisage a candidate externalist theory of knowledge that sets out solely to address the theoretical desideratum, perhaps offering a very promising candidate to satisfy this desideratum, yet does so in a way that flouts the regulative desideratum in its entirety (say, it permits very lucky or irresponsibly acquired knowledge – take Sartwell (1991, 1992) as a paradigm: the claim that knowledge is merely true belief). Or envisage a candidate internalist theory of rationality that sets out solely to address the regulative desideratum, perhaps offering a very promising candidate to satisfy this desideratum, yet does so in a way that flouts the theoretical desideratum in its entirety (say, it permits objectively entirely awry, radically false, yet putatively rational beliefs: take Foley (1993) as a paradigm). It will be protested (it routinely *is* protested) that suchlike theories ‘violate our intuitions’ (perhaps, our ‘core’ intuitions); and that is an awful thing – so awful that we must reject the abrupt divorce between theoretical and regulative accounts of epistemic value: an ordinary language / ‘conceptual analysis’ / intuitions-driven metaphilosophy establishes we must satisfy both desiderata in accounting for knowledge or rationality.

*Response: Abandon the indefensible metaphilosophy*

When our intuitions are outraged by, say, an account which has it that an agent *knows* yet wholly fails to satisfy the regulative desideratum, or *is rational* yet wholly fails to satisfy the theoretical desideratum, social-cognitive psychologists refer to (and dismiss) this type of phenomenon as a named species of error: a *halo effect*. We tacitly suppose a beautiful person must be good, and a good person must be wise … and a person who has achieved one epistemic success-state must possess another; but it is not so. Our feelings of oddness at attributing knowledge to someone who has not regulated her thought well (Lockie 2004), or rationality to someone operating under a massive framework of false beliefs (Foley 1993), are just that: mere feelings of oddness – and as such are not to the point. The view that an argument to a position that is driven by fundamental epistemic theory should nevertheless be put in full reverse when it militates against ordinary language intuitions that do not answer to anything like the constraints that shaped and motivated that argument’s development is unacceptable. Affording a priority to such intuitions over fundamental epistemic theory rests on an indefensible (and presently very hard-pressed) tacit, framework, metaphilosophy. There is no reason at all why those of us who are normative epistemologists should cede all the arguments against this metaphilosophy to the naturalists.

*The Conjunction of 5.1 & 5.2: They Don’t Sit Well Together*

Furthermore, when the virtues objection is conjoined with the ‘violated intuitions’ objection, an interesting tension is manifest: the ordinary language intuitions appear to militate against the virtues account. As the present author and others have previously noted (against, for instance, Zagzebski's (1996) account) the ordinary resonance and normative connotations of calling someone a 'virtuous' character suggest we strongly identify this state with a satisfaction of the regulative desideratum alone – whether in ethics or epistemology (Lockie 2008). This is a point emphasized by the Stoic (as opposed to the Aristotelian) tradition in virtue theory, and noted in their different ways by Russell (1996), Annas (2003) and others:

> men everywhere give the name of virtue to those actions, which amongst them are judged praiseworthy; and call that vice which they account blamable ... (Locke 1975: II, 28; cited in Goldman 2001: 30).

### 5.3 Objections to the Gordian approach: Impoverished justification is too cheap a notion

Nottelmann (2013) has an important discussion of blameless belief where this be cut away from other, more objective notions:

> blamelessness in the minimal sense of non-blameworthiness sometimes comes cheaply. In fact too cheaply, it would seem, to take a version of DCEJ [the deontic conception of epistemic justification] predicated on plain blamelessness seriously as a conception of EJ [epistemic justification]. The problem is that there could be beliefs which are blameless, only because it makes no sense to blame the believer for holding them. But intuitively it would then seem that it makes as little sense to evaluate such beliefs as epistemically justified.[22] ... suppose that each of us is for some reason born with an ineradicable[13] belief, for or against which we may never obtain relevant evidence ... Blaming us for this belief, if it is truly innate and ineradicable, seems strange. But so does declaring it somehow epistemically justified. I shall assume here that adherents of DCEJblame are willing to bite this bullet. Perhaps they will, like Goldman rest content that their conception of EJ captures "some chunks of intuition regarding "justification" (in its epistemic application).
>
> [His F.N. 22] One way out of trouble here, is distinguishing justified belief from responsible belief, and maintain that the former is a purely negative concept, e. g. consisting in the absence of obligations breached in holding the belief. In this vein, ineradicable beliefs may well be held justified, even if they are neither responsible, nor irresponsible. I hold this reply to be unconvincing: To me, justification seems more than a merely negative concept and cases of ineradicable belief provide decisive evidence against this view.

Nottelmann is right here, as far as it goes, but he has just identified that blamelessness (which in this context may hold place for any justificatory notion that solely answers to the regulative desideratum) is an *incomplete* notion of justification – and we established that ourselves. Of course an important axiological notion in epistemolgy concerns *objective* truth-conduciveness; but that is just a different notion to the deontic notion. Equally it may sometimes be important to conjoin our two notions of epistemic justification: internal and external – but they are distinct notions, then-conjoined. The terminology isn't important (so, Nottelmann's canvassed (footnoted) distinction between 'justified' and 'responsible' doesn't seem felicitously phrased to me)[14]. What matters is that we acknowledge that the principle 'ought' implies 'can' and our

[13] I have corrected throughout 'eradicable to 'ineradicable' as Nottelmann has confirmed (personal communication) that these are typos.

[14] Note that I don't define thin deontologism in terms of *blamelessness*. My deontologically justified subject has typically worked hard for his or her justificatory status – for instance, to the point where he or she could be *commended*

epistemic limitations forces on us a notion of perspectivally limited 'blameless' (if you must) justification however much it jars the ears: this jarring is just a halo effect. Suppose, to take Nottelmann's example, we consider ineradicable beliefs, or cognitive limitations forced upon us by our biological natures (à la Kant, Chomsky, McGinn). Were there such beliefs/limitations it would seem to me to be indeed correct to say we would then be, as a species, blameless and *as justified as we can be* in holding these (nevertheless false) beliefs. The same thing goes for beliefs that are a product of our cultural-historical situation rather than our biological limitations: Newton blamelessly believed in absolute simultaneity – he was justified in this (false) belief; Alston's tribesman blamelessly believes in his culture's metaphysical world-view, and so on. We all have our limits. Justification in this important sense, applies to us thinking as well as we can within these limits.

### 5.4 The Situation that Confronts Us

There is a general objection to any attempt to efface this distinction. Any insistence that we should elide or conjoin the twin desiderata we have identified thus far, faces the charge that this would be simply to ignore the situation that confronts us as epistemic agents. Recognition of the distinct nature of these desiderata is forced on us by consideration of our nature in facing decision. By hypothesis, in facing a decision, we have no 'marker' of which of our beliefs are true, which false – if we had, epistemology as an enterprise, and the questions we are addressing, would be superfluous. We have our beliefs, both true and false, and must move forward from these altogether to *find out* the justified ones – those likely to be of facts. This just is the regulative project. A consideration of one engaged with this project (and we are all engaged with this project) leaves us with no choice but to acknowledge the agent's epistemic limitations, her fallibility and frailty, the fact that her epistemic resources may not (and often will not) be up to the task. This situation simply confronts us, each of us, *qua* epistemic agent, *qua* inquirer in the world, and in acknowledging it we must recognise two things. First, notwithstanding the possibility that in such situations the agent may unavoidably be led into error, there is a core, vital, sense in which in such a situation she may nevertheless be justified – and *will* be justified should she marshal her resources as effectively as her perspectival limitations permit. Second, that in such a situation, she is nevertheless avowedly *in error* – that is, as much as she may have satisfied one desideratum and achieved justification thereby, she has failed to satisfy another undoubted aim of epistemology and lacks justification thereby. Conversely, when our agent, despite guiding her thought badly, has attained the truth, or some other objectively desirable factive state or relation to the world, we are describing at once both a type of epistemic success and a distinct type of epistemic failure. We have no option but to recognise these types of achievement (and failure) as distinct. There is a double dissociation between these two desiderata. Insisting that we must abandon two separate and distinct forms of assessment of the epistemic agent ignores much of what it is to be an agent in the world: the subject of decision and normative appraisal consequent upon that decision.

for this. But certainly he or she could be (radically) *wrong* – conceivably because of ineradicable beliefs. Note also, that justified beliefs in this sense are not usually punctate (e. g. indivisible, *sui generis* – say biologically 'ineradicable'). Usually, even if objectively false, such beliefs are a product of articulated reasoning, inferential relations and complex argument. Such argument is a cultural-cognitive product and the beliefs which eventuate from it may indeed be 'eradicated' by further argument or diachronic dialectic, albeit not necessarily for an objectively true belief in turn. Issues here intrude concerning Whig history, verisimilitude, and topics long discussed in the philosophy of science, but a punctate ineradicable belief that is yet apt for normative epistemic assessment – as justified or unjustified – would actually be rather an unusual thing.

# 6 Normative Epistemology in Light of this Distinction

## 6.1 Knowledge/Rationality and the Regulative-Theoretical Distinction

As stated at the outset, the existence of this third distinction is hardly esoteric knowledge in epistemology. Yet how seriously has it been taken by those seeking to develop accounts on either side of our first distinction – accounts of rationality or of knowledge? Does an awareness of the theoretical/regulative distinction really inform theory construction in epistemology? Periodically, the philosophical community approaches the insight that an epistemology which developed and shaped accounts of either rationality or knowledge in light solely of the appropriate desideratum for that item, might make space for Gordian treatments of certain epistemic problems: shocking accounts which otherwise would be scorned, marginalized or simply not entertained. But then, having approached such insights, the world of academic epistemology just seems to veer away[15].

So, over the last decade, a long overdue consideration of the *Meno* value problem has been underway (via, for example, the growing literature on the 'Swamping Problem'; and variants of Zagzebski's (2003) 'espresso machine' analogy). But the terms under which this problem has been discussed are desperately prescribed; participants in these debates seem, at the level of framework presupposition, not to entertain the possibility of a *radically* theoretical account – say, of the need to address the challenge represented by a genuinely Theaetetan[16] theory of knowledge (Sartwell 1991, 1992, Plato 1992: 187b, Plato 1956: 97a–b). Where such accounts are entertained, they are taken unargued to be merely a *reductio* of any theory that entails them (Chisholm 1988: 287). From the other side of the first distinction, the viability of extremely internalist accounts of rationality – e. g. 'Foley rationality': that rationality consists in being justified by our own deepest epistemic standards (Foley 1993) – are routinely disparaged on grounds which, for one who sees any such account as answerable solely to regulative considerations, are plainly non sequiturs.

## 6.2 Internalism/Externalism and the Regulative-Theoretical Distinction

The big challenge for intransigents on either side of the internalist-externalist distinction, is to ask how much will remain of their respective hostile arguments after fully acknowledging this distinction concerning the desiderata for epistemic theories; and in particular, after establishing that internalists and externalists alike are each already committed to their specific account answering only to one of the distinct component aspects of this distinction in desiderata. Isn't this just to effect a (possibly unwelcome) irenic resolution of much that was hitherto in vehement dispute? Internalism satisfies the regulative desideratum of adequacy and gives us our account of rationality. Externalism satisfies the theoretical desideratum of adequacy and gives us our account of knowledge. There is then plenty still to disagree about, like: What *is* the correct externalist theory of knowledge? What *is* the correct internalist theory of rationality? And in particular: What would an externalist theory of knowledge look like were this theory to be de-

[15] Foley (e. g. 2004) is the one, great, stand-alone exception to this rule, and I am right glad to acknowledge my debt to him. There are however, many others who approach his insight only then to retreat from it. For instance, Kvanvig considers an agent who possesses (with full understanding) true beliefs of, as he put it, a 'fortuitous ætiology' and notes *solely on grounds of their provenance* [which surely pertains to the regulative desideratum] that "we should not say that … she is lucky to have the knowledge she has, for knowledge [which surely answers to the theoretical desideratum] rules out this kind of luck" (Kvanvig 2003: 199). At least, we should rule this out "if we have learned our lessons from the Gettier literature" (Kvanvig 2003: 198). Given the plain fact that there exists some lucky knowledge, I suggest we might choose to learn a quite different lesson from, and about, this literature (c.f. Lockie 2004). For a further argument for Foley's position see Booth (2011).

[16] I mean: the theory advanced by Theaetetus in the *Theaetetus* (187b), not the theory advanced by Plato in the *Theaetetus*.

veloped wholly and solely with a view to addressing the theoretical desideratum? What would an internalist theory of rationality look like were this theory to be developed wholly and solely with a view to addressing the regulative desideratum? Should such questions be addressed with, and motivated by, an explicit awareness of the foregoing meta-epistemic arguments, the answers to them would be likely to prove *very interesting indeed*. But that such a state of affairs should come to pass would require a Rubicon to be crossed in modern epistemology.

What an informed awareness of these meta-epistemic issues should minimally lead us to question is how much point there is to the familiar dialectic which occurs when partisans for the one approach upbraid partisans for the other approach on the basis of, say, the inability of this (avowedly theoretical) account to satisfy this (clearly regulative) desideratum – or vice versa. Further, these meta-epistemic issues should lead us to question the requirement that one and the same normative epistemic theory should answer to (indeed maximise) both desiderata at once. There are strong reasons to doubt whether any one account can satisfy both desiderata. Within epistemology we need to confront this situation and entertain the radical conclusions which appear to follow from it.

## References

Alston, W. (1985), 'Concepts of epistemic justification', *Monist* **68**, 57–89.

Annas, J. (2003), 'The structure of virtue', *in* L. Zagzebski & M. de Paul, eds., 'Intellectual Virtue: Perspectives from Ethics and Epistemology', Oxford University Press, Oxford, pp. 15–33.

Aristotle (2000), *Nichomachean Ethics*, tr. R. Crisp, Cambridge University Press, Cambridge.

Bales, R. E. (1971), 'Act-Utilitarianism: account of right-making characteristics or decision-making procedure?', *American Philosophical Quarterly* **8** (3), 257–265.

Bergmann (2006), *Justification Without Awareness*, Clarendon Press, Oxford.

BonJour, L. (2003), Reply to Sosa, *in* L. BonJour & E. Sosa, 'Epistemic Justification: Internalism vs. Externalism, Foundations vs Virtues', Blackwell, Oxford, pp. 173–200.

Booth, A. (2011), 'The theory of epistemic justification and the theory of knowledge: a divorce', *Erkenntnis* **75**, 37–43.

Brink, D. O. (1986), 'Utilitarian morality and the personal point of view', *Journal of Philosophy* **83** (8), 417–438.

Chisholm, R. (1957), *Perceiving: A Philosophical Study*, Cornell, Ithaca.

Chisholm, R. (1988), 'The indispensability of internal justification', *Synthese* **74**, 285–296.

Chisholm, R. (1989), *Theory of Knowledge*, 3rd edn., Prentice Hall, Englewood Cliffs, NJ.

Clifford, W. K. (1947), The ethics of belief, *in* 'The Ethics of Belief and Other Essays', Watts and Co., London, pp. 70–96.

Conee, E. & Feldman, R. (2001), Internalism defended, *in* H. Kornblith, ed., 'Epistemology: Internalism and Externalism', Blackwell, Oxford, pp. 231–260.

Foley, R. (1993), *Working Without a Net: A Study of Egocentric Epistemology*, Oxford University Press, Oxford.

Foley, R. (2004), A trial separation between the theory of knowledge and the theory of justified belief, *in* J. Greco, ed., 'Ernest Sosa and His Critics', Blackwell, Malden, pp. 59–71.

Goldman, A. (1980), 'The internalist conception of justification', *Midwest Studies in Philosophy* **5**, 27–52.

Goldman, A. (1985), *Epistemology and Cognition*, Harvard University Press, Cambridge, MA.

Goldman, A. (1988), 'Strong and weak justification', *Philosophical Perspectives* **2**, 51–69.

Goldman, A. (2001), The unity of the epistemic virtues, *in* A. Fairweather & L. Zagzebski, eds., 'Virtue Epistemology: Essays On Epistemic Virtue And Responsibility', Oxford University Press, Oxford, pp. 30-48.

Goldman, A. (2009), 'Internalism, externalism and the architecture of justification', *Journal of Philosophy* **106** (6), 1–30.

Hieronymi, P. (2008), 'Responsibility for believing', *Synthese* **161** (3), 357–373.

Kornblith, H. (1983), 'Justified belief and epistemically responsible action', *Philosophical Review* **92** (1), 33–48.

Kvanvig, J. (2003), *The Value Of Knowledge And The Pursuit Of Understanding*, Cambridge University Press, Cambridge.

Locke, J. (1975), *Essay Concerning Human Understanding*, ed. P. H. Niddich, Oxford University Press, Oxford.

Lockie, R. (2004), 'Knowledge, provenance and psychological explanation', *Philosophy* **79**, 421–433.

Lockie, R. (2008), 'Problems for virtue theories in epistemology', *Philosophical Studies*, **138** (2), 169–191.

Nottelmann, N. (2013), 'The deontological conception of epistemic justification: a reassessment', *Synthese* **190** (12), 2219–2241.

Owens, D. (2000), *Reason Without Freedom*, Routledge, London.

Pettit, P. & Brennan, G. (1986), 'Restrictive Consequentialism', *Australasian Journal of Philosophy* **64** (4), 438–455.

Plantinga, A. (1993), *Warrant: the Current Debate*, Oxford University Press, Oxford.

Plato (1956), *Meno*, tr. W. K. C. Guthrie, Penguin, Harmondsworth, Middlesex.

Plato (1992), *Theaetetus*, tr. Levett (revised Burnyeat), Hackett, Indianapolis.

Russell, B. (1996), *A History of Western Philosophy*, Routledge, London.

Ryan, S. (2003), 'Doxastic compatibilism and the ethics of belief', *Philosophical Studies* **114**, 47–79.

Sartwell, C. (1991), 'Knowledge is merely true belief', *American Philosophical Quarterly* **28** (2), 157–165.

Sartwell, C. (1992), 'Why knowledge is merely true belief', *Journal of Philosophy* **89** (4), 167–180.

Sidgwick, H. (1907), *The Methods of Ethics*, Macmillan, London.

Smith, H. M. (1988), 'Making moral decisions', *Noûs* **22**, 89–108.

Vygotsky, L. (1978), *Mind in Society*, tr. M. Cole, Harvard University Press, Cambridge, MA.

Zagzebski, L. (1996), *Virtues of the Mind*, Cambridge University Press, Cambridge.

Zagzebski, L. (2003), 'The search for the source of epistemic good', *Metaphilosophy* **34** (1–2), 12–28.

abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 8, Number 1, pp. 15–29, 2014

# Reactive Commitments: Reasoning Dialectically about Responsibility[1]

David Botting

davidbotting33@yahoo.co.uk

**Abstract**

Philosophy has recently been presented with, and started to take seriously, sociological studies in which our 'folk concepts' are elaborated. The most interesting concepts studied are moral concepts, and results have been achieved that seem to sharply contradict the speculation of philosophers and to threaten the very way in which moral philosophy has been done in the past. In this paper, I consider these results and then sketch a version of a reactive attitude theory that allows for a genuine sense in which our intuitions about responsibility may be incoherent in a certain sense but without making moral reasoning radically contextual.

## 1 The Problem

### 1.1 The Data

In several studies, scenarios were described to people and they were asked whether the agent in the scenario was responsible for his actions. Judgments have been shown to be asymmetrical around several axes:

A. The abstract versus the concrete (Nelkin 2007, 247–48).

When given specific details, respondents are more likely to find a person responsible, but if questions are given in abstract form then respondents are less likely to find a person responsible.

B. Moral status asymmetries (Nelkin 2007, 248–50).

i) Emotion asymmetry

When the act performed is bad, we accept the presence of high emotions in the agent as an explanatory and mitigating factor. The bad act seems to us worse if done calmly, but the good act is not judged differently depending on whether it is done on impulse or deliberation.

ii) Side-effect asymmetry[2]

When a side-effect is unintended but foreseen, then people will usually say that the agent is responsible for it when it is bad, but not when it is good.

iii) Severity

[1] The author would like to acknowledge that funding for this paper was received from the FCT Portugal under grant awards "Argumentation, Communication and Context", PTDC/FIL-FIL/110117/2009 and "Is moral reasoning essentially dialogical?" SFRH/BPD/77687/2011.

[2] In this position Nelkin has "Intention and act asymmetry" that "When the act performed is bad, then the intention to perform it is usually held blameworthy. When the act performed is good, the tendency to praise the intention to perform it is less strong"; I will not discuss this here. This explains the difference between my labelling and Nelkin's.

> We judge acts more harshly when they have more harmful consequences than when they do not even when the act itself is the same. This means that a drunk driver who happens not to hit anything is the beneficiary of moral luck.

This data poses poses a threat to the whole way moral philosophy has been done so far. How has moral philosophy been done so far? It is widely agreed that it proceeds by *the method of cases*. What does this mean? What, according to the method of cases, is the relationship between the theory and the data? In the next section I will outline two approaches to moral philosophy – roughly, one that focusses on normative issues and one that focusses on descriptive issues – and try to show what follows from the data on these approaches.

### 1.2 The Relation of the Theory to the Data

On the first approach, moral philosophy can be seen as aiming at a theory or analysis of moral concepts. The explication of a moral concept may or may not involve a decision procedure. For instance, it does not follow necessarily from a utilitarian theory of 'the good' that agents can, or even should, try to calculate the net utility of all the possible consequences of their action; we would think there is actually something morally defective in an agent who, when his wife was drowning, tried to decide whether or not to save her by hedonic calculus or by wondering whether universalization of his maxim leads to a contradiction in conception. The theory is an account of what it is for such decisions to be correct and not a description of how such decisions are reached or what decision procedures should be used, although it does provide a norm for those procedures to aim at. This being so, it is possible that the decision procedure that should be used is one that does not, in the particular case in question, result in a decision that does, for example, maximize expected utility; rule-utilitarianism, for instance, may be the correct decision procedure even in cases where its application leads to sub-optimific results. Hence, there are two senses in which our intuitions about cases may be correct: they may result in the best outcome, or they may issue from the best procedure. This is important because what the data is really capturing – at least when we reason about whether to hold an agent responsible – is our decision procedures or, equivalently, our criteria for applying the concept. In consequence, we should expect some discrepancy between the data and the theory; futher argumentation is required to relate the data to the theory of responsibility, and this will be shown to rest on certain assumptions that the data will show to be questionable. However, we should not give this fact more than its due; it is not unreasonable to suppose that at least the embryo of a correct theory is discoverable in the data, yet this too will turn out to be problematic.

A theory of a moral concept will predict when something falls under that concept, e. g., when a scenario is fully described the concept will tell you whether the actor in that scenario is morally responsible. How are we to test such a theory? We use the *method of cases*: we test the scenario against the moral judgments of ourselves and others. Since whether that actor is responsible is not something observable, the prediction to be tested is not whether the actor is responsible but whether judges presented with the scenario will make a particular moral judgment, e. g., to hold the actor responsible, and this will depend on and reflect their decision procedures (arguably, simple intuition may qualify as such a procedure), and it is these that have a direct relation to the data. So, to be able to test the theory and indirectly to make the data relate to the theory we have to make the assumption that, most of the time at least, the moral intuitions of human subjects asked to judge the scenario will be correct.

However, this assumption has the further presupposition that intuitions are coherent amongst themselves, that subjects are not adversely influenced by non-formal features of the scenarios such as high affect, yet the data shows that scenarios that are formally identical elicit varying intuitions and one plausible explanation why this is is that these non-formal features are selecting different psychological processes/criteria for holding responsible/decision procedures, not only across different subjects but, what is more to the point, within the same subject across formally identical cases. A second possible explanation is that literally different concepts of responsibility are being selected, each with a single set of criteria. A third is that we have a single concept of responsibility but one that is highly sensitive to contextual features. In most cases the second and third possibilities will be indistinguishable.

It is this third possibility that seems to be the focus of debate, but I will show that this may be the result of a misunderstanding and a confusion between this and the first possibility, between a normative and a descriptive bias – criteria that are constitutive of a concept (its analysans) are not necessarily those of its application, e. g., we may attribute 'good' to an action on the basis that it follows a certain utilitarian rule without that action satisfying all the necessary conditions of goodness as the theory defines it, and it is not a mistake to do so. It cannot be assumed that such a procedure is the wrong one to use simply on the grounds that in this particular case its result is not the one the theory defines as "correct" – the theory of responsibility tells you what is the correct decision, not necessarily what is the correct decision procedure. These judgments/intuitions may be the correct ones to have.

Even so, the problem is that whatever the theory and whatever our definitions of correctness are, they are formal, and if intuitions do not depend on application of a single criterion but on different criteria depending on non-formal features of the individual case (and this is one possible explanation of the asymmetries in the data), then the relation between data and theory breaks down; the data cannot tell us anything at all about the theory *including, in particular, that we have different concepts of/a variantist concept of responsibility*. We *could* be selecting different concepts on the basis of non-formal features and this might explain the data, but the data itself does not and cannot show this. This point is important in what follows.

What is this "descriptive bias"? It is to take the second approach, in which the theorist sees his task less as providing a conceptual analysis and more as providing a philosophical clarification of the folk concept of responsibility. This can be seen as a descriptive, naturalizing approach, and is allied with the first possible explanation mentioned above.[3] The philosopher taking this approach does not talk about reponsibility itself but about our *attributions* of responsibility, making the relevant question "Is there a single criterion for attributing responsibility?" However, keeping conceptual analysis at bay does not mean that we cannot include a normative aspect of epistemology or that the issue of correctness is simply ignored. Even a descriptive approach can have normative consequences, here an account of what is the correct procedure. Nelkin (2007, 246) seems to be taking this approach when she assumes:

> Fit Assumption – The criteria for moral responsibility attributions fit well
> with all (or most) of our ordinary judgments.

This refers only to moral responsibility attributions and not to being responsible; the *concept* of responsibility is not mentioned at all. It is this assumption that seems to be threatened by the view of Doris, Knobe and Woolfolk (2007) who interpret the data of our ordinary judgments as showing that they are not made according to a single criterion and therefore there cannot be a single *invariant* criterion for *ascriptions* of moral responsibility (a view they call *invariantism*);

[3] When Mele (2003, 334) describes his project as "to construct a viable theory about how agents produce their intentional actions, *as I* (and many philosophers of action, I believe) *conceived of intentional actions*" [italics original] he seems to be moving towards this approach, but in a few sentences this has been re-construed as conceptual analysis of a 'core' concept. It seems that Mele has not really entered into the spirit of this approach.

there are different criteria depending on (e. g., psychologically and non-rationally selected by) non-formal features, hence the asymmetrical responses to formally symmetrical cases. That there are different criteria for *ascribing* the concept of responsibility (the first possibility mentioned above, called by them *variantism*) does not entail (the second and third possibilities mentioned above) that there are different concepts of responsibility or that the concept of responsibility is variantist, i. e., contextual.

In attributing to them the latter view and then refuting it, Warmke makes a straw man of their position, taking it as an conceptual analysis of responsibility rather than an empirical thesis about responsibility attributions. Warmke (2010, 2) puts the assumption as:

> *Conservativist Assumption*: The conditions for being morally responsible for an action should accord with all (or most) of our ordinary judgments about the conditions under which an agent is morally responsible and we can discover these conditions by considering these ordinary judgments.

This differs from the Fit Assumption in two ways.

Firstly, despite acknowledging this as a methodological assumption it should be noted that he refers here to the conditions for *being* morally responsible, i. e., the concept. Then he objects validly that nothing follows about the concept of responsibility unless we assume firstly that the asymmetrical responses are *correct* in that agents are correctly held responsible, and secondly that (most) agents that are correctly held responsible would also be responsible (as, arguably, they would be in a reactive attitudes theory). This is a problem for his own Conservativist Assumption but not for the Fit Assumption.

Secondly, the Conservativist Assumption claims that the conditions for being responsible can be discovered in the data. Obviously, the Fit Assumption does not say anything about discovering conditions. However, the *method of cases* itself does license an inference from a case satisfying a set of formal conditions to another case satisfying the same set of conditions (this is what is meant by their being 'relevantly similar') in the following way: the common methodology of the theoretician has been to present cases, state their own ordinary judgment about those cases, and by abstracting away the specific details of the cases, sort out a formal set of conditions for responsibility that is considered to be an adequate criterion for all cases independent of context, conditions that set out what similarities are relevant to judgments of responsibility. However, the side-effect case presents scenarios that are said to be relevantly similar, and the concrete/abstract cases present exactly the same scenario but described differently. Why, then, having decided that the agent was not responsible in one of these cases is this classificatory judgment not transferred from the 'source' to the 'target'? The data seem to show that being judged to be relevantly similar is not a stable basis for any inference from one case to another, or to put it another way, moral intuitions are not sufficiently coherent to determine what differences are or are not relevant; they resist formalization. This seems to be the principal threat posed by the data to the theory: it is tacitly to give up on the *method of cases*, at least as a means of theorizing about responsibility.

That Doris, Knobe and Woolfolk (2007) need not be interpreted as talking about variantism with regard to the concept of responsibility (the third possibility) does not mean to say that it is not a viable point of view and, as shown above, it is a possible avenue for explaining the data. To show that the concept of responsibility involved is variantist requires showing that more than one pattern of responses is actually correct, and if only one one pattern of responses is actually correct then the concept is invariantist. I will call these *metaphysical variantism* and *metaphysical invariantism*; these metaphysical theses require, in each case, an error theory for whatever patterns of responses are incorrect. Additionally, the metaphysical variantist would need to show that there is a single variantist concept rather than multiple invariantist concepts, such as responsibility as attributability and as accountability described by Watson (1996). But here it seems that the data helps the metaphysical variantist because the data is characterized by asymmetries, so for these asymmetries to be explained by multiple concepts it means that one and the same subject must shift from one concept to the other in considering one and the same scenario, which seems unlikely. In summary, both invariantism and variantism with regard to the concept of responsibility is consistent with variantism as Doris, Knobe and Woolfolk understand it which I will call *methodological variantism*. However, one moral we can draw from Warmke's critique is that even *methodological variantism* does not follow if all but one of the ascriptions are *performance errors*, defined as *mis*applications of a single criterion or *mal*functions of the same psychological process, perhaps caused by an affective bias.

Let me illustrate the possible strategies involved with an example. The side-effect asymmetry was discovered by eliciting folk intuitions on the following vignettes (Doris, Knobe and Woolfolk 2007, 193–94):

> The HARM scenario
> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment."
> The chairman of the board answered, "I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program."
> They started the new program. Sure enough, the environment was harmed.
>
> The HELP scenario
> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, and it will also help the environment."
> The chairman of the board answered, "I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program."
> They started the new program. Sure enough, the environment was helped.

Subjects were asked whether the chairman was blameworthy in the HARM scenario and praiseworthy in the HELP scenario. Obviously the possible answers are:

i) No blame/no praise.
ii) Blame/praise.
iii) Blame/no praise.
iv) No blame/praise.

We can forget about (iv) since this is counter-intuitive and hardly anyone answers in this way.

According to Knobe, the scenarios are identical in all features relevant to the chairman's relation to his act. Therefore, on the invariantist assumption of a single criterion we should get a symmetrical response, that is to say, either (i) or (ii). On the further assumption that one of the conditions for ascription of responsibility for an act is that the agent performed the act intentionally and in the vignettes the act is only a side-effect, the expected response is (i). However, most people by far give the asymmetrical response (iii). However, what is not always noted is that this does not actually support the claim that ascriptions of responsibility are variantist, because the same criterion that includes the act's being intentional seems to have been applied in both cases; subjects do not take themselves to be in violation of the intentionality condition since they are also prepared to say that the chairman harmed the environment intentionally but not that he helped the environment intentionally. Here, there is a single criterion for ascriptions of responsibility but, arguably, not for ascriptions of intentionality since in both cases the chairman had identical psychological states and so whether or not the action was performed intentionally should give a symmetrical response, but in fact the intentional/unintentional asymmetry mirrored the praise/blame asymmetry.

So, if we take the intentional/unintentional asymmetry as the prior explanandum, and we can explain it without referring to responsibility, then methodological invariantism with respect to responsibility follows, or at least the data do not provide any evidence against it. This preserves the traditional idea that only acts performed intentionally are blameworthy and that the intentionality judgment is prior – the question then is whether there is invariantism (methodological or metaphysical) with respect to intentionality; the data suggest that there is not. In contrast, Knobe takes the praise/blame asymmetry as the prior explanandum, explains it through variantism, and concludes that our responsibility judgments in some way influence our intentionality judgments and that the variance in the latter is explained by the variance in the former. This would imply an objectionable circularity when construed as saying that the concept of responsibility referred to intentionality and the concept of intentionality referred to responsibility, but construed only as saying that we can infer whether an action is intentional if we know whether the agent is responsible for it and vice versa, then circularity only occurs if these are parts of *a single chain of reasoning*. That there are some occasions when we infer from intentionality to responsibility and other occasions (other chains of reasoning) when we infer from responsibility to intentionality is perfectly acceptable and orthogonal to the issue of whether there is a single criterion for ascribing responsibility or intentionality.[4]

What does this mean for the *metaphysical invariantist* who does have a concept of responsibility and is trying to test his concept against the data? If he takes the intentionality condition as part of his concept then one might initially think that he is committed to (i) and must explain away (ii) and (iii) as errors. However, as I have shown above (iii) need not be an error if he modifies his concept of intentionality, and then it is (i) and (ii) that need to be explained away, perhaps by variantism with respect to intentionality or more simply as performance errors.

[4] This is most clearly the case when one inference is deductive and the other (weaker) inference is abductive. In itself, this is not circular, and provided that what is being inferred from is itself justified there is no error (unless you consider abductive reasoning erroneous) in drawing these inferences. The kind of circularity or error that I have in mind is where the interderivability of two propositions is erroneously taken to entail that they are both true or both justified, as would occur if the judgment of responsibility for an action is made *unjustifiedly*, an inference (itself justified) made that the action was performed intentionally, and then an inference made (also itself justified) back to the judgment of responsibility, as if by travelling in this circle something unjustified has or could become justified.

It is not entirely clear what Knobe means by the claim that ascriptions of responsibility and intentionality affect each other. Both he and Wright and Bengson (2009) seem to regard it as problematic but this seems to me to be either a mistake or they are arguing for a variantist concept after all. In the chairman case it could be argued (but I am not saying that anyone does) that some subjects use variantist criteria of intentionality and then make inferences about responsibility, and that other subjects use variantist criteria of responsibility and then make inferences about intentionality. Here we would have variantist criteria of application of both concepts, but in any particular case the directionality of the inference is one way; it is not that a judgment about responsibility is made on the basis of a judgment about intentionality itself made on the basis of a judgment about responsibility. Which out of

i) (responsibility → intentionality) exclusive-OR (intentionality → responsibility)
ii) (responsibility → intentionality → responsibility)
iii) (intentionality → responsibility → intentionality)

is being argued for? While (ii) and (iii) involve circularity, (i) does not when interpreted as concerning ascriptions.

This shows that the data as such is not inconsistent with invariantism of either variety. The threat posed by the data is actually different, and this is the assumption, explicit in the Conservativist Assumption but also implicit in the method of cases as such, that there is a coherent set of intuitions to begin with from which we can *discover* in ordinary judgments the conditions for responsibility or for ascribing responsibility, or in other words, that the folk concept can be the basis of our theorizing, and if this is the case then it is difficult to know how the theorist can get started. The cost of *variantism* then is to make the discovery of standards of correctness in ordinary judgments impossible, undermining firstly the otherwise plausible assumption that most of our ordinary judgments are correct and secondly an analogical form of reasoning that is used in moral and especially in legal contexts.

My claim will be that once we understand the nature of moral reasoning we will understand better the patterns of moral ascriptions they lead to. I accept the assumptions that only one pattern (I choose the asymmetrical pattern) of responses is correct and, by adopting a version of the reactive attitudes theory, that agents that are correctly held responsible (which is most of those held responsible) would also be responsible. The data is eliciting reactive attitudes and is to be explained by looking at the kind of reasoning from which those attitudes issue. My claim – and the essence of my position – is that the nature of moral reasoning is dialectical; it concerns the exchange and evaluation of reasons. This will be the subject of the next section.

# 2 The Solution

## 2.1 The Theory

The reactive commitment theory is the reactive attitude theory dialectified. A commitment is attributed to a speaker by his audience when the speaker makes the linguistic performance of an assertion. This is fair because in taking what is spoken as an assertion the audience must presuppose that all the conditions of satisfaction of the speech act of asserting have been satisfied, in particular the sincerity condition that requires that the speaker believe the propositional content of his utterance. A commitment is not, however, the same as a belief; if the utterance is *mis*taken as an assertion, for instance if the speaker is acting deceptively, then a commitment is still attributable to the speaker. Thus, a commitment can be thought of as being like a contract between speaker and audience and is entered into in the cases of both genuine assertions and its *misfires*.

In this way, discussions where reasons are given in support of a proposition $p$ are modelled dialectically in terms of generating commitments and the speaker trying to show that commitment to $p$ is a consequence of other commitments shared by speaker and audience. This is basically the process of proving, e. g., by natural deduction, a conclusion from given premises, but construed dialogically as a series of assertions of lemmas and questionings of those lemmas until commitments are appealed to that cannot be denied by the questioner (the audience) without violating their own contractual obligation (their shared commitments). The speaker then wins the discussion by appealing to the dialectical rule called the *closure rule* that prevents introduction of or appeal to commitments that are not shared and/or continuing to hold a commitment set that has been shown to be logically inconsistent. On the other hand, if $p$ is not shown to follow from the shared commitments then the audience appeals to the *closure rule* to win the discussion. Such appeals are not often modelled as dialectical moves in themselves, but my view is that they should be seen also as generating a commitment, but that this is a *reactive commitment*. The reactive commitment can be seen as a permission (which may or may not be exercised) to impose sanctions for breaking a contractual obligation. Unlike the kind of commitment discussed first where a commitment is generated by the mere performance of an assertion, a reactive commitment is only generated if the closure rule is used *correctly*.

This is linked to the reactive attitudes theory in the following way: an agent is responsible if there is a practice of *praise* or *blame* connected to his action. This is summed up in the phrase "an actor is responsible if she is held responsible." A moral sanction is analyzable as a speech act of blame.[5] To be a genuine blaming and not a misfire certain conditions must be satisfied, three of which are of paramount importance: the blamer must have *permission* to blame (there is a reactive commitment), the blamer must have *power* to blame (it must be physically possible that what is permitted can be done – a person who makes a 'threat' that he has no means to carry out does not really perform the speech act of threatening), and the blamer must not take the person blamed to be *morally unlucky*. It is *not* a condition that the blamer has the reactive attitude associated with blame, although he must have *some* reactive attitude, e. g., towards the discussion itself. A would-be blamer who has the reactive attitude associated with blame and desires to blame but lacks, for instance, the power to blame, cannot really blame. His attempts to blame would misfire – he would be merely letting off steam, 'playing' at blame. To answer "Yes" to the question "Is the chairman blameworthy?" in a questionnaire is not the same as blaming. Similarly, if he takes the agent to be morally unlucky but justifies blaming them on instrumental/pragmatic grounds (e. g., as a social deterrent) then he does not really blame even if, perhaps, this is the correct decision procedure. In such situations the agent is not responsible. However, the fact that there is a reactive commitment and yet the agent is not blamed does not necessarily mean that the agent is not responsible. There are cases where the conditions are met and yet the reasoner does not blame because of personal reasons such as the feeling that he is in no position to judge. It should be noted here that the reactive commitment makes blame permissible and not obligatory, or perhaps it would be better to say that the obligation to blame generated is one that can be defeated. In this situation the agent *is* responsible.

An agent is negatively responsible if open to blame and positively responsible if open to praise, or to put it another way, if the dialectical rules governing the kind of dialogue where an agent is obliged to justify his actions to a questioner *could* result in blame or praise respectively. This is not, of course, to suggest that such a dialogue actually takes place. The claim is rather that a reasoner reflecting on what reactive attitude to take, if any, reasons as if such a discussion was taking place. Moral reasoning is inherently dialogical.

The concept of responsibility given above is invariantist. It differs from most invariantist concepts because it considers the perspective of the judge as well as the actor, which allows for different judges to correctly give different responsibility ascriptions to the same actor. How this deals with the data of the first section will be the subject of the next section.

## 2.2 Accommodating the Theory to the Data

Explaining away the data has become something of a cottage industry, generating an impressive number of hypotheses which, even more impressively, have been found to have empirical support by their proponents. Hypotheses vary along the following axes:

1. Prior explanandum
   - Intentionality/unintentionality asymmetry is explanatorily prior (e. g., Hindriks)

[5] Note that the questioner cannot, in general, be sanctioned if the speaker wins the discussion, since the questioner who only questions commitments (in other words one who questions *p* without arguing for *not-p*) has not violated any rules unless they continue to question *p* after the discussion has been won. It is the one being questioned who has most of the obligations, e. g., the obligation to defend *p*. The dialectical rules govern how the obligations are divided and met in the course of the discussion.

  - Praise/blame asymmetry is explanatorily prior (e. g., early Knobe)
  - Some other asymmetry of which the praise/blame asymmetry is an epiphenomenon is explanatorily prior (e. g., Machery)
2. Internalist/externalist
  - Satisfiable by facts about the agent only (in the spirit of traditional accounts)
  - Not satisfiable by facts about the agent only

First, let us look at Knobe's position. He has since abandoned the position that the moral goodness or badness of the side-effect selects different psychological processes (a variantist position with regard to ascriptions of responsibility) and seems to have opted for conceptual revision of intentionality to include as a normative component whether some action is morally good or bad. This is an *externalist* criterion and seems to have a variantist concept of intentionality as its corollary. It is not a corollary of the superficially similar *internalist* criterion of whether the actor takes that thing to be good or bad (a psychological fact) as given, for example, in Hindriks' (2008, 638) *invariantist* account of intentional action where *S A*-s intentionally when *A* is a foreseen side-effect of *B* and *S B*-s in spite of the fact that he believes his expecting *A*-ing constitutes a normative reason against *B*-ing, and in Machery (2008) who treats the moral status as an epiphenomenon and has as his internalist criterion whether the side-effect is perceived by the actor as a cost. Considerations of cost (harming the environment is perceived by the actor as a cost but helping the environment is not a cost) would play the same role and produce the same results as Hindriks' account, without referring to anything normative such as moral badness.

Is there a case that can decide between hypotheses that take the actor's viewpoint and those that take the subject's viewpoint? Note that in the HARM condition the chairman, or at least his board, do recognize the moral badness of harming the environment, since in "it will increase our profits *but* harm the environment" the 'but' generates the implicature that what follows it is, from the speaker's point of view, a *contra* reason opposed to the *pro* reason preceding the 'but'. The alternative hypotheses seem to make the same predictions for this case.[6]

But consider the following:

> A terrorist discovers that someone has planted a bomb in a nightclub. There are lots of Americans ... who will be injured or killed if the bomb goes off. The terrorist says to himself, "Whoever planted that bomb in the nightclub did a good thing. Americans are evil! The world will be a better place when more of them are injured or dead."
>
> Later, the terrorist discovers that his only son, whom he loves dearly, is in the nightclub as well. If the bomb goes off, his son will certainly be injured or killed. The terrorist then says to himself, "The only way I can save my son is to defuse the bomb. But if I defuse the bomb, I'll be saving those evil Americans as well. ... What should I do?" After carefully considering the matter, he thinks to himself, "I know it is wrong to save Americans, but I can't rescue my son without saving those Americans as well. I guess I'll just have to defuse the bomb."
>
> He defuses the bomb, and all of the Americans are saved (Knobe 2007, 99-100).

Faced with this vignette, most subjects say that the terrorist *did not* save the Americans "intentionally" but that his saving the Americans can be explained by his reasons.

[6] Machery's hypothesis makes different predictions when the difference between the conditions is non-moral. The original studies did not show the asymmetry in non-moral cases, but these results have been contested (Machery 2008).

According to Knobe, this case creates problems for the externalist hypothesis that the *subject's* judgment of the badness of the action influences the ascription of intentionality. His reasoning seems to be that ascriptions of intentionality depend on the same psychological factors as ascriptions of reasons to the agent unless the outcome is morally bad, yet here the outcome is morally good and yet subjects are inclined to say that the agent's actions can be explained in terms of the reasons described but not that he saved the Americans intentionally. In other words, where the outcome is morally good (as perceived by subjects) then ascriptions of intentionality should track ascription of reason explanation. Thus, Knobe considers this to be a counter-example to his own hypothesis. I think that the premise Knobe relies on here is too simplistic; there can be reasons explanations without intentions. For instance, I intend to write a paper on experimental philosophy, and my having this intention explains why I wrote this paper and why there is such a paper before you. Does it explain why I wrote a *long* paper? This is a contrast question I find difficult to answer: I certainly did not set out with the intention of writing a long paper – it just turned out that way. But even if the reasons explanation of why I wrote a paper does not explain *fully* why I wrote a long paper, it does not seem to me that the explanation is false. When we *A* we bring about a host of effects and this bringing about can accurately be described as our actions *X*, *Y*, and Z. A full explanation of our *A*-ing is still a partial (arguably a so-called *straight*) explanation of our *X*-ing. So, I find it questionable whether this is really a counter-example: the terrorist has a reasons explanation for saving the Americans because he has a reasons explanation for saving his son and knew that he could not perform the one action without the other.

In contrast, this case does seem to work against hypotheses where it is how the actor perceives the moral status that is the question. Given that the actor perceived saving the Americans as a bad thing or as a cost, then subjects should say that the actor saved the Americans intentionally (consistently with the chairman in the HARM scenario) but they do not. This case does seem to decide in favor of taking the subject's point of view and variantism.

The result seems to be the expected one on Wright and Bengson's approach: given that we have judged the outcome as good, we should not judge it to have been performed intentionally or otherwise we would be forced to infer that the agent is positively responsible, i. e., praiseworthy. And we obviously do not want to consider the agent in this case as praiseworthy. They give the following formula (Wright and Bengson 2009, 27):

Good/bad action + intentionality = positively/negatively responsible actor

The formula can be used as a mathematical formula would, that is to say, it can be solved for whichever term is unknown. If the action has been judged as bad and as performed intentionally then the actor's being negatively responsible can be calculated by using the formula from left to right. On the other hand, if it is the 'intentionality' term that is missing and awaiting judgment then we can apply the formula to calculate it if we have the judgments on the action and on responsibility (and analogously when the missing judgment is whether the action is good or bad). The inference from the known values to the unknown value is non-deductive and seems to vary depending on which side of the '/' applies to the particular case. If the action is good and the actor is positively responsible (i. e., open to praise) then it can be inferred (defeasibly) that it was performed intentionally, and when the action is bad and the actor is negatively responsible (i. e., open to blame) it can be inferred (arguably more weakly) that it was performed intentionally.

Wright and Bengson exploit a perceived asymmetry in praise and blame that explains the praise/blame asymmetry. This is common to many different accounts of different types. According to McCann what we are blaming in the HARM condition is the chairman's attitude. This attitude violates a Kantian perfect duty not to harm the environment, whereas the corollary duty of helping the environment is an imperfect duty over which we have certain freedom to attend to as and when we wish. He sums up the chairman's attitude to harming the environment by saying that the chairman "means it" (McCann 2005). Similarly, Nadelhoffer stresses that what we praise or blame is in the first instance the agent and that "insofar as subjects judge that an *agent* is blameworthy, they are more inclined to say that any *negative* side effects brought about by the agent are intentional and any *positive* side effects brought about by the agent are not intentional" (Nadelhoffer 2004, 180); in both the HARM and HELP conditions the chairman is morally reprehensible because he does not take the moral considerations as motivating reasons for or against his action and this explains the intentional/unintentional asymmetry. If the agent were morally praiseworthy, says Nadelhoffer, then positive side-effects would be said to have been brought about intentionally, and this was in fact supported by empirical evidence.

The reactive commitment theory gives the same result. Assuming that the subjects in the study thought that saving Americans is a good thing, the question is whether the agent is praiseworthy in bringing it about, and this depends on what reasons the agent can give. For praise to apply those reasons must promote the moral values contained in the shared commitments by appeal to propositions contained in the shared commitments. Now, the subject would presumably consider the agent's saving his son to promote shared moral values and so praise him for that, but as for saving the Americans the best that the agent can sincerely say is that it was a side-effect of his saving his son. Because of the way that the scenario is presented the subject knows also that it was an undesired side-effect. So, whether the agent finds the side-effect undesirable (as in this case) or he simply doesn't care (as in the case of the chairman) he loses the discussion; he is not properly motivated. To put it another way, the agent was morally lucky from the subject's point of view that bad reasons led to a good result.

What does it say with regard to whether the chairman is blameworthy for harming the environment? Again, the agent cannot defend himself with the right kind of reasons and loses the discussion. But couldn't the chairman claim that he was morally *un*lucky? Subjects who give the "no praise/no blame" response probably reason something like this. But as Feltz and Cokely (2009a, 345) point out, what is often important from the group's point of view is social harmony, and this allows for an instrumental value of reactive attitudes and a looser interpretation of moral luck. Such subjects will tend to take foreknowledge as sufficient (given other conditions) for ascriptions of intentionality and respond also that the agent is blameworthy – they will give the asymmetrical "no praise/blame" and "unintentional/intentional" response. This could, in fact, be the correct decision procedure. But if the agent really is morally unlucky then although the subject would have the reactive attitude associated with blame and say in their questionnaires that the chairman was responsible this would not correspond to any reactive commitment. In the case as described, it is not clear whether the chairman is or is not the beneficiary of moral luck, but we do not need to decide this metaphysical issue in order to explain the pattern of responses: some such decisions will be correct, others will not.

What kind of explanatory hypothesis is this? It takes the praise/blame asymmetry as its explanandum and gives a single *invariantist* set of conditions for attributions of responsibility. Variation is explained by the fact that subjects may reason from different commitment sets and also have different ideas of what may defeat their obligation to sanction. It depends on facts about the judge as well as on facts about the agent, but – functionalized by the notion of a discussion governed by dialectical rules – this does not seem to lead to any problematic form of variantism.

On the connection between ascriptions of responsibility and intentionality I find Wright and Bengson's account of the inference between responsibility and intentionality attractive, and probably this can be also be given a dialectical turn. For instance, the closure rule could generate a commitment that the agent acted intentionally. A commitment, it has already been said, is not a belief, so it is not implied that either of the speakers believe that the agent acted intentionally. However, realizing that they have this commitment they may feel that it is necessary, when faced with the question, to indicate some kind of endorsement. This is a defeasible obligation and is overridden if you take having the intention, as opposed to mere foreknowledge, to be a necessary condition of ascribing intentionality. To put it another way, it could be a commitment of a type that can be retracted without sanction.

We must now consider the rest of the data. I will state my hypothesis before going through each asymmetry in turn. We have already seen in the case of the side-effect asymmetry that the man who foresees certain good or bad side-effects, but does not count those effects as reasons for acting, is lucky if those side-effects are good. My claim now is that moral luck is the common thread through all of the data.

A: The deterministic agent whose causal history is composed of only good acts is lucky.
B(i): The person who acts from emotion is lucky if his act turns out to be a good one (and presumably, that his emotion is a 'positive' one).
B(iii): The drunk motorist who gets home without hitting anything is lucky.

The issue of moral luck links all of these asymmetries. Asking a subject to praise such an act is to ask them to make a performative contradiction; it is a fallacy of many questions in that it demands a direct answer when its presupposition is not satisfied, to apply a concept that is not applicable. What we judge in such cases is not the act but, by reasoning dialogically about his or her reasons, the agent.

First, (A) the abstract/concrete asymmetry. Nahmias et al have done the experiment of describing the same deterministic background and varying whether an action is described as caused by his intentional states or by his neurophysical states, notwithstanding the fact that each are equally determined. They discovered that respondents were ready to attribute responsibility where psychological language was used but not when physical language was used. This seems to show that people are generally amenable to compatibilism between determinism and responsibility, contrary to the claims of incompatibilists, but not to the compatibilism between mechanism (the reduction of the description to a physicalist language) and responsibility. When examined more closely the folk conception of responsibility is not incompatibilist, it is argued, but only seems to be (Nahmias, Coates, and Kvaran 2007). Against this, Knobe argued that the folk conception of responsibility was incompatibilist and that what the studies (backed up with studies of his own) showed was that when presented 'concrete' cases subjects tended to exhibit compatibilist tendencies, but that when presented abstract cases subjects tended to exhibit incompatibilist tendencies. This in turn was explained by the fact that concrete cases elicited high affect which biased the psychological process; compatibilist intuitions were performance errors. But Nelkin (2007, 255–56) turns this around:

> With no other information given, people tend to assimilate determinism to coercion, but this suggestion is concealed when an intentional action is described in concrete terms. Determinism is also sometimes assimilated in people's minds to reductionism. But determinism does not preclude intentional, rational action. The concrete case avoids these faulty assimilations.

These assimilations being faulty, it is incompatibilist intuitions that are performance errors.

I will lay my cards on the table: I am an incompatibilist.[7] Granted that we may only feel ourselves coerced when it is another agent who prevents us from acting according to our will, or perhaps when there is a physical impediment to our so acting, this seems to me only an empirical fact about what we have to consider in everyday contexts and not a conceptual truth about responsibility. In the philosophical context I see no difference between our intentions being brought about by events out of our control and physical movements being brought about in the same way.

Scenarios described in physicalist language are simply not the kind of things about which making moral judgments is sensible or useful; "not responsible" in the abstract scenario here should be taken as a paraphrase of "responsibility is not an applicable concept in this scenario." In other words, asking for a moral judgment is at best infelicitous and at worst fallacious. For the same scenario described in psychological language, the agent can provide reasons and excuses of the type that the psychological description at least suggests and for that reason a dialogue where they are discussed can be imagined to take place because of which we may have a reactive attitude. But because there is no reactive commitment corresponding to this reactive attitude, this intuition (that the agent is responsible in the deterministic but non-reductionist scenario) is false.

On B(i) the emotion asymmetry Nelkin (2007, 252) claims that negative feelings can interfere with the reasons-responsive mechanism and cause us to fail to recognize the right reasons. Positive feelings, e. g., empathy, may in fact help us to see the right reasons. I find it oddly optimistic that so-called negative feelings can only hinder an accurate appraisal but positive feelings do not, and Nelkin does not seem to provide an argument for this claim. I am inclined to see this as a cultural bias – some groups may accept high emotion as a mitigating factor while others may not. This is one of those things that cannot be decided beforehand by a conceptual analysis but only emerges from the actual dialogue and what the commitment sets allow for.

The bad act seems to us worse if done calmly because the agent fails to appeal to the right values in giving reasons, and this is worse in the case where the agent has reasons and is not behaving non-rationally. On the other hand, if the presence of high emotion is among the shared commitments as an acceptable mitigating factor, then the agent may present a successful defence. Or, because of empathy, or because the agent has shown remorse and/or is unlikely to repeat the transgression, or the agent has performed the speech act of apologizing, or because the reasoner does not think he is in a position to judge (thus defeating his obligation to sanction) nobody actually blames the person even though the person is and sometimes is even explicitly recognized to be responsible.

On B(iii) the severity asymmetry Nelkin (2007, 254) comments that there cannot be moral luck. Intuitions that there must be – that attempted murder is not as blameworthy as murder – are due to under-description of the cases. If the situation is the same, the intuition should be the same. The asymmetry with regard to the negligent motorist could also be due to confusing degrees of responsibility with degrees of compensation. Thus, Nelkin seems to take the asymmetrical response to be incorrect in such a scenario.

But the negligent motorist case seems to differ from the chairman case in the following respects alone: (i) the chairman knew that the environment *would* be harmed, while the motorist knew only that some accident *might* occur, and (ii) the chairman explicitly considered, and excluded from his motives, the harm that would be caused, while the motorist might not

[7] I am also an introvert. If Feltz and Cokely (2009a; 2009b) are right there is a strong correlation between these things. In saying that the assimilation of determinism to coercion is faulty and attributing the intuitions to the fact that this is made clear in the non-reductionist scenario, Nelkin seems to be taking a compatibilist view.

have considered at the time he parked the car that an accident might result although he knew this in a *dispositional* sense. Now, the severity of the accident affects (i) since the worse the outcome the less likely you should think it able to occur. This is basically decision theory where sometimes you take a small chance because of a correspondingly large reward, or refuse to take a chance because, although you will probably win, the consequences of losing are severe. The more severe the side-effect, the more the motorist case is like the chairman case. However, in the motorist case the side-effect was not, we may suppose, foreseen, but was foreseeable. Obviously, it cannot be a moral requirement to foresee all the consequences of our actions. Whether foreseeability is considered as enough to avoid being morally unlucky depends, as before, on the extent to which social considerations play a part and the extent to which the motorist satisfices his duty to take precautions. Consideration of the latter may occur in two ways: it may be the case that the judge does not find the motorist's behaviour sanctionable and that having the brakelines checked would have been supererogatory, or it may be that the judge does find it sanctionable but defeats the sanction in virtue of some defeating obligation such as not believing himself to have the standing to sanction the motorist, as would probably occur if the subject behaved in a similar manner. On the other hand, if there are strict rules about when brakelines should be checked that the motorist has not observed then he will be held responsible.

## 3 Conclusion

An agent is *responsible for* his action if the group he is *responsible to* hold him responsible (in the sense described). An agent performs the action *intentionally* either if it is intended or if it is the foreseen side-effect of an action that is intended.[8] The empirical data can be explained by the interplay of reasons, obligations, and defeaters of obligations as regulated by dialectical rules. Attributions of blameworthiness made to agents in deterministic scenarios will be incorrect, but it is perhaps misleading to say on the contrary that they are not-blameworthy, i. e., the predicate-negation. It is better to say that blameworthiness is simply inapplicable in such cases, but there is no space to argue this point here. There is variantism in so far as attributions of intentionality can in some cases be made on the basis of attributions of responsibility as shown by Wright and Bengson, and there are different criteria for good actions and bad actions. This explains how we get from the no praise/blame response to the unintentional/intentional response. But this is a rather toothless variantism that does not lead to incoherence or to any objectionable form of circularity. Any thought that it does confuses claims about concepts with claims about their criteria of application.

Outside of conditions where the concept of responsibility is applicable, our reactive attitudes (unsurprisingly) do not imply anything about responsibility, but this does not, of course, mean that we simply cease to have them. Reactive attitudes arise out of our emotional, interpersonal interactions and cannot be reproduced by theoretical reasoning, but when the scenarios are described in concrete and psychological terms we can model them dialectically and thereby reason about them. This will tell us what reactive attitude we 'ought' to have. Our intuitions about responsibility (the 'ordinary judgments' of the Fit Assumption) are largely coherent within a particular dialogue or commitment set, but can be incoherent across dialogues or

[8] This seems to me the most perspicuous way of expressing the Single Phenomenon view. For discussion see Bratman (1984), Mele (2003) and Nadelhoffer (2006).

If this definition of "intentionally" is correct then the judgment that the terrorist did not save the Americans intentionally is incorrect – he surely did. However, the inference that he is not because he is not praiseworthy, as described by Wright and Bengson, still applies because it concerns the ascription of the term "intentionally" and not its analysis.

commitment sets. It is too quick to conclude from the empirical data that these intuitions correspond to distinct psychological processes.

# References

Bratman, M. (1984), 'Two faces of intention', *The Philosophical Review* **93** (3), 375–405.

Doris, J., Knobe, J. & Woolfolk, R. L. (2007), 'Variantism about responsibility' *Philosophical Perspectives* **2**, 183–214.

Feltz, A. & Cokely, E. T. (2009a), 'Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism', *Consciousness and Cognition* **18**, 342–350.

Feltz, A. & Cokely, E. T. (2009b), 'Individual differences, judgment biases, and theory-of-mind: Deconstructing the intentional action side effect asymmetry', *Journal of Research in Personality* **43**, 18–24.

Hindriks, F. (2008), 'Intentional action and praise-blame asymmetry', *The Philosophical Quarterly* **58** (233), 630–641.

Knobe, J. (2007), 'Reason explanation in folk psychology', *Midwest Studies in Philosophy* **31**, 90–106.

McCann, H. (2005), 'Intentional action and intending: recent empirical studies', *Philosophical Psychology* **18** (6), 737–748.

Machery, E. (2008), 'The folk concept of intentional action: philosophical and experimental issues', *Mind & Language* **23** (2), 165–189.

Mele, A. (2003), 'Intentional action: controversies, data, and core hypotheses', *Philosophical Psychology* **16** (2), 325–340.

Nadelhoffer, T. (2004), 'On praise, side-effect, and folk ascriptions of intentionality', *Journal of Theoretical and Philosophical Psychology* **24** (2), 196–213.

Nadelhoffer, T. (2006) 'On Trying to Save the Simple View', *Mind & Language* **21** (5), 565–586.

Nahmias, E., Coates, D. J. & Kvaran, T. (2007), 'Free will, responsibility and mechanism: experiments on folk intuitions', *Midwest Studies in Philosophy* **31** (1), 214–242.

Nelkin, D. (2007), 'Do we have a coherent set of intuitions about responsibility?', *Midwest Studies in Philosophy* **31** (1), 243–259.

Watson, G. (1996), 'Two faces of responsibility', *Philosophical Topics* **24**, 227–48.

Wright, J. C. & Bengson, J. (2009), 'Asymmetries in judgments of responsibility and intentional action', *Mind & Language* **24** (1), 24–50.

abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 8, Number 1, pp. 31–39, 2014

# The Theory of the Formal Discipline and the Possible Interpretations of Conditionals: Material Versus Defective Conditionals[1]

Miguel López Astorga

Universidad de Talca
Instituto de Estudios Humanísticos
milopez@utalca.cl

Keywords: conditional, defective, material, mathematics, mental models

**Abstract**

Attridge and Inglis try to check whether or not the 'Theory of Formal Discipline' is correct. This theory states that learning mathematics improves logical reasoning, and Attridge and Inglis review it by means of an experiment. Their conclusion is that, indeed, learning mathematics improves conditional inferences causing that conditionals are interpreted as defective. In this paper, I analyze Attridge and Inglis's experiment and hold that its results can be interpreted in a different way and that hence does not really prove that learning mathematics lead to defective interpretations of conditionals. Equally, the paper includes a brief reflection on how the mental models theory can explain the results achieved by Attridge and Inglis.

## 1 Introduction

Attridge and Inglis (2013) focus on the discussion, open by Plato, about whether mathematics improves reasoning. Thus they carry out an experiment and interpret that their results certainly demonstrate that idea, which is known as the 'Theory of the Formal Discipline'. However, they hold that, in addition, their results lead to other interesting findings. The improvement made by mathematics consists of understanding conditionals as defective, and not as material.

Nevertheless, Attridge and Inglis's (2013) results do not necessarily invalidate other alternative hypotheses and other interpretations of them (based on a material interpretation) are also possible. In my opinion, the problem is that Attridge and Inglis do not consider other interpretative possibilities and it is what causes that they consider that studying mathematics leads one to interpret conditionals as defective. This does not mean that the task used by them is not appropriate. In fact, it is a task that gave Evans, Clibbens, and Rood (1995) interesting results. The real problem is how Attridge and Inglis (2013) interpret the results that they obtain by means of that task. Their results certainly show that learning mathematics to some extent improves conditional reasoning (the trend to biconditionality decreases and fallacies tend to

[1] This paper is a result of the Project N. I003011, "Algoritmos adaptativos e inferencias lógicas con enunciados condicionales", supported by the Directorate for Research of Talca University, Chile (Dirección de Investigación de la Universidad de Talca, Chile). The author and lead researcher of this Project would like to thank the mentioned institutions for their help in funding this paper, and the anonymous expert that reviewed it for his (or her) comments, which, undoubtedly, improved this text.

disappear) but there are also crucial aspects involved in conditional reasoning that seem to be worse. In particular, those aspects refer to the use of the rule of *Modus Tollens* (from now on, MT), which appears not to be appropriately applied by the participants studying mathematics in Attridge and Inglis's experiment.

Maybe the problem is that Attridge and Inglis's (2013) participants are teenagers that have only had one year of advanced mathematics (post-compulsory level in England). In particular, they compare the results obtained by those participants in conditional reasoning tasks before and after completing such studies. In fact, they also compare these students' results to the results achieved by students that have not studied advanced mathematics. However, for this paper, as it can be noted below, only mathematics students' results are relevant.

I say that the participants can be the problem because it is obvious that their characteristics limit the scope of Attridge and Inglis's (2013) results. Such results are not useful for determining whether or not learning mathematics improves logical reasoning, but only for showing whether or not the mathematics post-compulsory level in England improves logical reasoning. Therefore, their conclusions and the conclusions that I will expose in this paper must be considered under this limitation. In this way, it should not be forgotten that those conclusions do not refer to learning mathematics in general, but to a concrete type of learning mathematics.

On the other hand, Attridge and Inglis (2013) also state that their results cannot explained by the mental model theory (Byrne & Johnson-Laird, 2009; Johnson-Laird, 1983, 2001, 2006, 2012; Johnson-Laird & Byrne, 2002; Johnson-Laird, Byrne, & Girotto, 2009; Orenes & Johnson-Laird, 2012). > From a semantic perspective, this theory (from now on, MMT) claims that human reasoning works by means of explicit and implicit models, but, according to Attridge and Inglis (2013), their findings are hard to explain based on it. Nevertheless, in my view, what happens is that Attridge and Inglis do not take certain essential aspects of MMT into account.

In this way, next pages will show the interpretation problem that I see in Attridge and Inglis's (2013) research, explain why their results do not demonstrate that learning mathematics is linked to a defective interpretation of conditionals, and, finally, expose the reasons why it can also be said that the semantic framework of MMT is coherent with such results. However, first of all, it seems opportune to describe in more details Attridge and Inglis's (2013) experiment.

## 2 Logical rules, fallacies, and interpretations of conditionals

Attridge and Inglis (2013) think that conditional can be interpreted in four ways. It can be taken as material, biconditional, defective, and conjunction. I briefly explain what these interpretations are (from now on, I assume that '1' stands for truth and '0' stands for falsehood):

- Material: This is the classical interpretation from logic and, as it is well-known, it establishes that $v(p \rightarrow q) = 0$ only if $v(p) = 1$ and $v(q) = 0$. Otherwise $v(p \rightarrow q) = 1$.
- Biconditional: The causes that make conditionals are considered as biconditionals have been studied in great details in the literature, and the conditional perfection phenomenon (e. g. van der Auwera, 1997a, 1997b; Geis & Zwicky, 1971; Horn, 2000; Moldovan, 2009) is especially interesting in this way. Nonetheless, what is important here is that $v(p \leftrightarrow q) = 0$ when either $v(p) = 0$ and $v(q) = 1$ or $v(p) = 1$ and $v(q) = 0$. Otherwise $v(p \leftrightarrow q) = 1$.
- Defective: This interpretation is related to probabilistic logic (e. g., Adams, 1998; Adams & Levine, 1975), but, in this paper, the basic aspect of it that needs to be considered is that, under a defective interpretation, $v(p \rightarrow q) = 1$ if $v(p) = 1$ and $v(q)$

= 1, and v(p → q) =0 if v(p) = 1 and v(q) = 0. If p is not present (i. e., ¬p happens), p → q is not a relevant relation.

- Conjunction: As it can be easily understood, under this interpretation, p → q is considered as p ∧ q, and, as it is also well-known, v(p ∧ q) = 1 when v(p) = 1 and v(q) = 1. Otherwise v(p ∧ q) = 0.

The case is that Attridge and Inglis (2013) provide equivalences between these four interpretations and two logical rules and two fallacies related to conditional. The two rules are the rule of *Modus Ponens* (from now on, MP) and MT, and the two fallacies are affirming the consequent (from now on, AC) and denying the antecedent (from now on, DA). Their equivalences are the following:

- The material interpretation only allows using MP and MT.
- The biconditional interpretation allows using MP, MT, AC, and DA.
- The defective interpretation only allows using MP.
- The conjunctive interpretation only allows using MP and AC.

Thus, taking this equivalences into account, Attridge and Inglis (2013) carried out an experiment in which, among other tasks and exercises, their participants had to solve reasoning tasks related to MP, MT, AC, and DA. In those tasks abstract conditional propositions (which referred to numbers and letters) were used as first premise. The second premise was the antecedent of the rule (MP), the denial of its consequent (MT), its consequent (AC), or the denial of its antecedent (DA). In this way, obviously, the conclusion was the consequent of the rule (MP), the denial of its antecedent (MT), its antecedent (AC), or the denial of its consequent (DA), and participants' task was to indicate, responding 'yes' or 'no', whether or not, in each of those four cases, the conclusion follows from the two premises.

As mentioned, Attridge and Inglis's (2013) participants solved those tasks twice, before and after completing a year of advanced mathematics, and the results relevant for this paper obtained by them were these: before completing the level, the participants tended to accept in large numbers the conclusions corresponding to the four inferences (a proportion between 0.7 and 0.8 in a scale from 0.2 to 0.8). Nevertheless, after completing it, they only tended to admit MP (a proportion between 0.5 and 0.6 in a scale from 0.2 to 0.8). This fact was interpreted by Attridge and Inglis (2013) as clear evidence that learning mathematics improves logical reasoning. However, in their view, the most important finding was that the improvement was related to the defective interpretation (which was that linked only to MP), and not to the material interpretation.

Nonetheless, as said above, I think that Attridge and Inglis's (2013) results do not show a global or general improvement of their participants' logical abilities. Their interpretation of their results is problematic, since it is also possible to interpret such results from an approach based on the material interpretation and to state that mathematics post-compulsory level in England only improves certain aspect of logical reasoning (it limits the tend to interpret conditional as biconditional). In my opinion, a material interpretation of such results can lead one to say that other aspects (those related to MT) are not improved by that mathematics level, and that, in a sense, they even worsen. This is explained in the next part.

## 3 The material interpretation and the rejection of MT

The main problem in the analysis of their results offered by Attridge and Inglis (2013) is that the link between the defective interpretation and the use of only MP is not obvious. When, in a conditional reasoning task, an individual responds 'yes' (that is, that the conclusion follows from the premises), his (or her) answer means that he (or she) thinks that the denial of the conclusion is not possible. The difficulty appears when he (or she) responds 'no' (that is, that the conclusion does not follow from the premises). This situation is a difficulty because we have no information on his (or her) thoughts, and hence we do not know the causes of his (or her) response. Certainly, the defective interpretation can be correct and, when the scenario refers to $\neg p$, he (or she) can respond 'no' because he (or she) considers that scenario to be irrelevant. However, we cannot be sure of that. It is also possible that the participant responds 'no' because he (or she) thinks that the conclusion is false or that the denial of the conclusion is possible. Thus, the acceptance of only MP is not necessarily linked to the defective interpretation. Therefore, it seems legitimate to assume other perspective and to interpret the answer 'no' in a different way in order to give an alternative explanation of Attridge and Inglis's (2013) results.

In my view, beyond the defective interpretation, the best alternative explanation to assume is that the participant responds 'no' because he (or she) thinks that the denial of the conclusion is possible. The answer 'no' does not imply that the individual thinks that the conclusion is necessarily false. Simply, it means that a scenario in which the denial of the conclusion is true is possible. Of course, the participant can respond 'no' because, in his (or her) opinion, the conclusion is false, but, given that we only know that he (or she) chose 'no', all we can state is that he (or she) admits the possibility that the denial of the conclusion is true, since we cannot know for sure whether or not he (or she) considers the conclusion to be false.

>From this perspective, in which the defective interpretation is not considered, Attridge and Inglis's (2013) results can have other meaning. This different meaning can be clear if we think about the logical structure of the tasks used by Attridge and Inglis (2013) and the semantic possibilities linked to both the answer 'yes' and the answer 'no'.

As far as MP is concerned, its premises are $p \rightarrow q$ and $p$, and the participant must indicate whether or not $q$ follows. If he (or she) responds 'yes', he (or she) is saying that $v(p \wedge \neg q) = 0$. However, if he (or she) responds 'no', he (or she) is not necessarily saying that $v\ (p \wedge q) = 0$. We can only be sure that he (or she) thinks that a scenario in which $v(p \wedge \neg q) = 1$ is possible.

On the other hand, the premises of MT are $p \rightarrow q$ and $\neg q$, and, in this case, the participant must decide whether or not $\neg p$ follows. If he (or she) responds 'yes', he (or she) is stating that $v(p \wedge \neg q) = 0$. However, if he (or she) responds 'no', he (or she) is not necessarily stating that $v(\neg p \wedge \neg q) = 0$. We can only be sure that he (or she) thinks that a scenario in which $v(p \wedge \neg q) = 1$ is possible.

The case of DA is similar. The premises are now $p \rightarrow q$ and $q$, and participants' task is to decide whether or not $p$ follows. If he (or she) responds 'yes', he (or she) is claiming that $v\ (\neg p \wedge q) = 0$. However, if he (or she) responds 'no', he (or she) is not necessarily claiming that $v(p \wedge q) = 0$. We can only be sure that he (or she) thinks that a scenario in which $v(\neg p \wedge q) = 1$ is possible.

Finally, in DA the premises are $p \rightarrow q$ and $\neg p$, and the participants must answer whether or not $\neg q$ follows. If he (or she) responds 'yes', he (or she) is saying that $v(\neg p \wedge q) = 0$. However, if he (or she) responds 'no', he (or she) is not necessarily saying that he (or she) thinks that $v(\neg p \wedge \neg q) = 0$. We can only be sure that he (or she) thinks that a scenario in which $v(\neg p \wedge q) = 1$ is possible.

Based on these arguments, it is correct to link biconditional to MP, MT, AC, and DA. The answer 'yes' means the following:

- In the cases of MP and MT: $v(p \wedge \neg q) = 0$.
- In the cases of AC and DA: $v(\neg p \wedge q) = 0$.

Therefore, given that $v(p \leftrightarrow q) = 1$ when $v(p \wedge q)= 1$ or $v(\neg p \wedge \neg q) = 1$, and that, by accepting MP, MT, AC, and DA, these last possibilities are not rejected, it can be said that, certainly, an individual that tends to consider those four inferences as valid is an individual that tends to interpret conditional as biconditional.

Equally, it is also appropriate to relate the material interpretation to only MP and MT. The answer 'yes' means the following:

- In the cases of MP and MT: $v(p \wedge \neg q) = 0$

But the answer 'no' means the following:

- In the cases of AC and DA: It is possible that $v(\neg p \wedge q) = 1$.

Therefore, given that $v(p \rightarrow q) = 0$ only if $v(p \wedge \neg q) = 0$, and that this possibility is precisely the only possibility that is explicitly rejected (that is in the cases of MP and MT), it can be said that an individual that tends to consider only MP and MT as valid is an individual that tends to the material interpretation of conditional.

Likewise, it is also right to link the conjunctive interpretation to only MP and AC. The answer 'yes' means:

- In the case of MP: $v(p \wedge \neg q) = 0$.
- In the case of AC: $v(\neg p \wedge q) = 0$.

And the answer 'no' means:

- In the case of MT: In principle, it is possible that $v(p \wedge \neg q) = 1$. However, given that this possibility is forbidden by MP, the answer 'no' can only mean here that $v(\neg p \wedge \neg q) = 0$.
- In the case of DA: In principle, it is possible that $v(\neg p \wedge q) = 1$. However, given that this possibility is forbidden by AC, the answer 'no' can only mean here that $v(\neg p \wedge \neg q) = 0$ as well.

Therefore, given that $v(p \wedge q) = 1$ only if $v(p) = 1$ and $v(q) = 1$, and that all other possibilities are rejected, it can be said that an individual that tends to accept only MP and AC is an individual that tends to interpret conditional as conjunction.

The problem is, as indicated, the relation of the defective interpretation with only MP. Of course, it can be thought that an individual that only accepts MP is an individual that rejects MT, AC, and DA because they are irrelevant (i. e., they refer to scenarios in which p is not). Nevertheless, given that, as commented, we do not know participants' thoughts, it can also be argued that the acceptance of only MP can also be linked to the material interpretation. I explain this idea below.

As said, in the case of MP, the answer 'yes' means that the individual thinks that $v(p \wedge \neg q) = 0$. However, in the other cases, the answer 'no' allows certain possibilities. As also mentioned, in the cases of AC and DA, $v(\neg p \wedge q) = 1$ is allowed, and this fact, evidently, is not a problem. The difficulty is given by MT. In its case, the answer 'no' means, apparently, that the participant accepts as possible a scenario in which $v(p \wedge \neg q) = 1$. Obviously, this is an inconvenience, since the combination $p \wedge \neg q$ is forbidden by MP. Of course, this inconsistency can lead one to think that the material interpretation cannot be linked to the participants that only accepted MP. The material interpretation not only requires the acceptance of MP and the rejection of AC and DA, but also the acceptance of MT. Therefore, this is the point that needs to be explained.

MT is very different from MP. MP is a simple and basic rule, but MT is complex and it cannot be used without other rule: *Reductio ad Absurdum*. If one wants to apply MT, he (or she) needs to adopt p as assumption to apply MP (considering the premise $p \rightarrow q$ and the assumption p), which leads to obtain q, and to note that the premise $\neg q$ is inconsistent with q, and that hence p is not possible. This process is far more complex than that of MP (which only needs one step) and this additional difficulty has been reported in the literature (e. g., Byrne & Johnson-Laird, 2009; López Astorga, 2013). In fact, Attridge and Inglis (2013) also refer to this issue. They state that this is the explanation of the difficulty of MT that the authors that support the idea of a mental logic often raise. They even seem to acknowledge that this explanation can be valid, even though the defective interpretation is assumed. Nonetheless, what is important is that, if this explanation is considered valid, Attridge and Inglis's (2013) results do not prove that mathematical study (at least, the kind of mathematical study corresponding to the post-compulsory level in England) causes a trend towards the defective interpretation. It is possible to interpret conditionals materially and, however, not to use MT. And this is because this last rule is hard to apply.

Thus, from this perspective, it can be said that mathematics only improves logical reasoning in a certain sense. Before the post-compulsory level, students tend to interpret conditionals as biconditionals and, according to Attridge and Inglis's (2013) results, learning mathematics corrects this problem. Nevertheless, logical reasoning ability tends to be worse in other sense after that same level, since students have difficulties to use MT. Such difficulties are not observed before the post-compulsory level because, under the biconditional interpretation, $v(\neg p \wedge \neg q) = 1$, which, apparently, must be noted to apply MT. Maybe it would be interesting to review the syllabi corresponding to the mathematics post-compulsory level. It is possible that they do not include the resolution of problems with processes similar to the application of MT, and that this fact is the cause that mathematics students do not often use this rule after the post-compulsory level.

## 4 MMT and learning mathematics

The main aim of this paper is not to propose arguments in favor of MMT. My more important goal is only show that Attridge and Inglis's (2013) results do not necessarily mean that mathematical study lead to a defective interpretation of conditionals. However, given that Attridge and Inglis (2013) state that their results are difficult to understand from the basic theses of MMT and that I think that that idea is not correct, it can be opportune to explain briefly, before concluding, why, in my view, Attridge and Inglis's (2013) research does not cause difficulties to MMT.

MMT is a wide theory on human reasoning. Nevertheless, what is relevant for this paper is only its explanation about conditionals. According to MMT, when an individual reasons about conditional propositions, he (or she) considers the possibilities, or models, corresponding to each conditional. Thus, a proposition such as $p \rightarrow q$ has three models:

A.– $p \,\&\, q$

B.– $\neg p \,\&\, q$

C.– $\neg p \,\&\, \neg q$

The key point is that only A is an immediate and explicit model. Both B and C need certain cognitive effort and the action of memory for becoming explicit models. In this way, faced to a conditional, individuals firstly take only one model (A) into account. B and C, as indicated, are models more difficult to consider.

Attridge and Inglis (2013) think that these theses are not coherent with their results. If their participants interpreted conditionals as biconditionals before mathematical advanced level, it means that they considered two models in that moment. In particular, the considered models were A (explicit) and C (implicit). A allowed the application of MP and AC, and C allowed the application of MT and DA. The problem is that, as I understand Attridge and Inglis's arguments, it is not easy for MMT to explain what happens at the later point in time, i. e., the acceptance of only MP. Certainly, it could be argued that C is eliminated after learning mathematics, but, if A is the only model that is considered, MP should not be the only rule that is accepted. AC should be accepted as well. This is, at least in my view, what Attridge and Inglis (2013) seem to mean as regards MMT.

However, I think that arguments such as these are not correct. A relevant datum that must be considered is that, after studying mathematics, a greater tendency to accept MP and AC (i. e., to the conjunctive interpretation) was also observed, which means that mathematical advanced study is somehow linked to the rejection of the biconditional interpretation and the acceptance of the conjunctive interpretation, that is, to the rejection of the model C and the acceptance of only A (the explicit model). Nevertheless, because what was most striking was that the mathematics post-compulsory level leaded the participants to admit only MP, it is obvious that MMP needs to clarify this last fact. In this way, I can say that, in my opinion, it is true that A and C are the models that must be taken into account if the interpretation is biconditional, but the acceptance of only MP can be explained, from the perspective of MMT, as a change of models. Indeed, if the considered models are not A and C, but A and B, MP must be accepted (by virtue of A), MT must be rejected (C is not considered), AC must also be rejected (by virtue of B), and DA must be rejected as well (by virtue of B too). Thus, based on MMT, it can be stated that learning mathematics also improves logical reasoning in a sense, since it causes that conditional relations are not interpreted as biconditional relations (the model B is added). Nonetheless, it is worse in another sense, because the model C is lost and hence it is not possible to accept MT.

Of course, this explanation proposed by me can be questioned by referring to certain theses of MMT. For example, Johnson-Laird and Byrne (2002) state that individuals only have two options: they can consider the explicit model (A) or all the models (A, B, and C). Thus, it does not seem possible that the participants only consider A and B (without C). However, as far as this difficulty is concerned, I can say that my arguments are only an interpretation of MMT that respects their main and basic theses and that, at the same time, is consistent with Attridge and Inglis's (2013) results. After all, MMT allows some models to be blocked in certain circumstances and, according to it, individuals only represent what they believe to be true. In this way, it can be thought that my explanation develops ideas that, although they are not clearly held by the adherents of MMT, do not contradict the core of this theory.

Thus, if my interpretation of MMT is right, it can be said that it continues to be clear that mathematics post-compulsory level in England corrects the mistake of interpreting conditionals as biconditionals. Equally, it can be stated that, from the perspective of MMT, the problem is also MT, since it seems that proofs related to the opposite of what must be proved (which are linked to C) are not practiced by students in post-compulsory level.

## 5 Conclusions

It is hence obvious that, if we do not assume the defective interpretation, it can be said that mathematical study, or mathematics post-compulsory level in England, only improves logical reasoning in a sense: $p \rightarrow q$ is not interpreted as $p \leftrightarrow q$. The ability of use MT is not improved.

This fact does not mean that the theory of the formal discipline can only be considered true if the defective interpretation is accepted. Maybe the problem is that the kind of mathematics taught in the post-compulsory level is not the type required for improving the use of some logical rules (especially MT). In addition, other possibility is that the post-compulsory level is not enough and that more mathematics levels are necessary for a better use of such rules.

In any case, there is no doubt that Attridge and Inglis's (2013) results do not conclusively prove that learning mathematics lead one to a defective interpretation of conditionals. As shown above, such results are coherent with the idea that their participants continue to interpret conditionals materially after the post-compulsory level. Furthermore, as also commented, MMT is also compatible with their results and can explain them.

Obviously, I cannot deny that the defective interpretation continues to be a possibility that must be taken into account. Undoubtedly, although Attridge and Inglis's (2013) results are consistent with a material interpretation and MMT, it is evident that they are also consistent with the defective interpretation proposed by Attridge and Inglis (2013). In this way, it can only be said that, if mathematical study leads to a defective interpretation of conditionals, that fact needs to be proved by means of further research. So far, we have no proofs that, certainly, that is the case. The discussion is open and other interpretations are also possible.

## References

Adams, E. W. (1998), *A Primer of Probability Logic*, Center for the Study of Language and Information Publications, Stanford.

Adams, E. W. & Levine, H. P. (1975), 'On the uncertainties transmitted from premises to conclusions in deductive inferences', *Synthese* **30**, 429–460.

Attridge, N. & Inglis, M. (2013), 'Advanced mathematical study and the development of conditional reasoning skills' *PLoS ONE* **8** (7), 269–399. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0069399 (March 11, 2014).

Auwera, J. van der (1997a), 'Pragmatics in the last quarter century: The case of conditional perfection', *Journal of Pragmatics* **27**, 261–274.

Auwera, J. van der (1997b), Conditional perfection, *in* A. Athanasiadou & R. Dirven, eds., 'On Conditional Again', John Benjamins, Amsterdam, pp. 169–190).

Byrne, R. M. J. & Johnson-Laird, P. N. (2009), '"If" and the problem of conditional reasoning', *Trends in Cognitive Science* **13**, 282–287.

Evans, J. St. B. T., Clibbens, J., & Rood, B. (1995), 'Bias in conditionals inference. Implications for mental models and mental logic', *The Quarterly Journal of Experimental Psychology* **48**, 644–670.

Geis, M. L. & Zwicky, A. M. (1971), 'On invited inferences', *Linguistic Inquiry* **2**, 561–566.

Horn, L. R. (2000), 'From if to iff: Conditional perfection as pragmatics strengthening', *Journal of Pragmatics* **32**, 289–326.

Johnson-Laird, P. N. (1983), *Mental Models. Towards a Cognitive Science on Language, Inference and Consciousness*, Harvard University Press, Cambridge, MA.

Johnson-Laird, P. N. (2001), 'Mental models and deduction', *Trends in Cognitive Science* **5**, 434–442.

Johnson-Laird, P. N. (2006), *How We Reason*, Oxford University Press, Oxford.

Johnson-Laird, P. N. (2012), Inference with mental models, *in* K. J. Holyoak & R. G. Morrison, eds., 'The Oxford Handbook of Thinking and Reasoning', Oxford University Press, New York, pp. 134–145.

Johnson-Laird, P. N. & Byrne, R. M. J. (2002), 'Conditionals: A theory of meaning, pragmatics, and inference', *Psychological Review* **109**, 646–678.

Johnson-Laird, P. N., Byrne, R. M. J., & Girotto, V. (2009), 'The mental models theory of conditionals: A reply to Guy Politzer', *Topoi* **28**, 75–80.

López Astorga, M. (2013), 'Are conditional and disjunction really comparable?', *Universum* **28** (2), 229–242.

Moldovan, A. (2009), 'Pragmatic considerations in the interpretation of denying the antecedent', *Informal Logic* **29** (3), 309–326.

Orenes, I. & Johnson-Laird, P. N. (2012), 'Logic, models and paradoxical inferences', *Mind & Language* **27** (4), 357–377.

abstracta *Linguagem, Mente & Ação* ISSN 1807-9792 Volume 8, Number 1, pp. 41–53, 2014

# Uma análise fregeana de expressões para eventos e resultados dos eventos

Ana Clara Polakof

PUC-Rio/Bolsista CNPq de doutorado – Brasil
anaclarapo@gmail.com

**Abstract**

This paper makes a Fregean analysis of certain expressions which, according to the way they are constructed, can refer, firstly, to events (*A tradução da Bíblia por João* demorou dois anos) and, secondly, to the results of these events (*A tradução da Bíblia de João* está sob a mesa). This paper intends to show that – if we take into account the above mentioned expressions and use Fregean notions such as meaning, reference, function, argument and concept – it is possible to consider events as objects. To do this we need to articulate developments made both from a linguistic perspective and a philosophical one. In this particular case, we take into account Fregean semantics; we try to analyze philosophically the linguistic phenomena called nominalization and arrive to interesting conclusions about the semantic behavior of the expressions that contain event and result nominalizations and ontological status of their respective referents.

## Introdução

Frege é um filósofo complexo e muitos têm dedicado parte de suas vidas para estudá-lo, e também para refutá-lo.[1] Este artigo não se centra no entendimento de Frege, mas usa os recursos que ele nos deu para fazer distinções ontológicas e semânticas entre as entidades às quais referem as expressões que contêm nominalizacões de evento e de resultado (objeto que resulta do evento). Isto implica usar o arsenal fregeano, mais especificamente as noções[2] de *sentido, referência, conceito, função* e *argumento* que são especialmente úteis no que diz respeito ao entendimento do relacionamento entre a linguagem e o mundo e, neste caso particular, das nominalizações de evento e resultado (do evento) e seus referentes.

As noções antes nomeadas só admitem elucidação e são de difícil entendimento.[3] Contudo, e isso tem que ser aclarado, não julgaremos as complexidades, nem faremos uma valoração das noções que Frege desenvolve, nem como positivas nem como negativas. Ele é utilizado em nossa análise porque sua proposta pode nos ajudar a compreender melhor as diferenças entre as nominalizações de evento e de resultado (do evento). Isso é devido, por um lado, a sua

[1] Cf. Chateaubriand (2001: capítulos 1 e 2). Para um estudo detalhado da teoria da linguagem em Frege pode-se consultar, Dummett (1981). Para uma discussão clássica a respeito da distinção entre sentido e referencia, pode-se consultar, Russell (1905).

[2] Utilizamos o termo *noção* para que ele não seja confundido com o termo fregeano *conceito*, ainda que, com o termo *noção*, designemos o que costuma ser chamado de *conceito*.

[3] Contudo, Frege as fez mais claras. A noção de *conceito* é fundamental em dois ensaios que são trabalhados nesse artigo, "Function and concept" (Frege, 1891) e "On Concept and Object" (Frege, 1892a). Neles, Frege desenvolve e explicita essa noção. Nesses trabalhos é possível encontrar uma definição de *conceito* de acordo com o gênero (função), a diferença específica (função que tem sempre como valor um valor de verdade), e que dá conta de seu caráter predicativo.

semântica bidimensional que tem em conta a diferença entre sentido e referência e, por outro, as noções que Frege utiliza que permitem avançarmos na distinção entre essas nominalizações, assim como no reconhecimento de que tanto os eventos como os objetos resultantes desses eventos são objetos (no sentido fregeano) no mundo, como mostraremos.

Este trabalho tem a seguinte organização: em primeiro lugar, introduziremos as diferenças sintáticas e linguísticas entre as nominalizações de evento e resultado (do evento); em segundo lugar, tentaremos mostrar que é possível trabalhar com o problema das nominalizações de evento e resultado (do evento) desde uma perspectiva fregeana e mostraremos quais são os problemas ou as objeções que podem se nos apresentar ao tentar elucidar o comportamento dessas nominalizações; em terceiro lugar, tentaremos mostrar que é possível, a partir das noções de função, argumento e conceito chegar a distintas conclusões sobre o comportamento semântico das nominalizações de evento e resultado (do evento); e em quarto e último lugar, tentaremos vincular o comportamento semântico com o estatuto ontológico dos entes aos que referem essas nominalizações.

## 1 Breve análise linguística das nominalizações

As nominalizações de evento e de resultado (do evento) com as quais trabalhamos neste artigo pertencem ao que foi chamado tradicionalmente *nomes deverbais* (de verbos), isto é, que foram formados a partir de verbos (por exemplo, *ampliação* deriva de *ampliar, construção* deriva de *construir*). Esses nomes têm sido amplamente estudados pelos linguistas, desde o âmbito da morfologia léxica[4], mas também têm sido estudados pelas propriedades sintáticas particulares que eles apresentam.[5] Essas nominalizações têm a peculiaridade de que expressam valores que são associados aos verbos, *isto é,* as nominalizações de evento expressam um evento e as de resultado, o resultado desse evento. Esses valores podem se encontrar em estruturas oracionais como é possível observar na construção *Os pedreiros ampliam a obra* na qual estamos diante do processo; enquanto que numa oração como *Os pedreiros ampliaram a obra*, é possível observar o término, a finalização do processo, seu resultado. Esse comportamento é próprio de verbos de *realização* que permitem expressar, no predicado, eventos delimitados[6] e que contêm no seu significado tanto o processo como o estado ou objeto que resulta desse processo. Embora os dois tipos de nominalizações tenham sido estudadas, as de resultado têm sido deixadas de lado na análise linguística. Isto é devido ao fato de que, como elas têm um comportamento sintático semelhante ao dos nomes que não se formam a partir de verbos, muitos linguistas consideraram que elas deveriam ter estruturas semelhantes a estes últimos e não aos seus pares eventivos.[7]

Nesta seção tentaremos fazer uma delimitação sintática clara entre os dois tipos de nominalizações e as expressões[8] que elas integram, para que os dados analisados a seguir, a partir de Frege, sejam claros. É importante estabelecer que a distinção feita entre essas duas nominalizações tem sido feita desde sempre, pois elas têm comportamentos diferentes, mas apresentam a mesma estrutura morfológica. Isto pode gerar ambiguidade linguística que pode ser evitada tendo em conta fatores linguísticos que explicaremos na continuação.

[4] Almela Pérez (1999); Lacuesta y Bustos Gisbert (1999); Varela (1990); entre outros.

[5] Grimshaw (1990); Picallo (1999); Resnik (2010); entre outros.

[6] *Cf.* De Miguel (1999: 3019).

[7] A maioria da bibliografia citada neste capitulo será hispânica, porque nossa pesquisa foi feita para o espanhol. Nesse artigo, para esclarecer, usaremos só nominalizações do português e nos centraremos só naquelas questões que ajudem na diferenciação entre as nominalizações, para que seja clara a proposta que fazemos a partir do Frege. Porém, deve ser aclarado que este tipo de nominalizações existem em línguas tao diversas como o grego, o espanhol, o inglês, o mocoví, só por nomear algumas.

[8] No sentido fregeano.

Semanticamente falando, a distinção é clara: as nominalizações de evento expressam o evento, o processo, e as nominalizações de resultado expressam o resultado desse evento que, no caso dos verbos de realização que nos interessam, tende a concordar com um objeto dado, como, por exemplo, uma tradução, uma construção, uma ampliação. Sintaticamente falando, a distinção também é clara: num caso se combinam com certos predicados, artigos, adjetivos, e no outro caso, com outros. É nessa diferença sintática que vamos nos concentrar. Não vamos discutir a estrutura morfológica das mesmas, isto pode ser visto em Polakof (2013). Nesse artigo queremos apenas esclarecer quais são as distinções para que as propostas que fazemos desde uma perspectiva fregeana não sejam mal interpretadas por questões sintáticas. Embora nas próximas seções nos centraremos nas diferenças sintácticas entre as nominalizações, devemos sempre saber que no primeiro caso estamos analisando aquelas nominalizações que podem integrar expressões que referem a eventos (*a tradução por João da Bíblia está sendo feita*) e que no segundo caso estamos analisando nominalizações que integram expressões que referem ao resultado desses eventos (*a tradução de João da Biblia está sobre a mesa*).

### 1.1 As nominalizações de evento

As nominalizações de evento foram caracterizadas há várias décadas, desde antes do trabalho de Chomsky (1970). O trabalho de Grimshaw (1990) sempre é tomado como ponto de partida, desde uma perspectiva gerativa. Nele fica demonstrado que o significado eventivo desse tipo de nominalizações fica evidenciado nas combinações sintáticas nas quais elas podem participar e que elas admitem no inglês. Basear-nos-emos na proposta da Grimshaw (1990), e as de Picallo (1999) e Resnik (2010) para propor as seguintes características das nominalizações de evento:[9]

(1) As nominalizações de evento

(a) nomeiam um processo ou evento e contêm os participantes do evento que são introduzidos pelas preposições *por* (agente) e *de* (paciente): *a tradução por João da Bíblia*

(b) podem combinar-se com nomes como *processo*

i. *O processo de absolvição demorou pouco tempo*

ii. *O processo da tradução da Bíblia por João foi longo*

(c) aceitam modificação aspectual (adjetival e não adverbial)

i. *A constante designação de problemas pelo professor molesta os alunos*

ii. *A destruição total da cidade em dois dias surpreendeu a todos.*

(d) só podem estar determinadas por artigos definidos no singular, mas podem não ter determinante

i. A/**uma tradução da Bíblia por João demorou dois anos*[10]

(e) Aceitam um controlador do evento

i. *A tradução da Bíblia para divulgar as ideias católicas foi um êxito*

(f) Podem ser o argumento do verbo presenciar

i. *Os assistentes presenciaram a construção da ponte pelos pedreiros*

Existem mais provas sintáticas que permitem estabelecer o comportamento das nominalizações de evento, mas como isso é somente uma introdução ao comportamento sintático que elas têm, achamos que é suficiente para poder ver a diferença linguística entre estas nominalizações e ver o significado eventivo que elas possuem. Estudaremos agora as nominalizações de resultado.

[9] Não entraremos nas dificuldades que essas propostas têm, só tomaremos alguns aspectos delas para mostrar claramente seu comportamento sintático. Para uma análise detalhada destas questões e diferenças, ver Polakof (2013).

[10] O asterisco (*) marca a agramaticalidade da expressão ou da oração.

### 1.2 As nominalizações de resultado

Como já mencionamos, estas nominalizações não têm sido muito estudadas na tradição linguística. Somente é nos últimos 20 anos que elas têm tomado alguma importância nas pesquisas linguísticas, sobretudo nas de corte gerativo. Então, faremos o mesmo que fizemos com as nominalizações de evento e as caracterizaremos brevemente segundo seu comportamento sintático-semântico:

(2) As nominalizações de resultado

α nomeiam o resultado do processo ou o objeto resultado do processo e contêm os participantes do evento que são introduzidos pelas preposições *de* (agente) e *de* (paciente): *a tradução de João da Bíblia*[11]

β Não podem combinar-se com nomes como *processo*

i. **O processo da tradução está bem escrito*[12]

χ Não aceitam modificação aspectual

i. **A constante tradução está sob a mesa*

δ Podem estar determinadas por qualquer artigo e qualquer determinante

i. *Eles estudaram a/uma/aquela/essa tradução da Bíblia de João*

ε Não aceitam um controlador do evento

i. **As traduções da Bíblia de João para divulgar as ideias católicas foramum êxito*[13]

φ Não podem ser o argumento do verbo presenciar

i. **Os assistentes presenciaram uma tradução*

Estas são as características que permitem-nos fazer uma comparação entre estas nominalizações e as de evento desde una perspectiva linguística, que será trabalhada desde um análise fregeano.

### 1.3 Comparações dos comportamentos sintáticos das duas nominalizações

Podemos então ver que as nominalizações de evento e resultado, embora compartilhem os morfemas que as compõem, têm comportamentos bem diferenciados e é com esses comportamentos bem diferenciados que vamos trabalhar. A partir de agora quando falamos dessas nominalizações sempre temos em conta essas diferenças. Trabalharemos com a impossibilidade que as nominalizações de evento possuem de aceitar artigos indefinidos, contra a possibilidade que as de resultado possuem de se combinarem tantos com artigos definidos como com indefinidos. Também usaremos as distintas combinações sintáticas que se podem estabelecer, como mostraremos logo.

## 2 Sentido e referência das expressões para evento e para resultado do evento

Nesta seção analisamos as expressões determinadas que são, para Frege, as que podem referir a objetos porque são para ele, de fato, nomes próprios. Isso implica que trabalhamos com

[11] Não confundir "de João" com o possuidor da tradução, pois é ele o agente da ação: foi ele quem traduziu a Bíblia.

[12] Essa oração é agramatical porque só o resultado do processo pode estar bem escrito, portanto não é possível fazer a combinação que se encontra nesse exemplo.

[13] Neste caso é importante ter em conta todo o contexto sintático, porque alguém poderia interpretar algo como *umas Bíblias para divulgar as ideias católicas* como gramatical. Mas nesse caso deveríamos interpretar uma elisão de uma subordinada do tipo *umas Bíblias (que foram feitas) para divulgar as ideias católicas* e essa não é a interpretação que procuramos.

as expressões com significado de evento e expressões com significado de resultado e não com as nominalizações (palavras) isoladas porque precisamos de construções que possam funcionar como nomes próprios e, também, como veremos, como funções. Então, vamos trabalhar com expressões como *a tradução da Bíblia por João* e *a tradução da Bíblia de João* que podem referir a um evento e a um objeto-resultado respectivamente.[14]

Devemos trabalhar brevemente com as noções de sentido e referência que permitem estabelecer uma semântica de duas dimensões. Mesmo que no desenvolvimento que Frege faz dessas noções é central a noção de igualdade,[15] queremos nos centrar no simples fato de que é possível que dois sentidos diferentes refiram a entidades diferentes sempre e quando não formem parte de uma relação de igualdade. Isto é o que acontece com as nominalizações de evento e resultado que estão no interior de expressões determinadas como as que já mostramos. Embora as expressões sejam muito similares, a preposição *por* que introduz o agente da ação naquela que tem como núcleo a nominalização de evento, diferencia seu sentido do sentido da expressão que contém a de resultado com o agente introduzido pela preposição *de*. Temos, então, duas expressões que se comportam como nomes próprios –de acordo com Frege[16]– que são similares, mas diferentes, e têm, nesse caso, referentes diferentes: num caso é o evento (pode ser *a tradução da Bíblia por João* que está sendo feita por João) e no outro caso é o objeto que resulta do evento (pode ser *a tradução da Bíblia de João* que foi feita por João).

Respeitante a esse tema, alguém poderia se perguntar se é suficiente só com uma diferença na preposição que introduz o agente afirmar que temos sentidos diferentes. Acreditamos que isso é assim, que usarmos a linguagem natural com todas as suas imperfeições e limitações é condição suficiente para que uma única preposição dê lugar a diferentes sentidos e, nesse caso em particular, a diferentes referentes. Se temos em conta outros elementos sintáticos que não sejam as combinações com distintos tipos de artigos, a diferença entre elas fica ainda mais clara. O fato de que elas sejam sujeitos de orações diferentes é uma mostra adicional de que elas devem ter referentes diferentes. Se comparamos *a tradução por Pedro da Bíblia está sendo feita* com *a tradução de Pedro da Bíblia está sob a mesa*, é muito simples ver que, tendo em conta tudo o que analisamos, no primeiro caso estamos fazendo referência ao evento e no segundo ao objeto-resultado.

Devemos, uma vez feitas as diferenças, aclarar o que é um objeto para Frege e por que podemos estabelecer tão facilmente que tanto os eventos como os objetos-resultados são objetos nessa ontologia. Frege tem uma noção ampla de objeto. Para ele, toda expressão que não tenha um espaço vazio, isto é, toda expressão que esteja saturada está por um objeto.[17] Então, tudo aquilo que possa funcionar como sujeito de uma oração terá um referente que será um objeto. Isto implica que tanto os eventos como os objetos-resultado são objetos, o qual é uma simplificação para o problema sobre a existência de eventos no mundo. Numa análise fregeana, não existiria nenhum problema em estabelecer que os eventos são objetos e que os objetos que resultam desses eventos também o são. Por essa razão, ela é particularmente atrativa para o tipo de perspectiva que tentamos defender e mostra que é possível utilizar a semântica de Frege para trabalhar com as expressões que contêm as nominalizações de evento e resultado.

[14] Não trabalharemos com a nominalização *tradução* que não tem a possibilidade de referir, além de que ela pode ser considerada um termo geral, pois ela não está determinada, e não é relevante nesta análise.

[15] Na introdução de "On sense and reference" (Frege, 1892b), é claramente estabelecido que a noção de igualdade é de importância inegável para conseguir entender a relevância que tem a possibilidade de usar distintos sentidos para um mesmo referente, mas não é isso o que analisaremos nesse artigo.

[16] Frege (1892b).

[17] Ver Frege (1891: 32)

Um leitor ávido poderia se perguntar qual é o problema nas expressões que contêm as nominalizações, se elas têm sentidos diferentes, referências diferentes e estão por objetos diferentes. Essa pergunta, que nos fizemos ao descobrir o anteriormente analisado, apresenta-nos um problema que tentaremos resolver tendo em conta as noções de função, conceito e argumento. Argumentaremos que a semântica/analítica fregeana pode ser usada para explicar a diferença existente entre o comportamento semântico das expressões que contêm nominalizações de evento e de resultado e que, considerar as noções antes mencionadas de função, conceito e argumento, permite-nos ir além do que seria possível se só trabalhássemos com as noções de sentido e referencia.

# 3 Função, conceito, argumento nas expressões para evento e resultado do evento

As noções de *função*, *argumento* e *conceito* são essenciais para poder diferenciar entre as expressões que contêm as nominalizações de evento e resultado. Nesta seção desenvolveremos essas noções e mostraremos qual é a importância que elas têm para poder cumprir o nosso objetivo. Em primeiro lugar, tentaremos dar uma solução a partir das noções de *função* e *argumento* que, em conjunção, são uma unidade completa, e em segundo lugar, trabalharemos com as noções de *conceito* e *palavra-conceito* a partir das quais consideramos que será possível chegar a uma resposta diferente às anteriormente dadas.

## 3.1 Função e argumento nas expressões para evento e resultado do evento

As noções de função e argumento que Frege propõe em “Function and Concept” têm mostrado ser um caminho possível para o entendimento das expressões que contêm nominalizações. Para entender como isso é possível, é necessário destacar que, mesmo que o desenvolvimento fundamental dessas noções seja feito por Frege para as matemáticas, ele mesmo afirma que essas noções podem se estender até a linguagem e é nessa extensão que basear-nos-emos.[18]

Com respeito à função, é necessário defini-la, relacionada à linguagem, como uma expressão que está insaturada e deve ser completada pelo argumento que tem autonomia própria. Frege se centra nas orações, mas ele mostra que é possível que uma expressão determinada seja tomada como uma função com o seu argumento (como em, por exemplo, *A capital da Inglaterra*, onde *Inglaterra* é o argumento e *A capital de X* a função). Esta noção permite-nos associar as expressões definidas que contêm as nominalizações de evento e resultado com a função e o argumento, pois, como temos visto, elas estão providas de participantes que podem ser, em termos fregeanos, argumentos de uma função.

As nominalizações poderiam ser pensadas, devido a sua formação a partir de verbos, como funções que têm dois lugares insaturados para argumentos, ou como funções que contêm o paciente da ação e devem ser completadas por um argumento que será o agente da ação –como no caso das orações, nas que Frege toma como função todo o predicado e como lugar para o argumento a posição do sujeito. Este agente se corresponderia, se transformássemos a nominalização determinada numa oração, com o sujeito da oração. Isto é, em vez de propor uma função com dois argumentos insaturados, propomos uma função com um argumento insaturado da mesma maneira que Frege propôs para a oração. Ou seja, não vamos propor uma função do tipo *X traduz Y*, com dois argumentos, mas vamos propor uma função do tipo *X traduz a Bíblia* que transportaremos para a análise das nominalizações.

Como mencionamos na seção anterior, as expressões determinadas nas quais as nominalizações de evento e resultado aparecem, mesmo se são similares, têm sentidos diferentes e referentes

[18] “We shall not stop at equations and inequalities. The linguistic form of equations is a statement” (Frege, 1891: 31).

diferentes. Isto deve evidenciar-se na função porque, se tivéssemos funções iguais, ao somarmos o argumento teríamos valores iguais e, como já mostramos, isto não é assim. Portanto, teremos funções diferentes se a soma entre estas e o argumento tiver como resultado referentes diferentes. Podemos, então, argumentar que teremos duas funções diferentes quando modificamos a preposição, o que implica que, quando um argumento que tem igual referência se soma a essa função, a soma entre a função e o argumento terá como resultado um referente diferente, como pode-se observar em:

a) *A tradução da Bíblia por X*
b) *A tradução da Bíblia de X*

Se X é *João*, no primeiro caso teremos uma função que ao somarmos ao valor do argumento X *João* terá como resultado um referente eventivo: o evento de traduzir João a Bíblia; no segundo caso o valor que a função toma terá um referente de objeto-resultado: o objeto que resulta da tradução que o João fez da Bíblia. Nesse sentido, não haveria nenhuma contradição com o que foi dito sobre o sentido e a referência na seção anterior. Contudo, ao separar as expressões em função e argumento, obtivemos um melhor entendimento do seu funcionamento e uma fundamentação mais profunda de por que num caso há referência a um evento e no outro caso a um objeto-resultado: porque no primeiro caso temos uma função que uma vez saturada por um argumento referirá a um evento, enquanto no segundo caso temos uma função que uma vez saturada pelo argumento referirá a um objeto-resultado.

Isto mostra, novamente, que estamos frente a objetos, pois se transformam em expressões saturadas uma vez que a função toma como argumento *João*. Por essa razão, eles podem funcionar como sujeito e ter como referente algum tipo de objeto. Mostramos que, a partir da análise em função e argumento, é possível avançar a um melhor entendimento de por que as expressões que contêm as nominalizações têm distintos comportamentos semânticos e distintos referentes; mas ainda não logramos explicar por que eles são diferentes. É necessário, então, usar outras noções que permitam-nos aproximar a essa diferencia. Na próxima seção mostraremos que se temos em conta a noção de *conceito* e a de *palavra-conceito* será possível mostrar essas diferencias.

### 3.2 O conceito e as expressões para evento e resultado do evento

O conceito, que pode ser entendido como um tipo especial de função que tem sempre como valor resultante um valor de verdade,[19] permite-nos acercar a uma diferença no comportamento das expressões que contêm as nominalizações de evento e resultado. Esse conceito que, numa oração, será o referente do predicado deverá diferenciar-se do objeto que nunca será o referente do predicado, pois, como vimos, é o referente do sujeito. É possível avançar, a partir da diferença entre objeto e conceito, à uma solução que nos ajude a entender que, de uma maneira ou outra, os eventos e os objetos-resultado têm comportamentos diferentes.

Mostramos que nas expressões que contêm as nominalizações de evento e as que contêm as de resultado estamos frente a distintas funções que uma vez saturadas com o mesmo argumento (refere ao mesmo) têm como resultado um valor diferente. Isto pode ser previsto desde uma proposta fregeana sobre funções e argumentos. Por sua vez, estabelecemos que mesmo se isso nos permite afirmar que tanto eventos como objetos-resultado são objetos no sentido amplo usado por Frege, não podemos estabelecer nenhum tipo de diferença no funcionamento que eles têm.

[19] Cf. Frege (1891).

Temos deixado claro, como Frege também o fez,[20] que o conceito e o objeto são referentes diferentes: estão no mundo, mas não tem o mesmo comportamento. O primeiro é o referente do predicado[21], o segundo o do sujeito. É, também, importante ter em conta que Frege faz uma clara distinção entre o que expressa uma igualdade e o que expressa uma afirmação. No primeiro caso, estamos frente a uma oração que mediante um verbo copulativo como *ser* expressa uma relação de igualdade (por exemplo, *O gato que come lasanha é Garfield*). Isto é, estabelece que a parte na esquerda da oração é igual à parte na direita do verbo. Nesse sentido, estamos frente a dois sentidos diferentes que têm o mesmo referente que nos permitem chegar a um conhecimento real do mundo (segundo Frege). No segundo caso, estamos frente a uma oração que nos permite construir uma afirmação (*O gato que come lasanha é laranja e preto*). Neste caso, se estabelece uma oração com sujeito e predicado na qual o verbo copulativo é –segundo Frege–[22] um "mere verbal sign of predication". Aqui, é possível afirmar que o objeto referido por *O gato que come lasanha* cai sob o conceito referido por *laranja e preto*. Enquanto no primeiro caso estamos frente a uma relação reversível (não importa qual dos nomes próprios esteja em primeiro ou em segundo lugar), no segundo caso estamos frente a uma relação irreversível (o intercâmbio de posições dos integrantes da oração faria que as pessoas construíssem uma oração agramatical, como em **laranja e preto é o gato que come lasanha*).[23] É possível estabelecer, então, que só no segundo caso estamos frente a uma construção que está constituída por um sujeito e um predicado no sentido fregeano.

Continuaremos aprofundando estas diferenças que nos permitirão estabelecer o fato de que as nominalizações de evento (ou, melhor, as expressões que as contêm) comportam-se semanticamente diferente às de resultado. O reconhecimento dessa diferença semântica permitirá nos aproximar a uma diferença ontológica entre os objetos que são referidos por essas expressões, algo que o mesmo Frege (1892a) faz quando ele distingue entre conceito e objeto como sendo os referentes do sujeito, por um lado, e do predicado, do outro.

Frege faz uma distinção entre *nomes próprios* e *palavras-conceito*. Os primeiros podem ser os chamados nomes próprios (como *João*), mas também expressões determinadas por um artigo definido (como *a tradução por João da Bíblia*), pois, de igual maneira como o fazem os nomes próprios, elas referem a objetos. As segundas, que são introduzidas em "On Concept and Object" (Frege, 1892a), são substantivos que estão determinados por um artigo indefinido (como pode ser *um gato que come lasanha*). Por essa mesma razão, não podem referir a objetos, mas podem fazer parte de um predicado e referir, no conjunto do predicado, a um conceito. Essa distinção permitirá nos fazer uso da demarcação proposta por Frege entre o que só pode ser objeto e todo o restante.[24]

É possível, depois de ter estabelecido as diferentes possibilidades de leitura de uma oração que tem um verbo copulativo como *ser* e a diferença entre os nomes próprios e as palavras-conceito, fazer a análise das estruturas que nos interessam seguindo um esquema fregeano que tentaremos estabelecer na continuação. Com respeito à relação de igualdade, como já mencionada, encontramos que tanto as expressões que incluem as nominalizações de evento como as que incluem as nominalizações de resultado podem funcionar como nomes próprios, o que reafirma o que temos visto até o momento, como se pode observar em:

[20] Cf. Frege (1892a).

[21] Sobre os problemas para caracterizar o *conceito* em Frege pode se ver Sluga (1980).

[22] Frege (1892a: 43), ainda que essa ideia possa ser rastreada, ao menos, até o Platão (Ackrill, 1997: 82–83).

[23] É necessário excluir qualquer leitura estilística que seja possível fazer a partir desta oração. Desde uma perspectiva fregeana, entre outros, esses dois adjetivos jamais poderiam ser sujeitos de uma oração e esta perspectiva é a única que importa nesse trabalho.

[24] Frege (1892a: 44).

c) *O evento é a tradução por João da Bíblia/ A tradução por Pedro da Bíblia é o evento*
d) *O objeto-resultado é a tradução de João da Bíblia/ A tradução de João da Bíblia é o objeto-resultado.*

Esses exemplos indicam que, mesmo se é possível alterar a ordem dos integrantes das sentenças c y d pois são relações de igualdade, não é possível alterar os integrantes de c com os integrantes de d pois eles não são iguais. Eles não tem os mesmos referentes e, portanto, intercambiar tais integrantes só teria como resultado uma alteração no valor de verdade das relações e as transformaria em algo falso (é falso, e agramatical, que **O objeto resultado é a tradução por João da Bíblia*). Podemos, então, seguir afirmando que, na noção ampla de objeto que o Frege usa, os dois tipos de entidades são objetos e como tais cada um deles pode ser referido mediante o uso de distintos sentidos que permitem-nos apreender conhecimento real do mundo, como já foi mencionado.

Tudo parece indicar que não há nada de novo a dizer, analisamos o sentido de uma maneira diferente mas obtivemos a mesma análise. Porém, uma vez que consideramos as orações que são afirmações, isto é, que estão constituídas por um sujeito e um predicado, teremos uma análise diferente que começará a ser observado a partir de e e f:

e) *A tradução por João da Bíblia é um evento*
f) *A tradução de João da Bíblia é um objeto-resultado*

Em e y f, onde as expressões continuam como sujeito da afirmação, não temos problemas de nenhum tipo. Mas, a análise muda quando estas expressões passam ao lado direito da oração, o tipicamente predicativo, como expressões indefinidas. Em g y h as temos como expressões definida e em i e j como indefinidas:

g) *O evento é a tradução por João da Bíblia*
h) *O objeto-resultado é a tradução de João da Bíblia*
i) **O evento é uma tradução por João da Bíblia*[25]
j) *O objeto-resultado é uma tradução de João da Bíblia*

Em g e h, mostramos que, para ter uma análise diferente, não é suficiente mudar o lugar da expressão, pois nesses casos estamos frente a uma relação de igualdade entre dois sentidos que podem referir a um mesmo objeto, como vimos em c e d. Devemos transformar a expressão definida em uma indefinida para que o problema surja, como em i e j. É possível observar que i não é uma oração admissível porque ela é, de fato, agramatical, enquanto j é admissível e gramatical. Estes exemplos mostram que, embora a nominalização de resultado possa integrar tanto um nome próprio, desde a perspectiva fregeana, como uma palavra-conceito; a nominalização de evento só pode integrar um nome próprio e não pode ser transformada em palavra-conceito. Isto indica que, ademais de referir a objetos diferentes, as nominalizações têm comportamentos semânticos diferentes, o que é um resultado interessante. Por um lado, temos uma nominalização que, nas fórmulas estabelecidas por Frege e estendidas neste artigo, só pode funcionar como sujeito. Por outro lado, temos uma nominalização que pode funcionar como sujeito e como palavra-conceito. Desta maneira, ela referirá a um objeto quando integra um nome próprio e poderá referir quando esteja dentro do predicado, com o verbo *ser,* a um conceito. Temos, então, no primeiro caso uma entidade que só pode ser objeto: o evento de realização referido do pela expressão que inclui *tradução* e seus participantes; e no segundo caso temos algo que pode ser introduzido no grupo de “todo o restante” de que Frege fala, pois a nominalização

[25] Porque, como vimos, as nominalizações de evento não podem estar determinadas por um artigo indefinido.

de resultado com seus participantes, como vimos, pode funcionar tanto como sujeito como parte do predicado. Aprofundamos nessas diferenças, que até agora tem ficado no âmbito de semântica, no próximo capitulo.

# 4 Diferenças ontológicas entre as expressões que contêm as nominalizações de evento e resultado desde uma perspectiva fregeana

Mostramos que é possível trabalhar com as nominalizações de evento e resultado desde uma perspectiva fregeana e fazer um verdadeiro aporte ao seu entendimento. À princípio, não parecia ser rentável trabalhar com esta aproximação devido ao fato de que não parecia haver nenhum conflito, pois simplesmente nos deixava estabelecer que tanto os eventos como os objetos-resultado nomeados pelas nominalizações eram objetos. Esta aproximação foi útil uma vez que logramos nos introduzir numa semântica na qual o sentido e a referência são essenciais e numa ontologia na qual há, ademais de objetos, funções.

O fato de que as expressões que contêm nominalizações de resultado possam funcionar como palavras-conceito, que possam funcionar como parte de um predicado numa afirmação mostra que elas têm um comportamento semântico diferente que o das expressões que incluem as nominalizações de evento. Elas têm um comportamento similar ao dos nomes comuns, como *planeta*, que também podem se comportar como palavras-conceito; coisa que não acontece com as expressões para eventos. Portanto, é possível observar que, mesmo numa semântica que privilegia o sentido e a referência, é possível reconhecer um funcionamento diferenciado entre os sentidos que temos para referir aos eventos e entre os sentidos que temos para referir aos objetos-resultado.

Estas nominalizações compartilham uma estrutura similar na qual a mudança de uma única preposição nos permite reconhecer não apenas um sentido diferente, mas também uma estruturação diferente.[26] São diferentes funções que, ao tomarem o mesmo argumento, têm referentes diferentes. A introdução da noção de conceito, que é puramente predicativa, permite-nos estabelecer que, como as nominalizações de evento não podem estar determinadas por um artigo indefinido, as expressões eventivas não podem formar predicados com o verbo *ser* e, portanto, não referem nunca a conceitos. Isto é, elas não podem ser transformadas em palavras-conceitos, como pode acontecer com um nome próprio como *Viena*.

Para que as diferenças sejam mais evidentes, voltamos à ideia de palavra-conceito junto com a possibilidade de usar as nominalizações de resultado dessa maneira e a impossibilidade de usar as de evento de mesma maneira. Retomemos os exemplos seguintes:

k) *O objeto-resultado é uma tradução de João da Bíblia*
h) **O evento é uma tradução por João da Bíblia*

A palavra-conceito que está determinada pelo artigo indefinido, no exemplo k, está sendo usada de uma maneira que busca especificar as propriedades que tem o sujeito da oração. Isto é, o sujeito tem as características de ser 'uma tradução de João da Bíblia' e não, por exemplo, 'uma mesa'. Desta maneira, seu comportamento é semelhante ao que vimos no exemplo *O gato que come lasanha é laranja e preto*, no qual *ser laranja e preto* pode indicar a(s) 'propriedade/es do gato que come lasanha'. Algo similar ocorre com *uma tradução de João da Bíblia*: é uma 'propriedade do objeto-resultado' que ajuda na caracterização do sujeito da oração e que poderia

[26] Esta diferença faz-se mais evidente em casos de línguas como o Mocoví que apresentam estruturas morfológicas diferentes para as nominalizações de evento e de resultado (cf. Carrió 2009), o que evidenciaria mais a diferença entre as funções que deveriam ter as expressões para evento e resultado.

ajudar, também, no reconhecimento do objeto no mundo: não é qualquer objeto-resultado, é aquele que é uma tradução de João da Bíblia.

Isto não acontece no caso de h. Por mais que alguém quisesse afirmar alguma coisa das propriedades do evento, resulta claro –pela agramaticalidade da oração– que não é possível caracterizar semanticamente um evento com outro evento. Isto é, se tivéssemos o desejo de enunciar as propriedades do objeto referido por *o evento*, deveríamos pensar na sua duração ou sua localização, mas não é possível caracterizá-lo com outro evento. Poderíamos formar uma afirmação como *o evento é divertido*, mas não uma como a que está escrita em h. Isto pode ser visto como evidencia de que os eventos devem ser diferentes que os resultados, mesmo se eles podem ser nomeados por nomes com sentidos similares, não só não têm a mesma referência mas eles não têm o mesmo comportamento semântico. O fato de que as expressos eventivas formadas a partir de nominalizações de realização não possam se comportar como palavras-conceito e, portanto, não possam ser usadas para caracterizar outro objeto-evento permite-nos argumentar que esse comportamento semântico é um reflexo do fato de que o evento é único e irrepetível.

Este desenvolvimento nos permite afirmar que os objetos que são referidos pelas expressões que incluem as nominalizações de resultado são semelhantes aos dos nomes comuns. Isto implica que é possível transformar uma expressão como *a tradução de João da Bíblia* numa palavra-conceito como *uma tradução de João da Bíblia* porque, embora o objeto-resultado possa ser diferente de outras traduções, ele tem certas características que são próprias de todas as traduções o que é refletido no seu comportamento semântico. O mesmo ocorre, por exemplo, se tentamos caracterizar *o cão de meu vezinho* como *um cão ruidoso*, pois ele compartilha com o resto dos cães ruidosos a propriedade de ser ruidoso. Contudo, isto não ocorre com os eventos que são nomeados por nominalizações a partir de verbos de realização. Se tomarmos a ideia de Davidson (2001 [1981]) de que os eventos são entidades necessárias para dar conta do mundo, que são irrepetíveis e verdadeiros particulares, poderíamos explicar a impossibilidade de que sejam transformados em palavras-conceito devido a que o evento referido por *a tradução por João da Bíblia* é um evento único e irrepetível –como mencionamos– e, portanto, não é possível usá-lo para caracterizar nenhum outro evento.[27] Isto, por sua vez, mostra que a única maneira na qual essa expressão pode ser usada, mediante o verbo copulativo *ser*, na parte direita da oração é quando temos uma relação de igualdade, de semelhança, na qual temos um mesmo referente com sentidos diferentes. Esses nos permitiriam chegar a um conhecimento real como, por exemplo, pode ser reconhecido que *a tradução por João da Bíblia* é o mesmo que *o evento de traduzir João a Bíblia*.

## 5 Conclusões

Podemos concluir que é possível, desde uma análise fregeana, estabelecer uma diferença no comportamento das expressões que incluem as nominalizações de evento e resultado. Embora a peculiaridade dessa diferença não se estabeleça através do sentido e da referência, pois, como vimos, elas têm sentidos e referentes diferentes, é possível estabelecer que –ademais de referir a entidades diferentes– elas não têm o mesmo comportamento semântico. As nominalizações de resultado a partir de verbos de realização, nas estruturas corretas, poderão referir tanto a

[27] Tomemos nota de que isto não sucede com o nome *evento*. Este pode ser usado como nome próprio e como palavra-conceito. Isto é devido ao fato de que, diferentemente da tradução que deve ser complementado pelo seu agente e paciente, o nome *evento* não necessita desses complementos, razão pela qual se comporta, desde esta perspectiva, como um nome comum. Nesse sentido e devido à sua generalidade, ele nos permite caracterizar a todos os eventos, porque ele é o termo geral que inclui a todos, além da particularidade e da irrepetibilidade que os respectivos eventos formados a partir de verbos de realização possam ter.

objetos-resultado como a conceitos; enquanto que as de evento só podem referir a eventos e nunca a conceitos.

Essa perspectiva nos permite estabelecer as diferenças antes mencionadas e nos permite reconhecer certas características especiais dos eventos que coincidem com as que são propostas na ontologia de Davidson. Isto é, é possível defender –a partir de uma análise que parte do arsenal fregeano– que os eventos são entidades únicas e irrepetíveis, que são verdadeiros particulares, pois eles não apresentam características que podam ser usadas para caracterizar outros eventos; o que é refletido no fato de que as expressões eventivas não podem ser transformadas em palavras-conceito. Ademais, neste artigo argumentamos que o comportamento das nominalizações de resultado, com respeito ao sentido, à referência e ao conceito, assemelha-se ao resto dos objetos, o que era esperável, pois, ontologicamente falando, são objetos como os demais objetos.

Terminamos este trabalho afirmando que, neste caso em particular, se pensamos nas condições suficientes e necessárias para que estas entidades existam, podemos afirmar que dado que os eventos e os objetos-resultado são referidos por expressões definidas saturadas, ambos cumprem com as condições necessárias e suficientes para serem objetos na análise fregeana. Por fim, podemos concluir que os eventos que são nomeados pelas expressões de evento a partir de verbos de realização e os objetos que resultam desses eventos que são nomeados pelas expressões de resultado são objetos no mundo e que têm estatutos ontológicos diferentes que refletem-se na diferença no comportamento semântico.

## Bibliografia

Ackrill, J. L. (1997), *Essays on Plato and Aristotle*, Oxford University Press, New York.

Almela Pérez, R. (1999), *Procedimientos de formación de palabras en español*, Ariel, Barcelona.

Black, M. & Geach, P., eds. (1960), *Translations from the Philosophical Writings of Gottlob Frege*, Basil Blackwell, Oxford.

Bosque, I. & Demonte, V., eds. (1999), *Gramática Descriptiva de la Lengua Española*, Espasa, Madrid.

Carrió, C. (2009), *Mirada Generativa a la Lengua Mocoví (Familia Guaycurú)*, Tesis presentada para aspirar al grado de Doctor en Letras, Universidad Nacional de Córdoba.

Chateaubriand, O. (2001), *Logical forms. Part 1. Truth and description.*, Campinas, Coleção CLE.

Chomsky, N. (1970), Remarks on nominalizations, *in* R. Jacobs & P. Rosenbaum, 'Readings in English Transformational Grammar', Ginn and Company, Waltham, MA, pp. 184–221.

Davidson, D. (2001 [1981]). *Essays on Actions and Events*, Oxford University Press, New York.

De Miguel, E. (1999), El aspecto léxico, *in* Bosque & Demonte, eds. (1999), pp. 2977–3060.

Dummett, M (1981), *Frege: Philosophy of Language*, Harvard University Press, Cambridge, MA.

Frege, G. (1891), Function and Concept, *in* Black & Geach (1960), pp. 21–41.

Frege, G. (1892a), On Concept and Object, *in* Black & Geach (1960), pp. 42–55.

Frege, G. (1892b), On Sense and Reference, *in* Black & Geach (1960), pp. 56–68.

Grimshaw, J. (1990), *Argument Structure*, MIT Press, Cambridge, MA.

Lacuesta, R. & Bustos Gisbert, E. (1999), La derivación nominal, *in* Bosque & Demonte, eds. (1999), pp. 4505–4594.

Picallo, M. C. (1999), La estructura del sintagma nominal: las nominalizaciones y otros sustantivos con complementos argumentales, *in* Bosque & Demonte, eds. (1999), pp. 365–393.

Polakof, A. (2013), ‘La estructura funcional de las nominalizaciones deverbales de evento y resultado a partir de verbos de realización’, *Anuari de Filologia. Estudis de Lingüística* **3**, 113–144.

Resnik, G. (2010), *Los nombres eventivos no deverbales en español*, Tesis de doctorado, Universitat Pompeu Fabra, Barcelona.

Russell, B. (1905), ‘On Denoting’, *Mind* **14**, 479–493

Sluga, H. (1980), *Gottlob Frege*, Routledge and Kegan Paul, London and Henley.

abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 8, Number 1, pp. 55–73, 2014

# An Inferentialist Approach to Paraconsistency

James Trafford

University for the Creative Arts at Epsom
jtrafford2@ucreative.ac.uk

**Abstract**
This paper develops and motivates a paraconsistent approach to semantic paradox from within a modest inferentialist framework. I begin from the bilateralist theory developed by Greg Restall, which uses constraints on assertions and denials to motivate a multiple-conclusion sequent calculus for classical logic, and, via which, classical semantics can be determined. I then use the addition of a transparent truth-predicate to motivate an intermediate speech-act. On this approach, a liar-like sentence should be "weakly asserted", involving a commitment to the sentence and its negation, without rejecting the sentence. From this, I develop a proof-theory, which both determines a typical paraconsistent model theory, and also gives us a nice way to understand classical recapture.

## Introduction

This paper develops and motivates a paraconsistent approach to semantic paradox from within a modest inferentialist framework. There are many different forms of inferentialism. By modest inferentialism, I will mean a view on which a specified set of inference rules are taken to determine the truth-conditional content of logical constants, and where those rules have a substantive connection with ordinary inferential practices (Belnap and Massey, 1990; Boghossian, 2003; Garson, 2010; Peacocke, 1986a,b, 1987). Much of the existing work on paraconsistent logic emphasizes the construction of many-valued semantic consequence (Beall, 2013; Priest, 2006, 2008). By expanding upon the bilateralist theory of inferentialism, the aim is to develop a paraconsistent proof-theory that itself determines the model theory.

In §1, I outline the bilateralist framework developed by Greg Restall (Restall, 2005, 2009), which uses constraints on assertions and denials to motivate a multiple-conclusion sequent calculus for classical logic. I also show how classical semantics can be determined by the calculus. §2 introduces a transparent truth-predicate into the calculus, which, following advocates of non-classical logic (e.g Parsons 1984; Priest 2006), is taken to motivate a non-primitive attitude intermediate between assertion and denial. On this approach, a liar-like sentence should be "weakly asserted", involving a commitment to the sentence and its negation, without rejecting the sentence. §4 outlines a corresponding model-theory (broadly this is Beall's (2013) $LP^{+}$), before showing that any ordinary sequent calculus fails to completely determine that semantics. To deal with this, in §5, I outline a 3-sided proof-theory, which both determines the semantics $LP^{+}$, and also gives us a nice way to understand classical recapture in limit cases. Whilst there are a few novel technical results scattered throughout, the real novelty lies in the philosophical setting for a generalized program of paraconsistent modest inferentialism.

# 1 Modest inferentialism

This section briefly sketches the bilateralist approach to inferentialism. The modest version of inferentialism that we are interested in here allows that the meanings of logical constants in a language $\mathscr{L}$ are explained in terms of which inferences are valid in $\mathscr{L}$. §1.1 introduces the position and uses it to motivate a multiple-conclusion sequent calculus for classical logic. In a way that will be specified in §1.2, these inferences can be said to determine classical model-theory. Finally, in §1.3, the notion of absoluteness is introduced to characterize when a model-theory is completely determined by a sequent calculi.

## 1.1 Motivating classical sequent calculus

Logic tells us something about the way that agents' rational commitments are combined and constrained over arguments. For example, an agent asserting $\alpha, \beta \vdash \alpha \wedge \beta$, may be said to be rationally committed to not simultaneously asserting $\alpha$, $\beta$ and denying $\alpha \wedge \beta$. Note that this way of putting things is deliberately inequivalent to saying that the agent asserting $\alpha$, $\beta$ is thereby rationally committed to asserting $\alpha \wedge \beta$. This is because, whilst such rational commitments play a key role in the fixation of beliefs, they neither commit an agent to logical omniscience, nor do they rationally oblige an agent to assert $\alpha \wedge \beta$ where, for example $\alpha \wedge \beta$ is unreasonable given the antecedent context of assertion. So, logic, on this view, constrains an agent's commitments by saying that it is rationally prohibitive to deny $\alpha \wedge \beta$, given the assertion of $\alpha$, $\beta$.

Importantly, this view takes logical consequence to tell us not just about assertion, but about both assertion and denial, and the connection between the two. Restall's (2005; 2009) suggestion is to think of logical consequence as governing positions involving asserted and denied statements.

**Definition 1.** (Position) A position $[\Gamma : \Delta]$ is a pair of sets of formulae where $\Gamma$ is the set of asserted formulas, and $\Delta$ the set of denied formulas.

A position expressed in a language may be used to represent an agent's rational commitments in terms of the coherence between assertions and denials. Where $[\Gamma : \Delta]$ is a position, we allow that $[\Gamma, \alpha : \Delta, \beta]$ is the state adding the formula $\alpha$ to the left set $\Gamma$, and $\beta$ to $\Delta$. Think of the above coherence constraints over rational commitment as saying that, a position $[\Gamma : \Delta]$ is incoherent if it contains some formula in both the left set and the right set, so that $\Gamma \cap \Delta \neq \emptyset$. Thinking of this in terms of an agent, such a position indicates that some statement is both asserted and denied, and so incoherent. Incoherence allows us to characterize sequent provability.

**Definition 2.** (Sequent provability) If $[\alpha : \beta]$ is incoherent, then $\alpha \vdash \beta$.

This follows because, if a position consisting of asserting $\alpha$ and denying $\beta$ is incoherent, then $\alpha \vdash \beta$, and an agent who asserts $\alpha$ and denies $\beta$, as we said above, has made a mistake.

The definition generalizes in cases involving sets of assertions and denials. In a multiple-conclusion (SET-SET) framework, $\Gamma \vdash \Delta$ may be read in terms of the underlying atomic formulae $\{\alpha_1, \ldots, \alpha_n\} \vdash \{\beta_1, \ldots, \beta_n\}$, which is (classically) equivalent to $\{\alpha_1 \wedge \alpha_2 \ldots . \wedge \alpha_n\} \rightarrow \{\beta_1 \vee \beta_2 \ldots . \vee \beta_n\}$. Then, any position $[\Gamma : \Delta]$ for which an agent who asserts each member of $\Gamma$, and denies each member of $\Delta$ is incoherent. In that case, $\Gamma \vdash \Delta$, and an agent is mistaken to assert all $\alpha \in \Gamma$ and deny all $\beta \in \Delta$.

**Definition 3.** (Sequent provability generalized) If $[\Gamma : \Delta]$ is incoherent, then $\Gamma \vdash \Delta$.

The general idea is to build-up a sequent calculus out of these constraints over assertions and denials. First, consider the addition of structural constraints. Since both asserting and denying the same formula is incoherent, from $[\Gamma, \alpha : \Delta, \alpha]$ and Definition 3, we have the usual identity axiom for sequent-calculi.

$$\alpha \vdash \alpha \text{ for all } \alpha \text{ (Identity)}$$

We also have weakening, since, if a position is incoherent, the addition of assertions or denials will not bring it back to a coherent position. Contra-positively, if $[\Gamma : \Delta]$ is coherent, and $\Gamma' \subseteq \Gamma$ and $\Delta' \subseteq \Delta$, then $[\Gamma' : \Delta']$ will be coherent. This gives us:

$$\frac{\Gamma \vdash \Delta}{\Gamma, \Gamma' \vdash \Delta, \Delta'} \text{ (Weakening)}$$

Think of extensibility constraints on assertions and denials. For a position $[\Gamma : \Delta]$ is coherent, if the positions $[\Gamma, \alpha : \Delta]$ or $[\Gamma : \Delta, \alpha]$ are incoherent, then the original position $[\Gamma : \Delta]$ must already be incoherent. In other words, if a position is coherent, it should be extensible by a formula $\alpha$ to a coherent position where $\alpha$ is either asserted or $\alpha$ is denied. So, where $[\Gamma : \Delta]$ is coherent, either $[\Gamma, \alpha : \Delta]$ or $[\Gamma : \Delta, \alpha]$ is coherent. This gives us:

$$\frac{\Gamma, \alpha \vdash \Delta \qquad \Gamma \vdash \alpha, \Delta}{\Gamma \vdash \Delta} \text{ (Cut)}$$

Operational rules for the connectives can also be constructed fairly naturally out of positions. For example, if the position $[\Gamma : \Delta, \alpha \wedge \beta]$ is coherent, then $[\Gamma : \Delta, \alpha]$, $[\Gamma : \Delta, \beta]$, or both, are coherent. Contra-positively, if $[\Gamma : \Delta, \alpha]$ and $[\Gamma : \Delta, B]$ are incoherent, then so to is $[\Gamma : \Delta, \alpha \wedge \beta]$. In this case, we know that $\Gamma \vdash \Delta, \alpha$, and $\Gamma \vdash \Delta, \beta$, so that $\Gamma \vdash \Delta, \alpha \wedge \beta$. This gives us:

$$\frac{\Gamma, \alpha, \beta \vdash \Delta}{\Gamma, \alpha \wedge \beta \vdash \Delta} \wedge\text{-L} \qquad \frac{\Gamma \vdash \Delta, \alpha \qquad \Gamma \vdash \Delta, \beta}{\Gamma \vdash \Delta, \alpha \wedge \beta} \wedge\text{-R}$$

Keep in mind that, on this bilateralist approach, the meanings of connectives are built up from the primitive speech acts of both assertion and denial, in contrast to the assumption that denial of $\alpha$ is simply the assertion of $\neg\alpha$. We construct the rules for classical negation by taking a negation $\neg\alpha$ to be assertible when $\alpha$ is deniable, and vice-versa. So, if $[\Gamma : \Delta, \alpha]$ is incoherent, then so too is $[\Gamma, \alpha : \Delta]$. This gives us Gentzen's classical negation rules:

$$\frac{\Gamma \vdash \alpha, \Delta}{\Gamma, \neg\alpha \vdash \Delta} \text{ } (\neg\text{-L}) \qquad \frac{\Gamma, \alpha \vdash \Delta}{\Gamma \vdash \neg\alpha, \Delta} \text{ } (\neg\text{-R})$$

Analogous accounts can be provided for all of the classical sequent rules (Restall, 2005). This gives us a construction of the classical sequent rules in multiple-conclusion form, which is built out of a simple and plausible account of agents' rational commitments.

## 1.2 A determination theory

Given that we are working toward a modest inferentialism, we can think of these rules as also determining the truth-conditional content of the connectives. The general idea is to let the assertion and denial conditions governing a logical connective determine a meaning for that connective when the rules completely determine truth-conditional content. This view echoes that suggested in Peacocke (1986b), with the general requirement that:

> **General requirement**: The given rules of inference, together with an account of how the contribution to truth-conditions made by a logical constant is determined from those rules of inference, fixes the correct contribution to the truth-conditions of sentences containing the constant (Peacocke, 1993, p.172).

What follows draws heavily upon Hardegree (2005); Hjortland (2014); Humberstone (2011), and the details developed in Trafford (2014). In providing a recipe for the interaction between a sequent calculi and truth-conditions we should not talk of the truth-value of a sequent. Nonetheless, we may understand a valuation as providing a counterexample, or not, to the potential validity of a sequent in terms of whether or not truth is preserved when passing from *l.h.s* formulae to right. The idea is to let incoherent positions, and therefore, provable sequents, determine a set of valuations from the universe of possible valuations, $U$, over a language $\mathscr{L}$. In this way, a logic induced by a proof-system can be said to determine a valuation-space $V \subseteq U$ on $\mathscr{L}$, defined as any subset of the set of all possible valuations. First, define a logic as consisting of all provable sequent inferences.

**Definition 4.** For a set of formulae $S$ in a language $\mathscr{L}$, a multiple-conclusion sequent is an ordered pair, Γ, Δ, (where $\Gamma \cup \Delta \in WFF$, and where Γ, Δ are sets of formulae of $S$). A multiple-conclusion logic $L$ is an ordered pair $\langle S, L \rangle$, where $L$ is the set of binary relations $\vdash_L$ between finite subsets of $S$ and finite subsets of $S$. A rule, $R^{\#}$, defined for a logic $L$ consists of a set of sequent premises and a set of sequent conclusions $\{SEQ_P\} \rightarrow \{SEQ_C\}$. We call the set of provable sequent in $L$, $L$-valid, such that $\Gamma \vdash \Delta =_{df} \{\langle \Gamma, \Delta \rangle$ is $L$-valid$\}$.

**Definition 5.** $\mathcal{V}$ is a set of truth-values, and $\mathcal{D} \subseteq \mathcal{V}$ designated values. A valuation $v$ is a function on $\mathscr{L}$ assigning a truth-value $\in \mathcal{V}$ to a formula in $S$ where $v : S \rightarrow \{\mathcal{V}\}$. Classically, we have $\mathcal{V} = \{1, 0\}$, and $\mathcal{D} = \{1\}$.

Before defining a relation between the two, first, note that (Definition 3) an incoherent position $[\Gamma : \Delta]$ allows that $\Gamma \vdash \Delta$, which we understood as saying that an agent is mistaken to assert all $\alpha \in \Gamma$ and reject all $\beta \in \Delta$. Equivalently, $[\Gamma : \Delta]$ is incoherent just in case some $\alpha \in \Gamma$ is denied, or some $\beta \in \Delta$ is asserted. This, and given that we read $\Gamma \vdash \Delta$ as (classically) equivalent to $\{\alpha_1 \wedge \alpha_2 \ldots . \wedge \alpha_n\} \rightarrow \{\beta_1 \vee \beta_2 \ldots . \vee \beta_n\}$, gives us that $\Gamma \vdash \Delta$ is valid just in case some $\alpha \in \Gamma$ is denied, or some $\beta \in \Delta$ is asserted.

We use this to define a relationship between sequents and valuations in terms of *satisfaction*.

**Definition 6.** A sequent $\Gamma \vdash \Delta$ is *satisfied* by a valuation $v$ just in case $v(\alpha) = 0$ for some $\alpha \in \Gamma$, or $v(\beta) = 1$ for some $\alpha \in \Delta$, otherwise $v$ *refutes* the argument.

Because we have Cut, this is easily extensible to every formula in $\mathscr{L}$.

**Example 7.** For a sequent $\Gamma, \alpha \vdash \Delta$ to be valid, we know that each $v \in V$ has $v(\alpha) = 0$ for some $\alpha \in \Gamma \cup \alpha$, or $v(\beta) = 1$ for some $\beta \in \Delta$. Similarly, for $\Gamma \vdash \Delta, \alpha$ to be valid, each $v \in V$ has $v(\alpha) = 0$ for some $\alpha \in \Gamma$, or $v(\beta) = 1$ for some $\beta \in \Delta \cup \alpha$.

Weakening plays the role of ensuring that each formula $\alpha \in S$ appears on either the *l.h.s* or the *r.h.s* of a sequent. Cut ensures that no formula $\alpha \in S$ appears on both. Think of this in terms of valuations. Then Identity tells us that all formulas $\alpha \in S$ takes either 1 or 0, and Cut tells us that no formula takes both. In effect, cut allows for a partition of formulas into those that take the value 1, and those that take 0.

Then, let $L$ determine a valuation-space by determining the set of admissible valuations $V$ that are consistent with $L$.

**Definition 8.** ($L$-consistency) A valuation $v \in U$ is $L$-consistent iff $v$ satisfies each provable sequent in $L$.

**Theorem 9.** *A logic $L$ determines a valuation-space $V$ which consists of the set of valuations that are consistent with every provable sequent in $L$.*

*Proof.* A valuation space $V \subseteq U$ is determined by $L$ when each $v \in U$ is $L$-consistent, so $\mathbb{V}(L) =_{df} \{v \in U : v \text{ is } L\text{-consistent}\}$. Then $V$ consists of the set of valuations that are consistent with every provable sequent in $L$ by definition. □

### 1.3 Completeness

We are not quite home and dry here, since, whilst Theorem 9 may hold for a logic, it is possible that a logic fails to determine a unique valuation space. In which case, we will not have achieved a modest inferentialist account for that logic, since the meanings of the connectives will, to some degree, be underdetermined. As is well known (Belnap and Massey, 1990; Carnap, 1943; Dunn and Hardegree, 2001; Garson, 2010; Hardegree, 2005; Hjortland, 2014; Humberstone, 2011; Shoesmith and Smiley, 1978), single-conclusion formulations of classical logic fail in this way.

We clarify this by noting that a valuation-space may also be defined semantically. On this view, classically speaking, we expect $V_{CPL}$ to be induced by the truth-conditional interpretation of each classical connective. Then $V_{CPL}$ contains those valuations $v : S \to \{1, 0\}$ that obey the truth-conditional clauses for the connectives of classical propositional logic. We can then use this valuation-space to determine a logic.

**Definition 10.** ($V$-validity) An sequent $\langle\Gamma, \Delta\rangle$ is $V$-valid iff, for all $v \in V$, $v$ satisfies $\langle\Gamma, \Delta\rangle$.

**Theorem 11.** *Starting from a valuation-space, $V$ determines a logic $L$ w.r.t $V$ where all the $V$-valid arguments are $L$-valid.*

*Proof.* Let the set of all $V$-valid arguments constitute the logic $L$, which is now determined by $V$: $\mathbb{L}(V) =_{df} \{\langle\Gamma, \alpha\rangle \in L : \langle\Gamma, \alpha\rangle \text{ is } V\text{-valid}\}$. All arguments valid on $V$ are $\mathbb{L}(V)$-valid by definition. □

So, it is possible to transition between a valuation-space and a logic, and vice-versa.[1] This allows us to define an abstract completeness theorem which, following Dunn and Hardegree (2001), I call *absoluteness*.

**Fact 12.** *(Hardegree, 2005) For any $L$, $V$;*

- $L$ is absolute iff $L = \mathbb{L}(\mathbb{V}(L))$
- $V$ is absolute iff $V = \mathbb{V}(\mathbb{L}(V))$

When absoluteness holds, we have a guarantee that the determining relationship between $L$, $V$ is complete. We know (Theorem 11) that a semantic structure $\langle S, V\rangle$ is determined by an inferential structure $\langle S, L\rangle$ when the set of $L$-valid sequents are such that $V$ comprises the set of valuations consistent with each sequent in $L$. Absoluteness on $V$ tells us that $V \subseteq U$ is the only valuation-space consistent with $\langle S, L\rangle$, and absoluteness on $L$ tells us that $L$ is the only set of sequents that can be associated with $\langle S, V\rangle$. So, absoluteness provides a standard by which to analyze the determining relationship between an inferential structure and a semantic structure.

**Theorem 13.** *(General determination theory) A semantic structure $\langle S, V\rangle$ is completely determined by an inferential structure $\langle S, L\rangle$ when $L$, $V$ are absolute.*

[1] Details can be found in appendix 1.

We can utilize absoluteness to show that multiple-conclusion classical sequent logic completely determines the classical semantics.

**Definition 14.** ($V$-consistency) A valuation $v$ is $V$-consistent when $v$ satisfies every $V$-valid argument (Hardegree, 2005).

**Lemma 15.** *By the definition of absoluteness, $V$ is absolute iff $V$ contains every $V$-consistent valuation.*

*Proof.* See appendix 1. □

**Theorem 16.** *For any $V \subseteq U$, $V = \mathbb{V}(\mathbb{L}(V))$ (in the SET-SET framework).*

*Proof.* See appendix 1. □

An immediate corollary is that the bilateralist framework gives us a modest inferentialist account of classical logic where the valuation-space determined by the classical proof-theory uniquely determines the valuation-space $V_{CPL}$. Note that absoluteness does not hold for a classical proof-theory in single-conclusion format, since the valuation-space determined by the connectives $\neg, \vee$ is compatible with valuations $\notin V_{CPL}$.[2]

# 2 Adding Transparent Truth

One of the motivating issues in the development of paraconsistent logics is the addition of a truth predicate to classical logic. This section looks at a typical paraconsistent response that expands the set of truth-values. §2.1 shows what happens when we introduce a transparent truth-predicate into the classical system; §2.2 outlines the model theory for a multiple-conclusion paraconsistent semantics based on the logic $LP^{+}$. §2.3 goes on to show that the system fails to be absolute, which suggests that the bilateralist account can not provide an inferentialist home for paraconsistent logics.

## 2.1 Transparent truth

Following Beall (2009), a transparent truth predicate is a notion of truth that is "see through", such that $T(\ulcorner\alpha\urcorner)$ and $\alpha$ are intersubstitutable in all transparent contexts, and for all $\alpha \in S$.[3] Given Identity, this gives us familiar rules for $T$: $T(\ulcorner\alpha\urcorner) \vdash \alpha$; and $\alpha \vdash T(\ulcorner\alpha\urcorner)$. In terms of positions, we read this as saying for any incoherent position in which $T(\ulcorner\alpha\urcorner)$ appears, $\alpha$ is also incoherent. This gives us the following sequent rules:

$$\frac{\Gamma, \alpha \vdash \Delta}{\Gamma, T(\ulcorner\alpha\urcorner) \vdash \Delta}\ (T\text{-L}) \qquad\qquad \frac{\Gamma \vdash \alpha, \Delta}{\Gamma \vdash T(\ulcorner\alpha\urcorner), \Delta}\ (T\text{-R})$$

Taken alone, and given that we are working in SET-SET, both $L_{\neg}$, $L_{T}$ are absolute. Taken together, they result in typical paradoxes. Supposing self-reference to be available in the language, we can construct a formula, $\Theta$, of the form $\Theta : \neg T(\ulcorner\Theta\urcorner)$. In the classical framework that we are working in thus far, this quickly gets us into trouble:

[2] See appendix 1.

[3] Corner brackets indicate a name-forming device.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{T(\ulcorner\Theta\urcorner)\vdash T(\ulcorner\Theta\urcorner)}{\neg T(\ulcorner\Theta\urcorner)\vdash \neg T(\ulcorner\Theta\urcorner)}}{\Theta\vdash\Theta}}{T(\ulcorner\Theta\urcorner)\vdash\Theta}}{\vdash\neg T(\ulcorner\Theta\urcorner),\Theta}}{\vdash\Theta}\qquad\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{T(\ulcorner\Theta\urcorner)\vdash T(\ulcorner\Theta\urcorner)}{\neg T(\ulcorner\Theta\urcorner)\vdash \neg T(\ulcorner\Theta\urcorner)}}{\Theta\vdash\Theta}}{\Theta\vdash T(\ulcorner\Theta\urcorner)}}{\Theta,\neg T(\ulcorner\Theta\urcorner)\vdash}}{\Theta\vdash}}{\vdash}$$

The final step uses a single application of Cut, yielding the empty sequent, which is extensible to all $S$. So successive application of the rules allows us to infer any conclusion from any premise. If we attempt to let $L_{\neg T}$ determine a valuation space, it will be incoherent since the only $v \in U$ that are consistent with it are $v_t =_{df} \{\alpha \in S : v(\alpha) = 1\}$, and $v_f =_{df} \{\alpha \in S : v(\alpha) = 0\}$. In other words, the addition of $T$ to $L_{CPL}$ entails that $L_{CPL}$ fails to determine a coherent $V$, and so fails to determine the meanings of the connectives.

### 2.2 *LP* model theory

A typical paraconsistent response to the above employs a logic such as Graham Priest's Logic of Paradox ($LP$), which allows for some sentences, such as Liar sentences to be "gluts", that is, both true and false. This has been extensively argued for in the literature (Priest, 2006). This section outlines the basic model theory for $LP$. However, in keeping with the discussion above, we will remain within a multiple-conclusion structure, and, following Beall (2013), denote this $LP^+$.

We extend Definition 5 to allow three truth-values so that $\mathcal{V} = \{1, b, 0\}$, and $\mathcal{D} = \{1, b\}$. Again, we expect $V$ to be induced by the truth-conditional interpretation of the standard connectives. Then $V$ contains those valuations $v : S \rightarrow \{1, b, 0\}$ that obey the truth-conditional clauses for the connectives.

**Definition 17.** The connectives $\{\wedge, \vee, \neg\}$ satisfy the truth-conditional clauses:
*(i).* Conjunction: $v(\alpha \wedge \beta) = min\,\{v(\alpha), v(\beta)\}$.
*(ii).* Disjunction: $v(\alpha \vee \beta) = max\,\{v(\alpha), v(\beta)\}$.
*(iii).* Negation: $v(\neg\alpha) = 1 - v(\alpha)$

Let $V_{LP}$ be the valuation-space comprising the set of valuations induced for each $\alpha \in S$.

**Definition 18.** A sequent, $\Gamma \vdash \Delta$, is refuted by a valuation $v$ iff, when $v(\alpha) \in D$ for each $\alpha \in \Gamma$, $v(\beta) \notin D$ for all $\beta \in \Delta$); and otherwise satisfied by $v$.

**Definition 19.** A sequent is $V$-valid iff it is satisfied by each $v \in V$. Then;

- $\mathbb{L}(V) =_{df} \{\langle\Gamma, \alpha\rangle : \langle\Gamma, \alpha\rangle$ is $V$-valid$\}$.

**Definition 20.** For a set of formulae $S$ in a language $\mathscr{L}$, an $LP^+$ sequent is an ordered pair, $\Gamma$, $\Delta$, (where $\Gamma \cup \Delta \subseteq S$, and where $\Gamma$, $\Delta$ are sets of formulae of $S$). The logic $LP^+$ is an ordered pair $\langle S, LP^+\rangle$, where $LP^+$ is the set of binary relations $\vdash_{LP^+}$ between finite subsets of $S$ and finite subsets of $S$. We call the set of provable sequent in $LP^+$, $LP^+$-valid, such that $\Gamma \vdash \Delta =_{df} \{\langle\Gamma, \Delta\rangle$ is $LP^+$-valid$\}$.

We note some features of the logic $LP^+$. The logical truths of $LP^+$ are precisely those of classical propositional logic.

**Proposition 21.** $\vdash_{LP^+} \alpha$ *iff* $\alpha$ *is a classical tautology.*

*Proof.* LRD. Since all classical models are also $LP^+$ models, this is clear. RLD. Take a valuation of $LP^+$, $v_{LP^+}$ and a two-valued valuation $v_{CPL}$ which assigns 1 to $\alpha$ if $v(\alpha) \in D$. We can prove by induction that, if $v_{CPL}(\alpha) = 1$ then $v_{LP^+}(\alpha) \in \{1, b\}$, and if $v_{CPL}(\alpha) = 0$ then $v_{LP^+}(\alpha) \in \{b, 0\}$. Hence, if $v_{CPL}(\alpha) = 1$, for each two-valued valuation $v_{CPL}$ then $v_{CPL}(\alpha)$ is designated for each three-valued $v_{LP^+}$. □

**Proposition 22.** *$LP^+$ is paraconsistent. Both (EFQ) and disjunctive syllogism (DS) are invalid in $LP^+$; i.e. $\alpha \wedge \neg\alpha \nvdash_{LP^+} \beta$; $\neg\alpha, \alpha \vee \beta \nvdash_{LP^+} \beta$.*

*Proof.* A counterexample to both is easily given when $v(\alpha) = b$, and $v(\beta) = 0$. □

### 2.3 Absoluteness for $LP^+$

Given that we are working within a modest inferentialist framework, what we are interested in is whether or not our position structures yield a proof-theory that adequately determines $LP^+$ models. In other words, what we require is that, starting from position structures, we can build a logic that completely determines $V_{LP^+}$ as we have for bilateralist classical logic. Without going into any details regarding such a proof-theory, it is simple to show that this is not possible.

**Theorem 23.** *For any logic L of the form SET-SET, and any $V \subseteq U$ (for which $\mathcal{V} = \{1, b, 0\}$), $V \neq \mathbb{V}(\mathbb{L}(V))$.*

*Proof.* See appendix 2. □

The obvious issue is that we can no longer rely upon the partitioning of formulas into those taking $v = 1$, $v = 0$. For $LP^+$, it is possible to construct a kind of absoluteness proof, but only by defining a partition over formulas into those that take designated values and those that do not. That is, where, $D = \{\alpha \in S : v(\alpha) \in \mathcal{D}\}$ and $D^- = \{\alpha \in S : v(\alpha) \in \mathcal{D}^-\}$, (where $\alpha \in \mathcal{D}^- =_{equiv} \alpha \notin \mathcal{D}$). But, whilst partitioning into $D$, $D^-$ tells us something about consequence relations for many-valued logics (which preserve $\mathcal{D}$), it involves a loss of grasp on the semantical clauses over $V_{LP^+}$ where we want to distinguish between designated values $\{1, b\}$.

What this means is that any logic that determines $V_{LP^+}$ will also be consistent with the valuation-space $V_{LP^+} \cup v_b$. So that logic, however we construct it, will underdetermine the semantics of the connectives that it defines since it fails to determine a unique valuation-space. It is not too much of a stretch to say that, working from a modest inferentialist account, we will have failed to adequately determine the meanings of the connectives for any logic dealing with truth-values beyond $\{1, 0\}$. In the next section, I suggest that the problem comes down to the way in which formulas are located in the SET-SET framework. Then, by expanding bilateralist positions to trilateralist positions, we may be able to expand the locations that formulas can take in sequent arguments.

## 3 Expanding positions

This section provides motivation for the expansion of positions to trilateralist structures of the form $\langle \Gamma : \Theta : \Delta \rangle$, where $\Gamma$ indicates the set of asserted statements, $\Delta$ the denied statements, and $\Theta$, a set of "weakly asserted" statements. A statement $\alpha$ is weakly asserted when asserting $\neg\alpha$ does not rule out the content $\alpha$ (i.e. without also denying $\alpha$). First, I clarify why $LP^+$ fails absoluteness by looking at the roles of Cut and Identity in determining the locations of formulas over sequent arguments. This, together with a typical paraconsistent view of assertions and denials motivates the addition of the third, intermediate, location within positions.

### 3.1 Partitions and cut

We clarify why absoluteness fails for many-valued semantic structures by looking at the roles played by Cut and Identity. First, recall that an ordinary SET-SET sequent $\alpha_1, \ldots, \alpha_n \vdash \beta_1, \ldots, \beta_m$ is equivalent to $\{\alpha_1 \wedge \alpha_2 \ldots \wedge \alpha_n\} \rightarrow \{\beta_1 \vee \beta_2 \ldots \vee \beta_m\}$. This, by ordinary propositional logic, is equivalent to $\neg\alpha_1 \vee \neg\alpha_2 \ldots \vee \neg\alpha_n \vee \beta_1 \vee \beta_2 \ldots \vee \beta_m$.

**Definition 24.** Definition 18 is equivalent to saying that a valuation $v$ satisfies a sequent $\Gamma \vdash \Delta$ iff either one of the formulae in $\Gamma$ is not designated or one of the formulas in $\Delta$ is designated. Spelling this out disjunctively over $\mathcal{V}_{LP^+}$, $\Gamma \vdash \Delta$ is $V_{LP^+}$-valid iff, for each $v \in V_{LP^+}$, either $v(\alpha) = 0$, or $v(\alpha) = n$ for some $\alpha \in \Gamma$, or $v(\beta) = 1$, or $v(\beta) = b$ for some $\beta \in \Delta$.

On this interpretation we can think of a sequents as having two locations, with the *l.h.s* of the sequent being the undesignated location, and the *r.h.s* the designated location. So, $\Gamma \vdash \Delta$ may be rewritten $\Gamma_{D^-} \vdash \Delta_D$.

**Example 25.** For a sequent $\Gamma, \alpha \vdash \Delta$ to be valid, we know that each $v \in V$ has $v(\alpha) = 0$ or $v(\alpha) = n$ for some $\alpha \in \Gamma \cup \alpha$, or $v(\beta) = 1$ or $v(\beta) = b$ for some $\beta \in \Delta$. Similarly, for $\Gamma \vdash \Delta, \alpha$ to be valid, each $v \in V$ has $v(\alpha) = 0$ or $v(\alpha) = n$ for some $\alpha \in \Gamma$, or $v(\beta) = 1$ or $v(\beta) = b$ for some $\beta \in \Delta \cup \alpha$.

As this formulation makes clear, Identity plays the role of ensuring that each formula $\alpha \in S$ appears on either the *l.h.s* or the *r.h.s* of a sequent, and Cut ensures that no formula appears on both. Think of this in terms of valuations. Then Identity tells us that all formulas take a designated value or a non-designated value, and Cut that no formula takes both. So, if, as in the case of liar-like sentences, we end up in a position where a formula $\Theta$ is forced to be located on both the *l.h.s* and the *r.h.s* of a sequent, Cut tells us that something has gone wrong. No formula can be forced to be both designated and undesignated.

Inevitably, then, over ordinary sequent structures, formulae will be partitioned into those that are designated, and those that are not, with no way of discriminating between, for example, $\{1, b\}$. Whilst the system behaves nicely for two-valued cases, it is broken by the addition of admissible values $\in \mathcal{V}$. It is for this reason that absoluteness fails for ordinary sequent inferential structures determining many-valued semantic structures. What we require then, is a way of expanding the location of formulas such that the structural rules do not ride roughshod over the finer-grained distinctions between truth-values. Then, we can expect Identity to ensure that every formula $\alpha \in S$ takes some truth-value, and Cut to ensure that no formula takes more than one.

### 3.2 Weak assertion

Once we have allowed that assertion is not the only game in town, the expansion of positions to accommodate a third location is fairly well motivated. Think of liar-like sentences in terms of rational commitment. The presence of liar-like sentences threatens to break down the bilateralist account since $\Theta$ is not something that we can either assert or deny. In the above presentation, $\Theta$ attempts to force an overlap between assertion and denial, which, because of the role played by Cut, is impossible. Advocates of paraconsistent approaches usually take assertion and denial to be exclusive states, so that denying $\alpha$, in a sense, rules out the content of $\alpha$. Nonetheless, an agent can be inferentially committed to asserting certain contradictions such that $\alpha, \neg\alpha \nvdash$. This is because asserting $\neg\alpha$ is not taken to be the the dual of denying $\alpha$.[4] Asserting $\neg\Theta$ will not, then have us also deny $\Theta$. Rather, we should assert both $\Theta$ and $\neg\Theta$. To

[4] For discussion, see Parsons (1984).

follow this line of thought, $\Theta$ would then not force an overlap between assertion and denial, but between assertion of $\Theta$ and $\neg\Theta$.

On the paraconsistent response, denying $\alpha$ must, therefore, be stronger than asserting $\neg\alpha$. In this respect, we might think of this in terms of an additional speech act that is weaker than assertion in that it does not cancel the content of the opposite proposition. As analogy, whilst ordinary assertion and denial behave much like "exclusion negation", weak assertion will behave much like "choice negation" in a language.[5] Think of this is in terms the grounds, or evidence, that agents have for asserting or denying a proposition $\alpha$.[6] When an agent asserts $\alpha$, there should be grounds in the language (given a specific context) supporting $\alpha$ such that $\alpha$ can not also be denied. But in paradoxical cases such as liar-like sentences $\Theta$, there are grounds in the language that provide support for $\neg\Theta$, but not in such a way as to also support the denial of $\Theta$. An agent who is rationally committed to $\neg\Theta$ is not thereby committed to the denial of $\Theta$. In order for that to be the case, we would also require some sort of grounds for ruling out the content $\Theta$, but this is precisely what the liar-reasoning fails to provide. So, we have reason to consider a third speech-act corresponding to cases of this kind.

**Definition 26.** (Weak assertion) A statement is weakly asserted when asserting $\neg\alpha$ does not rule out the content $\alpha$ (i.e. without also denying $\alpha$).

We should, of course, be wary of introducing *ad hoc* distinctions into a theory. But, in this case, there are significant reasons for making these distinctions given that we are working in a modest inferentialist framework. Even aside from the technical problems discussed above, distinguishing between ordinary and weak assertion provides a way of understanding what an agent becomes inferentially committed to in accepting some contradictions ($\alpha, \neg\alpha \nvdash$). Some formulae, such as $\Theta$ can be weakly accepted.[7]

If this is plausible, then we have a trilateralist position structure $[\Gamma : \Theta : \Delta]$, with $\Gamma$ indicating the set of asserted statements, $\Delta$ the denied statements, and $\Theta$, the set of "weakly asserted" statements. The relation between incoherent trilateralist positions and provable sequents will be pretty much as above, except that we will have to allow for a third location in our sequent structure, and the structural rules will need to account for this distinction.

# 4 Sequent calculi for trilateralist positions

This section develops a sequent calculus for $LP^+$. We being from trilateralist positions, and draw upon $n$-sided sequent calculi as developed in Baaz et al. (1993a,b, 1998), Hjortland (2013), and Zach (1993) to construct a proof-theory which is absolute on $V_{LP^+}$. The resulting proof-theory has significant benefits in addition to the fact that it is suitable for the modest inferentialist. Primarily, it offers an account of the necessary and sufficient conditions under which certain classical valid, though paraconsistently invalid, arguments can be "recaptured".

## 4.1 From trilateral positions to $n$-sequents

We rewrite the ordinary SET-SET sequent $\Gamma_0 \vdash \Gamma_1$ as a two-sided sequent $\Gamma_0|\Gamma_1$. Whilst the indices indicate values, informally this should be read in terms of attitudes, indicating that

[5] See Tappenden (1999) for cases in which the use of "not" in a natural language context indicates the rejection of an assertion without also indicating the assertion of the negation of the relevant sentence; e. g. "Some men are not chauvinists. All of them are", "John isn't wily or crazy. He's wily and crazy".

[6] See, for example, Pagin (2012).

[7] Whether or not further distinctions amongst commitments may be warranted is left for further investigation. The obvious further distinction would be the dual of weak assertion, weak rejection, which may be thought to correspond roughly to a truth-value gap (see Remark 36).

either something in $\Gamma_0$ is rejected, or something in $\Gamma_1$, accepted. Then, we simply carry over from Definition 26:

**Definition 27.** $\Gamma_0|\Gamma_1$ is $V$-valid iff either $v(\alpha) = 0$ for some $\alpha \in \Gamma_0$, or $v(\beta) = 1$ for some $\beta \in \Gamma_1$.

The idea is to expand this strategy for three formula locations using $n$-sided sequents.

**Definition 28.** An $n$-sided sequent $\Gamma$ is an ordered $n$-tuple of finite sequences $\Gamma_1|\dots|\Gamma_n$where $\Gamma_n$ is the $n$-th component of $\Gamma$.

Where $\Gamma$ is a sequent, $\Gamma_i$ denotes the $i$-th component of the sequent, with the sequent interpreted as a disjunction of statements saying that a particular formula takes a particular location in the structure of the sequent. Then, it is straightforward to define a three-sided sequent corresponding to the logic $LP^+$.

**Example 29.** A three-sided sequent for $L_{LP^+}$ is written as:

$$\Gamma_1|\Gamma_b|\Gamma_0$$

again, read disjunctively as saying that either $v(\alpha) = 1$ for some $\alpha \in \Gamma_1$, or $v(\theta) = b$ for some $\theta \in \Gamma_b$, or $v(\beta) = 0$ for some $\beta \in \Gamma_0$.

The relation between incoherent trilateral positions and sequent provability can now be characterized.

**Definition 30.** (Three-sided sequent provability) If $[\Gamma : \Theta : \Delta]$ is incoherent, then $\Gamma_1|\Gamma_b|\Gamma_0$ is valid.

We need to spell out incoherence disjunctively. A position $[\Gamma : \Theta : \Delta]$ is incoherent if an agent either denies or weakly asserts some $\alpha \in \Gamma$, or asserts or denies some $\theta \in \Theta$, or asserts or weakly asserts some $\beta \in \Delta$. We use this to define the logic $LP^+$.

**Definition 31.** A three-sided logic is an ordered pair $\langle S, L\rangle$, where $L$ is a set of relations between finite sets of three-sided sequents of $S$, and each sequent argument in $L$ is called $L$-valid. For the set of truth-values $\mathcal{V} = \{1, b, 0\}$, and for each location $\Gamma_i$ (which is a possibly empty set of formulae $\in S$), a valuation $v$ satisfies a three-sided sequent iff for some $\Gamma_i$, when $i \in \{1, b, 0\}$, and some formula $\alpha \in \Gamma_i, v(\alpha) = i$.

**Definition 32.** An three-sided sequent is $V$-valid iff it is satisfied by each $v \in V$. A valuation $v \in U$ is $L$-consistent iff $v$ satisfies each provable sequent in $L$.

- $\mathbb{L}(V) =_{df} \{\langle\Gamma_1|\Gamma_b|\Gamma_0\rangle \in L : \langle\Gamma_1|\Gamma_b|\Gamma_0\rangle$ is $V$-valid$\}$
- $\mathbb{V}(L) =_{df} \{v \in U : v$ is $L$-consistent$\}$

By considering locations in trilateral positions pair-wise, we are able to ensure that the structural rules play the required roles, now formulated as follows (Baaz et al., 1993a).

$$\alpha|\alpha|\alpha(\text{Identity})$$

For each sequent location $i$:

$$\frac{\Gamma}{\Gamma, [i : \alpha]}\,(\text{Weakening})$$

For each couple of truth-values where $v_i \neq v_j$:

$$\frac{\Gamma, [i : \alpha] \qquad \Delta, [j : \alpha]}{\Gamma, \Delta}\ (\text{Cut}(i, j))$$

Since the structure has more than two locations for formulae, Identity now makes sure that each formula takes a valuation, and Cut operates on pairs of truth-values. So, Cut still partitions formulae by ensuring that each formula takes a single truth-value only. For any $v_i \neq v_j$ for a formula $A$ such that $v(\alpha) = v_i = v_j$, $\alpha$ is removed:

$$\frac{\Gamma_1|\dots|\Gamma_i, \alpha|\dots|\Gamma_n \qquad \Delta_1|\dots|\Delta_j, \alpha|\dots|\Delta_n}{\Gamma_1, \Delta_1|\dots|\Gamma_n\Delta_n}$$

The difference is that for ordinary sequents, we had only two values to worry about, $\{1, 0\}$ (or two locations).

## 4.2 Proof-theory for $LP^+$

We construct an 3-sided proof-system in a uniform way (Baaz et al., 1993a). The structural rules remain as above and operational rules for each connective are given for each location in the three-sided sequent as follows:

$$\frac{\Gamma_0|\Gamma_b|\Gamma_1, \alpha \qquad \Gamma_0|\Gamma_b|\Gamma_1, \beta}{\Gamma_0, |\Gamma_b|\Gamma_1, \alpha \wedge \beta}\ (\wedge 1) \qquad \frac{\Gamma_0, \alpha, \beta|\Gamma_b|\Gamma_1}{\Gamma_0, \alpha \wedge \beta|\Gamma_b|\Gamma_1}\ (\wedge 0)$$

$$\frac{\Gamma_0, \alpha|\Gamma_b, \alpha|\Gamma_1 \qquad \Gamma_0, \beta|\Gamma_b, \beta|\Gamma_1 \qquad \Gamma_0|\Gamma_b, \alpha, \beta|\Gamma_1}{\Gamma_0, |\Gamma_b, \alpha \wedge \beta|\Gamma_1}\ (\wedge b)$$

$$\frac{\Gamma_0, \alpha|\Gamma_b|\Gamma_1 \qquad \Gamma_0, \beta|\Gamma_b|\Gamma_1}{\Gamma_0, \alpha \vee \beta|\Gamma_b|\Gamma_1}\ (\vee 0) \qquad \frac{\Gamma_0|\Gamma_b|\Gamma_1 \alpha, \beta}{\Gamma_0|\Gamma_b|\Gamma_1, \alpha \vee \beta}\ (\vee 1)$$

$$\frac{\Gamma_0, \alpha|\Gamma_b, \alpha|\Gamma_1 \qquad \Gamma_0, \beta|\Gamma_b, \beta|\Gamma_1 \qquad \Gamma_0|\Gamma_b, \alpha, \beta|\Gamma_1}{\Gamma_0|\Gamma_b, \alpha \vee \beta|\Gamma_1}\ (\vee b)$$

$$\frac{\Gamma_0|\Gamma_b|\Gamma_1, \alpha}{\Gamma_0, \neg\alpha|\Gamma_b|\Gamma_1}\ (\neg 0) \qquad \frac{\Gamma_0, \alpha|\Gamma_b|\Gamma_1}{\Gamma_0|\Gamma_b|\Gamma_1, \neg\alpha}\ (\neg 1)$$

$$\frac{\Gamma_0|\Gamma_b, \alpha|\Gamma_1}{\Gamma_0|\Gamma_b, \neg\alpha|\Gamma_1}\ (\neg b)$$

The logic $L_{LP^+}$ comprises the set of valid sequent arguments determined by the proof-system. For some $V \subseteq U$ (where $U$ has $\mathcal{V} = \{1, b, 0\}$), $V_{L_{LP^+}} = \mathbb{V}(L_{LP^+}) = \{v : v \text{ is } L_{LP^+}\text{-consistent}\}$.

Soundness and completeness for $\mathbb{V}(L_{LP^+})$ are relatively simple. First, soundness can be proved, in the usual way, by induction over proofs, since the operational rules preserve validity by definition, and the structural rules are valid. For completeness, if a sequent is $V_{LP^+}$-valid, then it is provable in the three-sided construction without cuts.[8] The proof can be carried over from Baaz et al. (1993b), which uses the method of reduction trees.[9]

Most importantly for our purposes, $L_{LP^+}$ is absolute *w.r.t* $V_{LP^+}$ as outlined semantically in §2.2.

**Theorem 33.** *In general, for a valuation space $V \subseteq U$, $V = \mathbb{V}(\mathbb{L}(V))$ (Hjortland, 2014).*

*Proof.* See appendix 3. □

[8] I note a complication regarding cut-elimination below.

[9] Priest (2008) uses the tableau system to give a completeness proof for $FDE$, which, though not constructed using $n$-sequents, can be carried over with a little tweaking.

As an immediate corollary, the 3-sided proof-theory for $LP^+$ yields an absoluteness result. Then, we have a modest inferentialist account that, beginning with trilateralist positions, provides a proof-theory that uniquely determines the valuation-space $V_{LP^+}$.

## 4.3 Notable features of the proof-theory

### 4.3.1 Finer-grained distinctions

An immediate advantage of the 3-sided proof-theory (in addition to its inferentialist scruples), is that it offers a way of maintaining fine-grained distinctions across sequent arguments. By way of illustration, we can see that there is, despite absoluteness, no way of retaining these distinctions in the ordinary sequent set-up.

**Example 34.** For $\mathbb{L}_{LP^+}(V_{LP^+})$, typically, we will say that $\Gamma \vdash \Delta$ is valid iff when $v(\alpha) \in D$ for each $\alpha \in \Gamma$, $v(\beta) \in D$ for some $\beta \in \Delta$, or equivalently, either $v(\alpha) \notin D$ for some $\alpha \in \Gamma$ or $v(\beta) \in D$ for some $\beta \in \Delta$. We can spell out the latter by saying that either $v(\alpha) \neq 1$ and $v(\alpha) \neq b$ for some $\alpha \in \Gamma$ or $v(\beta) = 1$ or $v(\beta) = b$ for some $\beta \in \Delta$. But, as we saw in §3.2, as long as $\vdash_{LP^+}$ obeys Cut (ensuring transitivity for the two-sided system), the differentiation amongst $D$ is lost. This is clearer when we consider that by the definitions of $V_{LP^+}$ and $V_{LP^+}$-validity, $\Gamma \vdash_{LP^+} \Delta$ when $\Gamma|\Gamma|\Delta$ is $V_{LP^+}$-valid. In other words, we can switch back from $\Gamma_0|\Gamma_b|\Gamma_0$ to the two-sided $\Gamma_0 \to \Gamma_b \cup \Gamma_1$ where the latter is $V_{LP^+}$-valid when $v(\alpha) \notin D$ for some $\alpha \in \Gamma_0$ or $v(\beta) \in D$ for some $\beta \in \Gamma_b \cup \Gamma_1$. But, there is no route back from a $V_{LP^+}$-valid sequent in this two-sided incarnation to a specific $V$-valid 3-sided sequent. In essence, this is due to the fact that, whilst Cut holds for each of the three-locations in the three-sided structure, the three-sided version of Cut does not guarantee transitivity in two-sided systems because of the way in which the three-sided locations overlap in the two-sided structure.

This provides additional reason to think that pursuing issues related to paraconsistent logics in three-sided constructions may be preferable to standard proof-theories (including those in SET-SET) because of the fine-grained nature preserved by the logic.

### 4.3.2 Transparent truth

Take $L_{LP}$, and additionally define operational rules for $T$:

$$\frac{\Gamma_0, \alpha|\Gamma_b|\Gamma_1}{\Gamma_0, T(\ulcorner\alpha\urcorner)|\Gamma_b|\Gamma_1}\ (T0)$$

$$\frac{\Gamma_0|\Gamma_b, \alpha|\Gamma_1}{\Gamma_0|\Gamma_b, T(\ulcorner\alpha\urcorner)|\Gamma_1}\ (Tb) \qquad \frac{\Gamma_0|\Gamma_b|\Gamma_1, \alpha}{\Gamma_0|\Gamma_b|\Gamma_1, T(\ulcorner\alpha\urcorner)}\ (T1)$$

Now, cut is no longer eliminable and the subformula property fails since $T$ applies to formulae of any complexity, including the liar sentence. It is not clear whether or not this is problematic. For example, above, it was suggested both that $\mathcal{T}$ may be considered part of the basic set of constraints on rational commitment, and it is required for absoluteness proofs. As such, eliminating cut, whilst of technical interest, loses some philosophical motivation, particularly given that any calculus involving $T$ suffers from a loss of the subformula property in any case.[10]

[10] It may be possible to rectify this by considering sequents in terms of multisets rather than sets, and develop a construction without the structural rule of contraction.

### 4.3.3 Classical recapture

As resultant from this finer-grained distinctions between formulas over sequent arguments, the three-sided proof-theory also provides a nice way of understanding "classical recapture". $LP^+$ is notoriously weak in comparison with classical logic. As above, both EFQ and DS are invalid in $LP^+$. The failure of such inferences is counter-intuitive, and it threatens to undermine significant features of ordinary reasoning. Resultantly, paraconsistent logics have come under significant criticism (e. g. Parsons 1984; discussed in Priest 2006). By way of response, various forms by which classical reasoning can be "recaptured" have been suggested (Priest, 2006, §8), where:

> **Methodological maxim:** Unless we have specific grounds for believing that the crucial conditions in a piece of quasi-valid reasoning are gluts, we may accept that reasoning. (116)

Whilst the maxim is certainly plausible, working out the logical details underlying it has proven difficult. For example, we can not force consistency by somehow adding consistency to the premise set. A natural thought would be to employ the truth-predicate, with a formulation of the law of non-contradiction, such as $\neg T(\alpha \wedge \neg\alpha)$ to "force" the consistency of truth. But, $\neg T(\alpha \wedge \neg\alpha)$ does not logically rule out the possibility that $T(\alpha \wedge \neg\alpha)$. A prominent suggestion (Priest, 2006) has been to add a stronger than material conditional to the stock of logical connectives, where, for all $\alpha$, $(\alpha \wedge \neg\alpha) \rightarrow\perp$. But, arguably, this is too strong as it requires a logical connection between antecedent and consequent derived from relevant logic (Beall, 2012).

In any case, the three-sided proof-theory above has a significant advantage over these suggestions in that it provides us with necessary and sufficient conditions for when certain classically valid inferences are provable, and when they are not. For example, take the negation rules. $\neg$0 and $\neg$1 give the conditions for negation in $L_{LP^+}$ corresponding to assertion and denial. A formula may be denied if its negation is asserted, and asserted if its negation is denied. This corresponds to the classical rules $\neg$-L or $\neg$-R. The difference (and the reason why $\neg$-L and $\neg$-R no longer force the equivalence of asserting $\neg\alpha$ with denying $\alpha$) is that we also have the rule $\neg$b. This gives the conditions for negation corresponding to weak assertion. The negation of a formula may be weakly asserted when the formula is weakly asserted. Taken together, the rules impose constraints on $\mathbb{V}(L_{LP^+})$ corresponding to the semantic definition of negation (Definition 17).

**Example 35.** Take disjunctive syllogism (DS) as example. As we saw in Proposition 22, (DS) is invalid in $LP^+$. Let us analyze this in detail in the three-sided logic. First, consider the disjunction rule $\vee b$ (since $\vee 0$ and $\vee 1$ are familiar). The rule provides three (necessary and sufficient) conditions under which a disjunction $\alpha \vee \beta$ may be weakly asserted: $\alpha$ or $\beta$ is weakly asserted; and $\alpha$ is denied or weakly asserted; and $\beta$ is denied or weakly asserted. As before, we have a counterexample to (DS) where $\alpha$ is weakly asserted, and $\beta$ denied. Since $\alpha$ is weakly asserted, we know that $\neg\alpha$ is weakly asserted. Then, we can derive the provable sequent argument $\Gamma, \neg\alpha, \alpha \vee \beta \vdash \alpha, \neg\alpha, \Delta$:

$$\dfrac{\dfrac{\Gamma_0|\Gamma_b, \alpha \vee \beta|\Gamma_1}{\Gamma_0, \beta|\Gamma_b, \alpha|\Gamma_1}\,(\vee b) \qquad \Gamma_0|\Gamma_b, \neg\alpha|\Gamma_1}{\Gamma_0|\Gamma_b, \alpha, \neg\alpha|\Gamma_1}\,(\neg b)$$

But, notice that we also have the resources in the proof-system providing the necessary and sufficient conditions for when it is permissible to derive (DS). When, for example, $\alpha$ and $\beta$ are asserted, we know (by the rule $\vee 1$) that $\alpha \vee \beta$ may be asserted. We also know (by $\neg 1$) that whenever $\alpha$ is asserted, $\neg\alpha$ may be denied. This allows us to derive (DS) $(\Gamma, \neg\alpha, \alpha \vee \beta \vdash \beta, \Delta)$:

$$\dfrac{\dfrac{\Gamma_0|\Gamma_b|\Gamma_1, \alpha \vee \beta}{\Gamma_0|\Gamma_b|\Gamma_1, \alpha, \beta}\,(\vee 1) \qquad \dfrac{\Gamma_0|\Gamma_b|\Gamma_1, \neg\alpha}{\Gamma_0, \alpha|\Gamma_b|\Gamma_1}\,(\neg 1)}{\Gamma_0|\Gamma_b|\Gamma_1, \beta}\,(\text{Cut})$$

*Remark* 36. It is not difficult to see that the 3-sided construction for $LP$ is analogous to a construction for the Kleene 3-valued logic, $K_3$. $K_3$ similarly has three algebraic truth-values, with the middle value typically denoted $i$ for indeterminate, so $\mathcal{V} = \{1, i, 0\}$. In distinction with $LP$, $K_3$ has $\mathcal{D} = \{1\}$. It is well known that $K_3$ is paracomplete (law of excluded middle may not hold), whereas $LP$ is paraconsistent. So, the two logics have distinct consequence relations, since, whilst they share the same interpretation of standard connectives, they differ with respect to the interpretation of the truth-values. This fact is typically reflected in standard proof-theoretic constructions of the two logics.[11] However, in an $n$-sequent construction, the two coincide apart from the decoration of the middle sequent. Nonetheless, whilst the decoration is arbitrary, it reflects a distinction between the two structures at the level of provability.[12] By the translation in Example 34, we say that $\Gamma \vdash_{LP} \Delta$, whenever $\Gamma|\Delta|\Delta$ is derivable in $L_{LP}$; in distinction, $\Gamma \vdash_{K_3} \Delta$, whenever $\Gamma|\Gamma|\Delta$ is derivable in $L_{K_3}$, where $L_{K_3}$ is equivalent to $L_{LP}$ (just decorate the middle sequent with $i$ in place of $b$). This difference allows us to distinguish between the two structures, so that, for example, law of excluded middle is derivable in $L_{LP}$, but not in $L_{K_3}$. Additionally, we know, by Theorem 33, that the semantics for $L_{K_3}$, $\mathbb{V}(L_{K_3})$, will be absolute, and, since the designated values differ from that of $\mathbb{V}(L_{LP})$, the consequence relation differs accordingly.

We can make sense of this approach to $L_{K_3}$ by considering the dual to weak assertion, "weak denial", where a statement $\alpha$ is weakly denied when denying $\neg\alpha$ does not rule in the content $\alpha$ (i.e. without also asserting $\alpha$). This is typical in discussions of paracomplete logics. Of course, if we allow weak denial alongside weak assertion, then it is possible to construct a four-sided sequent structure by simply expanding sequents to: $\Gamma_1|\Gamma_b|\Gamma_i|\Gamma_0$. Unsurprisingly, the semantic structure determined by the full (four-sided) sequent structure is precisely that of first degree entailment ($FDE$), and, again, by Theorem 34, their relationship is absolute.

# Conclusion

This paper has developed a modest inferentialist approach to paradox. Starting with a bilateralist framework, an account of absoluteness for multiple-conclusion classical logic was provided. This, however, does not carry over to many-valued logics such as $LP^+$, primarily due to structural deficiencies in both bilateral positions and ordinary sequent formulas. By way of response, I suggested expanding bilateralism to trilateralism, which incorporates a third, intermediate, speech-act, "weak assertion". In doing so, we can construct, out of trilateral positions, a proof-theory using three-sided sequents. This proof-theory for $LP^+$ is absolute, and also provides a simple way of understanding classical recapture.

# References

Baaz, M., Fermüller, C. G., Salzer, G. & Zach, R. (1998), 'Labeled calculi and finite-valued logics', *Studia Logica*, **61** (1), 7–33.

[11] See, for example, Priest (2008).

[12] See, for example, Hjortland (2013), where the relation between the two structures is discussed in the context of logical pluralism.

Baaz, M., Fermüller, C. G. & Zach, R. (1993a), 'Elimination of cuts in first-order finite-valued logics', *Elektronische Informationsverarbeitung und Kybernetik*, **29** (6), 333–355.

Baaz, M., Fermüller, C. G. & Zach, R. (1993b), Systematic construction of natural deduction systems for many-valued logics, *in* D. M. Miller, Q. Hong, A. Lloris-Ruiz et al., eds., 'Proceedings of the Twenty-Third International Symposium on Multiple-Valued Logic', IEEE, New York, pp. 208–213.

Beall, J. (2012), 'Why priest's reassurance is not reassuring' *Analysis* **72** (3), 517–525.

Beall, J. (2013), 'Free of detachment: Logic, rationality, and gluts' *Noûs*. http://dx.doi.org/10.1111/nous.12029.

Beall, J. C. (2009), *Spandrels of Truth*, Oxford University Press, Oxford.

Belnap, N. D. & Massey, G. J. (1990), 'Semantic holism', *Studia Logica* **49** (1), 67–82.

Boghossian, P. A. (2003), 'Epistemic analyticity: A defense', *Grazer Philosophische Studien* **66** (1), 15–35.

Carnap, R. (1943), *Formalization of Logic*, Harvard University Press, Cambridge, MA.

Dunn, J. M. & Hardegree, G. (2001), *Algebraic methods in philosophical logic*, Oxford University Press, Oxford.

Garson, J. W. (2010), 'Expressive power and incompleteness of propositional logics', *Journal of Philosophical Logic* **39** (2), 159–171.

Hardegree, G. M. (2005), 'Completeness and super-valuations', *Journal of Philosophical Logic* **34** (1), 81–95.

Hjortland, O. T. (2013), 'Logical pluralism, meaning-variance, and verbal disputes', *Australasian Journal of Philosophy* **91** (2), 355–373.

Hjortland, O. T. (2014), 'Speech acts, categoricity, and the meanings of logical connectives', *Notre Dame Journal of Formal Logic* **55** (4), 445–467.

Humberstone, L. (2011), *The Connectives*, MIT Press, Cambridge, MA.

Pagin, P. (2012), 'Assertion, inference, and consequence', *Synthese* **187** (3), 869–885.

Parsons, T. (1984), 'Assertion, denial, and the liar paradox', *Journal of Philosophical Logic* **13** (2), 137–152.

Peacocke, C. (1986a), *Thoughts: An Essay on Content*, Blackwell, Oxford.

Peacocke, C. (1986b), What determines truth conditions?, *in* P. Pettit & J. McDowell, eds., 'Subject, Thought, and Context', Clarendon Press, Oxford, pp. 181–207.

Peacocke, C. (1987), 'Understanding logical constants: A realist's account' *Proceedings of the British Academy* **73**, 153–200.

Peacocke, C. (1993), Proof and truth, *in* J. Haldane & C. Wright, eds. 'Reality, Representation and Projection', Oxford University Press, New York, pp. 165–190.

Priest, G. (2006), *In Contradiction: A Study of the Transconsistent*, Oxford University Press, Oxford.

Priest, G. (2008), *An Introduction to Non-Classical Logic: From If to Is*, Cambridge University Press, Cambridge.

Restall, G. (2005), Multiple conclusions, *in* P. Hájek, L. Valdés-Villanueva & D. Westerståhl, eds., 'Logic, Methodology and Philosophy of Science: Proceedings of the Twelfth International Congress', Kings College Publications, London, pp. 189–205.

Restall, G. (2009), 'Truth values and proof theory' *Studia Logica* **92** (2), 241–264.

Shoesmith, D. J. & Smiley, T. J. (1978), *Multiple Conclusion Logic*, Cambridge University Press, Cambridge.

Tappenden, J. (1999), Negation, denial and language change in philosophical logic, *in* D M. Gabbay & H. Wansing, eds., 'What is Negation?', Springer, Berlin, pp. 261–298.

Trafford, J. (2014), 'Compositionality and modestinferentialism', *Teorema* **33** (1), 39–56.

Zach, R. (1993), *Proof theory of finite-valued logics*, Master's thesis, Technische Universität Wien.

# Appendices

## Appendix 1

The below follows the account given in Trafford (2014).

Thinking of $L$ and $\mathbb{V}(L)$ in the abstract (as not yet constrained by any proof-system), we have, in effect, two partially ordered sets defined over $\mathscr{L}$ (Hardegree, 2005):

(P1) The set of all valuation-spaces $V \subseteq U$ on $\mathscr{L}$, ordered by set-inclusion;

(P2) The set of all logics $L \subseteq L'$ on $\mathscr{L}$, ordered by set-inclusion.

The relation between the two induces an antitone Galois connection between valuation spaces and logics (Dunn and Hardegree, 2001; Hardegree, 2005; Hjortland, 2014; Humberstone, 2011), where a generalized Galois connection is an adjunction of maps between partially ordered sets in terms of order preservation functions.

**Definition.** A Galois connection between posets $P$, $Q$ is a map: $f_1 : P \to Q$ and $f_2 : Q \to P$ where the following conditions hold for all subsets $P_n$, $Q_n$ of $P$, $Q$:

$$P_0 \subseteq f_2(f_1(P_0)) \tag{4.1}$$

$$Q_0 \subseteq f_1(f_2(Q_0)) \tag{4.2}$$

$$P_0 \subseteq P_1 \Rightarrow f_1(P_1) \subseteq f_1(P_0) \tag{4.3}$$

$$T_0 \subseteq T_1 \Rightarrow f_2(T_1) \subseteq f_2(T_0) \tag{4.4}$$

It follows that $f_1 \subseteq f_1 f_2 f_1 \subseteq f_1$, so $f_1 = f_1 f_2 f_1$ and also $f_2 = f_2 f_1 f_2$.

For our purposes, here $P$ is the set of all valuations over $\mathscr{L}$, and $Q$ the set of all sequents in $L$, with satisfaction being the relation defining the functions between them. For any valuation space $V$, $f_1(V)$ consists of the set of sequents satisfied by each $v \in V$, i.e. $f_1(V) = \mathbb{L}(V)$. For any logic $L$, $f_2(L)$ will consist of the set of valuations that satisfy every sequent in $L$, i.e. $f_2(L) = \mathbb{V}(L)$.

With this, we can define a closure operator $cl$ as a function on the posets $\langle V, L\rangle$, given that $cl$ obeys the following clauses for all $x$, $y$ on $\langle V, L\rangle$:

(c1) $x \leq cl(x)$

(c2) $cl(cl(x)) \leq cl(x)$

(c3) $x \leq y \to cl(x) \leq cl(y)$

This ensures that, where $cl$ is a closure operator on a poset $\langle P, \leq\rangle$, and $x$ is an element of $P$, then $x$ is closed iff $cl(x) = x$. In our context, this gives us an abstract completeness theorem over $\langle \mathbb{V}, \mathbb{L}\rangle$.

**Fact.** *For each $V \subseteq U$ and $L \subseteq L'$ (for some $S$):*

$$L \subseteq \mathbb{L}(\mathbb{V}(L)) \tag{4.5}$$

$$V \subseteq \mathbb{V}(\mathbb{L}(V)) \tag{4.6}$$

$$L \subseteq L' \Rightarrow \mathbb{V}(L') \subseteq \mathbb{V}(L) \tag{4.7}$$

$$V \subseteq U \Rightarrow \mathbb{L}(U) \subseteq \mathbb{L}(V) \tag{4.8}$$

*Proof.* Given at length in Hardegree (2005). □

With this, we define *absoluteness*.

**Fact.** *(Hardegree, 2005) For any $L$, $V$;*

- $L$ is absolute iff $L = \mathbb{L}(\mathbb{V}(L))$
- $V$ is absolute iff $V = \mathbb{V}(\mathbb{L}(V))$.

*Proof.* By the fact that $L, V$ form a Galois map, and the definition of Galois closure (c1–3). □

Resultantly, we can give general soundness and completeness theorems for the construction of any normal, finite logic.

**Fact.** *Let $\Gamma$ be any set of formulas in $S$. Define $v_\Gamma$, as: $v_\Gamma(\alpha) = 1$ if $\Gamma \vdash \alpha$, and $v_\Gamma(\alpha) = 0$ otherwise. Then $v_\Gamma$ is $L$-consistent and $v_\Gamma \in \mathbb{V}(L)$.*

*Proof.* (Hardegree, 2005) If not, there must be a sequent, $\Delta \vdash \beta$ in $L$ that is refuted by $v_\Gamma$, so that $v_\Gamma(\Delta) = 1$ and $v_\Gamma(\beta) = 0$. Given the way in which $v_\Gamma$ is defined, this means that $\Gamma \vdash \Delta$. But, $\Delta \vdash \beta$ is $L$-valid, and given that the $\vdash$ associated with $L$ is closed under transitivity, it follows that $\Gamma \vdash \beta$, so by the definition of $v_\Gamma$, $v_\Gamma(\beta) = 1$, so $v_\Gamma$ does not refute $\Gamma \vdash \beta$. □

**Fact.** *In general, for any finite normal logic $L$, $L = \mathbb{L}(\mathbb{V}(L))$.*

*Proof.* (Hardegree, 2005) Suppose that some $\langle \Gamma \vdash \beta \rangle \notin L$, to show that $\langle \Gamma \vdash \beta \rangle \notin \mathbb{L}(\mathbb{V}(L))$ (in other words, it is refuted by $\mathbb{V}(L)$. Take the valuation $v_\Gamma$, which by Lemma 17 is in $\mathbb{V}(L)$. By definition, $v_\Gamma$ satisfies all derivable sequents of $L$. Since $L$ is reflexive, each element of $\Gamma$ is derivable in $L$, so $v_\Gamma$ satisfies $\Gamma$. But, since $\langle \Gamma \vdash \beta \rangle$ is not $L$-valid, $\beta \notin \Gamma$, so $v_\Gamma$ refutes $\beta$. Then $v_\Gamma$ refutes $\langle \Gamma \vdash \beta \rangle$, and so too does $\mathbb{V}(L)$, thus $\langle \Gamma \vdash \beta \rangle \notin \mathbb{L}(\mathbb{V}(L))$. □

**Theorem.** *For any $V \subseteq U$, $V = \mathbb{V}(\mathbb{L}(V))$.*

*Proof.* (Dunn and Hardegree, 2001, p. 200) We prove contra-positively by defining a valuation $v_0 \notin V$ (in order to show that $v_0 \notin \mathbb{V}(\mathbb{L}(V))$). Then define $T = \{\alpha \in S : v_0(\alpha) = 1\}$ and $F = \{\alpha \in S : v_0(\alpha) = 0\}$. For any $v \in V$, $v \neq v_0$, so either $v(\alpha) = 0$ for some $\alpha \in T$ or $v(\alpha) = 1$ for some $\alpha \in F$. Then $v$ satisfies $T \vdash F$, and it follows that $T \vdash F$ is valid on $V$. But, by definition, $v_0$ refutes $T \vdash F$, so $v_0 \notin \mathbb{V}(\mathbb{L}(V))$. □

**Example.** Absoluteness does not hold for $V_{CPL}$ because the classical proof-system defining the connectives $\neg, \vee$ is compatible with valuations $\notin V_{CPL}$. For example, say we define negation in this framework as:

$$\frac{\Gamma, A \vdash B \wedge \neg B}{\Gamma \vdash \neg A}\ (Reductio) \qquad\qquad \frac{\Gamma \vdash A \qquad \Gamma \vdash \neg A}{\Gamma \vdash B}\ (EFQ)$$

>From this, induce $L_\neg$, and determine a corresponding valuation space $V_\neg$. Then $V_\neg \neq V_{CPL}$ if the latter is supposed accord with the truth-functional definition $f^\neg$:

$$f^{\neg}(x) = \begin{cases} 1 & if\ x = 0 \\ 0 & if\ x = 1 \end{cases}$$

More precisely, $\mathbb{V} \neq \mathbb{V}(\mathbb{L}(V_{CPL\neg}))$ because, whilst $L_{\neg}$ is sound and complete *w.r.t* $V_{CPL}$, it is also sound and complete *w.r.t* alternative semantic structures. For example, Hardegree (2005) defines a super-valuation associated with a valuation-space $V$ to be the valuation $v_V$ where, for every $WFF$, $\alpha$, $v_V(\alpha) = 1$ if $v(\alpha) = 1$ for every $v \in V$ and $v_V(\alpha) = 0$ otherwise. Clearly, $L_{\neg}$ is sound and complete *w.r.t* both $V_{CPL}$, and $V_{CPL} \cup v_V$.

## Appendix 2

**Theorem.** *For any logic L of the form SET-SET, and any $V \subseteq U$ (for which $\mathcal{V} = \{1, b, 0\}$), $V \neq \mathbb{V}(\mathbb{L}(V))$.*

*Proof.* We show that this is the case by showing that the absoluteness proof when $\mathcal{V} = \{1, 0\}$ (Theorem 16) is inadequate when we add intermediate truth-values. First, we define a valuation $v_b \notin V_{LP^+}$ (in order to show that $v_b \in \mathbb{V}(\mathbb{L}(V_{LP^+}))$). Let $v_b$ be defined such that, for every $\alpha \in S$, $v_b(\alpha) = b$ (clearly, $v_b \notin V_{LP^+}$ since $v_b$ makes everything a glut at once). Since $v_b \notin V_{LP^+}$, for each $v \in V_{LP^+}$, there is some formula $\alpha$ for which $v(\alpha) \neq v_b(\alpha)$. For such a formula, let $\Gamma = \{\alpha \in S : v(\alpha) \notin \mathcal{D}\}$ and $\Delta = \{\alpha \in S : v(\alpha) \in \mathcal{D}\}$. Keep in mind that $v_b(\Gamma) = b$, and $v_b(\Delta) = b$. Where $L = \mathbb{L}(V_{LP^+})$, $\Gamma \vdash \Delta$, since either $\alpha \in \Gamma$, and so $v(\alpha) \notin \mathcal{D}$, or $\alpha \in \Delta$, and so $v(\alpha) \in \mathcal{D}$. However, unlike the case where $\mathcal{V} = \{1, 0\}$, $v_b$ also satisfies $\Gamma \vdash \Delta$, so we can not use that partition to rid ourselves of inadmissible valuations. □

## Appendix 3

**Theorem.** *In general, for a valuation space $V \subseteq U$, $V = \mathbb{V}(\mathbb{L}(V))$ (Hjortland, 2014).*

*Proof.* We proceed by supposing that $v_0 \notin V$, with the intent to show that $v_0 \notin \mathbb{V}(\mathbb{L}(V))$. First, define $\Gamma_1 = \{A \in WFF : v_0(A) \neq v_1\}$; $\Gamma_2 = \{A \in WFF : v_0(A) \neq v_2\}$; $\ldots$ ; $\Gamma_n = \{A \in WFF : v_0(A) \neq v_n\}$. As is clear, for each $v \neq v_0$, $v$ satisfies the sequent $\Gamma_1|\Gamma_2|\ldots|\Gamma_n$, since $v \neq v_0$, there is a formula $A$ where $v(A) \neq v_0(A)$. If we assume that $v_0(A) = v_i$, then, for some $j \neq i$, $v(A) = v_j$. By definition, $A \in \Gamma_k$ for each $\Gamma_k$ where $k \neq i$, so $A \in \Gamma_j$. Hence $v$ satisfies $\Gamma_1|\Gamma_2|\ldots|\Gamma_n$. However, $v_0$ fails to satisfy $\Gamma_1|\Gamma_2|\ldots|\Gamma_n$, since if it did, then for some $i$ there is a formula $A \in \Gamma_i$ such that $v_0(A) = v_i$. But, by definition, if $A \in \Gamma_i$, then $v_0(A) = v_i$. Hence, $\Gamma_1|\Gamma_2|\ldots|\Gamma_n$ is $V$-valid since it is satisfied by each $v \neq v_0$, and, because it fails to be satisfied by $v_0$, $v_0 \notin \mathbb{V}(\mathbb{L}(V))$. □

abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 8, Number 1, pp. 75–82, 2014

# Which-Object Misidentification

Max Seeger

Heinrich-Heine-University
Institute of Philosophy
seeger@phil.hhu.de

Keywords: immunity to error through misidentification, self-consciousness, I-thoughts, first person

**Abstract**

This paper examines the relation between *de re* and which-object misidentification. I argue that the most natural reading of which-object misidentification, according to which the two kinds of error are mutually exclusive, is inconsistent with Pryor's claim that immunity to which-object misidentification implies immunity to *de re* misidentification. My argument undermines Pryor's strategy to focus his discussion of error through misidentification on which-object misidentification and raises more general issues concerning the nature of introspective grounds.

A belief of the form '*a* is F' is immune to error through misidentification when it cannot be wrong with respect to the question who or what instantiates (or seems to instantiate) the property in question. For instance, if I believe that I am seeing a canary, based on my visual impression as of a canary in front of me, I may be wrong in many respects, but I cannot be wrong about it being me who is (or seems to be) seeing a canary (cf. Shoemaker 1968: 557). While the exact definition of immunity to error through misidentification is disputed, many authors support the claim that introspection-based self-ascriptions of mental states are so immune (*loci classici*: Wittgenstein 1958, Shoemaker 1968, Evans 1982). Further, I-thoughts which are so immune are generally taken to be fundamental to self-consciousness.

In his seminal (1999) paper, James Pryor distinguishes two varieties of error through misidentification, *de re* misidentification and *which-object* misidentification, and two corresponding varieties of immunity to error through misidentification. This paper examines the relations between both *de re* and which-object misidentification as well as between the two corresponding varieties of immunity. I argue that the most natural reading of which-object misidentification, according to which the two kinds of error are mutually exclusive, is inconsistent with Pryor's claim that immunity to which-object misidentification implies immunity to *de re* misidentification.

## 1 Two Notions of Misidentification

The standard case of misidentification is *de re* misidentification. One of Pryor's paradigmatic cases of error through *de re* misidentification is this.

> [Sam the gun] I see that the blue-coated man is carrying a gun under his coat. Because I've misidentified the blue-coated man as Sam, I form the *de re* belief, of Sam, that he is carrying a gun. (275[1])

[1] All references are to Pryor (1999), unless otherwise noted.

The belief 'Sam is carrying a gun' is in error through *de re* misidentification because its justification rests on justification for the identity belief 'the blue-coated man = Sam'. Generally, a belief of the form '*a* is F' is in error through *de re* misidentification iff its justification rests on justification for a false identity belief of the form '*a* = *b*', where 'a' and 'b' are singular *de re* concepts.

Pryor pointed out that there is another kind of misidentification which he labelled *which-object misidentification*. Here is his well-known example.

> [Skunk] I smell a skunky odor, and see several animals rummaging around in my garden. None of them has the characteristic white stripes of a skunk, but I believe that some skunks lack these stripes. Approaching closer and sniffing, I form the belief, of the smallest of these animals, that it is a skunk in my garden. This belief is mistaken. There are several skunks in my garden, but none of them is the small animal I see. (281)

In this case, justification for the belief 'this animal is a skunk in my garden' does not rest on justification for an identity belief containing two *de re* concepts. Rather, the subject moves from justification for an existential belief ('there is a skunk in my garden') to a singular belief. So, in cases of which-object misidentification, the subject believes a property to be instantiated by something or other, and then goes wrong in singling out who or what is the witness of that property.

To repeat, in *de re* misidentification, justification for the singular belief '*a* is F' rests on justification for the singular belief '*b* is F' and justification for the false identity belief '*a* = *b*'; in which-object misidentification, justification for the singular belief '*a* is F' rests on justification for the existential belief 'something is F' and justification for believing that *a*'s being F is, so to speak, the relevant truthmaker of that existential belief. In cases of *de re* misidentification, the subject's justification for believing the property to be instantiated allows her to entertain a singular *de re* thought about the bearer of the property (e. g. 'the blue-coated man is carrying a gun'). In cases of which-object misidentification, the subject's initial justification for believing the property to be instantiated does not yet allow her to entertain a *de re* thought about the bearer of the property, but only an existential thought (e. g. 'there is a skunk in my garden').

It will ease exposition to distinguish two notions of identification which correspond to the two notions of misidentification (cf. Lafraire 2013). A belief is based on *de re identification* iff its justification rests on justification for an identity belief '*a* = *b*', where 'a' and 'b' are *de re* concepts; a belief is based on *singling-out identification* iff its justification involves justification for the move from an existential belief to a singular *de re* belief (where justification for the existential belief does not yet allow the subject to entertain a *de re* belief).

Some authors have questioned Pryor's distinction or its usefulness, and have proposed alternative distinctions.[2] I am not going into that. Let us assume that there is an interesting difference between cases like Sam the gun and Skunk which can be captured by Pryor's distinction.

## 2 Which-Object Misidentification

The most natural reading of 'which-object misidentification', the one I suggest Pryor had in mind, construes which-object misidentification in a way that renders the two kinds of misiden-

[2] Coliva (2006) argues that the distinction does not reveal a difference in the structure of the grounds, but a difference in the nature of the concepts involved. Recanati (2012) proposes to distinguish between misidentification based on singular grounds and misidentification based on general grounds. Wright (2012) defends Pryor's distinction, and Smith (2006) criticizes the general idea that there are two kinds of immunity.

tification mutually exclusive. It assumes that which-object misidentification necessarily involves a singling-out identification, i. e. a move from justification for an existential belief only, to a singular *de re* belief. I dub this the *disjunctive reading*. As I will show, the disjunctive reading is inconsistent with Pryor's claim that immunity to which-object misidentification implies immunity to *de re* misidentification. Before I make this point, let me establish that the disjunctive reading gets the notion of which-object misidentification right.

First, in describing which-object misidentification, and in particular in contrasting it to *de re* misidentification, Pryor makes several remarks that strongly suggest the disjunctive reading. For instance, he contrasts the singling-out identification in which-object misidentification with reidentification in *de re* misidentification:

> In cases of *de re* misidentification, I form a mistaken belief of *a* that it is F, and my epistemic right to this belief rests on my justification for believing, of some *particular* object *y*, that *y* is F and that *a* and y are the same object. [. . . ] In cases of *wh*-misidentification, I also form a mistaken belief of *a* that it is F, but here my epistemic right to that belief *does not* rest on my justification for believing, of any *other* object, that *it* is F. In cases of *wh*-misidentification, I go wrong not in *reidentifying* the thing I know to be F as some other thing; rather, I go wrong in figuring out *which thing* is F, in the first place. (282 f.)

Second, Pryor explicitly claims that the two kinds of misidentification can occur independently of one another: "*Which*-misidentification can occur without *de re* misidentification. [. . . ] And *de re* misidentification can occur without *which*-misidentification" (285). The first claim is beyond question: cases like Skunk involve which-object misidentification but no *de re* misidentification. It is the second claim that is under discussion in this paper. Strictly speaking, this claim does not imply the disjunctive reading, for this claim is also compatible with a reading that takes which-object misidentification and *de re* misidentification to simply be logically independent. However, as will become clear shortly, the relevant alternative reading holds that any case of *de re* misidentification is also a case of which-object misidentification. The quoted passage supports the disjunctive reading in ruling out this alternative.

Third, attributing the disjunctive reading to Pryor is consistent with the literature. For instance, Lafraire explicitly claims that there can be cases of *de re* misidentification that do not involve which-object misidentification (cf. 2013: 51 f.).[3] The disjunctive reading is arguably also implied by Wright, who writes that in cases of which-object misidentification "a thinker goes into a situation equipped with grounds for a unique existential claim" (2012: 255). Even more explicitly, Prosser states that "[i]n cases of *wh*-misidentification [. . . ] there are no genuine grounds for any *de re* judgment of the form '*b* is *F*', but at most genuine grounds for a judgment of the form $\exists x\, Fx$" (2012: 161).

Thus, it has been shown that the notion of which-object misidentification is best construed in its disjunctive reading. However, I should note that the disjunctive reading is inconsistent with Pryor's explicit definition. To wit, Pryor's definition implies that any case of *de re* misidentification is ipso facto a case of which-object misidentification. This obviously contradicts the disjunctive reading. However, I take this to be a mere lapse. Let me quickly set the definition straight, as this will also provide the necessary background for the ensuing discussion. Here are the original definitions in full.

[3] Again, this supports the disjunctive reading in ruling out the relevant alternative.

> *De re* misidentification occurs when:
>
> [d$^{\text{i}}$] There is some singular proposition about $x$, to the effect that it is F, that a subject believes or attempts to express. [...]
>
> [d$^{\text{ii}}$] The subject's justification for believing this singular proposition rests on his justification for believing, of some $y$, and of $x$, that $y$ is F and that $y$ is identical to $x$. [...]
>
> [d$^{\text{iii}}$] However, unbeknownst to the subject, $y \neq x$. (274 f.)
>
> Wh-misidentification occurs when:
>
> [wh$^{\text{i}}$] A subject has some grounds G that offer him knowledge of the existential generalization $\exists x$ F$x$.
>
> [wh$^{\text{ii}}$] Partly on the basis of G, the subject is also justified, or takes himself to be justified, in believing of some object $a$ that *it* is F.
>
> [wh$^{\text{iii}}$] But in fact $a$ is not F. Some distinct object (or objects) $y$ is F, and it's because the grounds G "derive" in the right way from this fact *about y* that they offer the subject knowledge that $\exists x$ F$x$. (282)

Now, how is which-object misidentification implied in cases of *de re* misidentification? Consider Sam the gun. Pryor takes this to be an example of *de re* misidentification that does not involve which-object misidentification (cf. 285). His idea must be that since there is no singling-out identification, there is no which-object misidentification. But in fact, no singling-out identification is required by Pryor's definition of which-object misidentification. Indeed, Sam the gun does, pace Pryor, satisfy the criteria of which-object misidentification.

Criterion (wh$^{\text{i}}$) is satisfied since I clearly have grounds that offer me knowledge of the existential generalization that someone is carrying a gun. Note that for the criterion to be fulfilled I need not actually entertain the existential belief (cf. 281), but only need grounds that would justify me in believing the existential claim. Criterion (wh$^{\text{ii}}$) is equally satisfied since it is partly on the basis of G (my visual impression of the blue-coated man carrying a gun) that I take myself to be justified in believing of Sam that he is carrying a gun. Finally, criterion (wh$^{\text{iii}}$) is satisfied since, first, it is not in fact Sam, but the blue-coated man, who is carrying a gun, and, second, it is because my visual impression derives in the right way from this fact about the blue-coated man that it offers me knowledge of the existential claim that someone is carrying a gun.

These considerations are not specific to the case. When a subject has grounds for knowledge that a particular object is F, as in cases of *de re* misidentification, then the subject *ipso facto* has grounds for knowledge of the existential belief that someone or something is F. Any case of *de re* misidentification will also satisfy the criteria for which-object misidentification. Put simply, the definition fails to restrict its scope to cases involving a mistaken singling-out identification. Therefore, if we take the definitions literally, *de re* misidentification implies which-object misidentification.[4]

Of course, there is an easy fix for this. All we need to do is add to (wh$^{\text{i}}$) the clause that grounds G offer knowledge for an existential belief *only* (i. e. they do not yet offer knowledge for a *de re* belief). With this amendment, the kinds of grounds we find in cases of *de re*

[4] Actually, that is not quite correct. There are two further differences between the definitions. Only the definition of *de re* misidentification allows for error through misidentification to be compatible with (a) error through misprediction, and (b) accidentally true belief. I think that these differences are inadvertent, but in any case they are not relevant to the issue under discussion.

misidentification do not satisfy ($\text{wh}^1$) and hence these cases are not in error through which-object misidentification. In what follows, I am assuming that the amended definition is the one Pryor had in mind.

## 3 Immunity to Which-Object Misidentification

According to the disjunctive reading, which-object misidentification and *de re* misidentification are mutually exclusive categories. This has direct implications for the relation between the corresponding varieties of immunity. I now show that the disjunctive notion of which-object misidentification is inconsistent with Pryor's claim that which-immunity implies *de re* immunity:

> (whide) Immunity to error through which-object misidentification implies immunity to error through *de re* misidentification.

Consider Pryor's illustration of (whide). He claims, contraposing (whide), that "any beliefs which are vulnerable to *de re* misidentification when justified by certain grounds are ipso facto also vulnerable to *wh*-misidentification when justified by those grounds" (285). As an example he claims that, "in [Sam the gun], where my belief of Sam that he is carrying a gun exemplifies *de re* misidentification, this belief must therefore also be vulnerable to *wh*-misidentification, when believed on the grounds described in the example" (285). Now, Sam the gun is a paradigm example of *de re* misidentification. Pryor further claims that this case does not involve which-object misidentification (cf. 285). How, then, is this case vulnerable to which-object misidentification in virtue of being in error through *de re* misidentification?

What are the grounds in Sam the gun? The justification for the belief 'Sam is carrying a gun' is based on justification for believing 'the blue coated man is carrying a gun' and justification for believing 'the blue coated man = Sam'. But if these are the grounds, then we should expect that, on the disjunctive reading, the belief is immune to error through which-object misidentification. For, on the disjunctive reading, a belief is in error through which-object misidentification only if it involves a false singling-out identification. But the grounds in Sam the gun do not contain a singling-out identification, only a *de re* identification. If the case involves any misidentification at all it must be a *de re* misidentification, given its grounds.

This is a general point. Any case of *de re* misidentification will involve grounds that contain a false *de re* identification. But on the disjunctive reading, a belief cannot be vulnerable to error through which-object misidentification in virtue of being based on a *de re* identification, for *de re* misidentification and which-object misidentification are mutually exclusive. Rather, cases of *de re* misidentification are immune to which-object misidentification, because the grounds of the belief do not contain a singling-out identification.[5]

A different way of making the same point is this. Given the amended definition of which-object misidentification, a judgment is in which-object misidentification only if it is based on identity-neutral grounds (cf. Coliva 2006: 411). That is to say, for a judgment to be in which-object misidentification its justification must be partially based on grounds that allow for an existential claim only. However, cases of *de re* misidentification precisely do not involve

[5] More precisely, this only holds for regular cases of *de re* misidentification. There may of course be beliefs that involve both kinds of misidentification. Suppose, for instance, that I have justification for believing that someone or other called my name, go wrong in singling out *this man* demonstratively as the one who called, and make a further mistake in believing of this man that he is Sam. My eventual belief that Sam called my name would then be both in error through which-object misidentification (singling out *that man*) and in error through *de re* misidentification (taking this man to be Sam). However, the belief's vulnerability to which-object misidentification is independent of the belief's being in error through *de re* misidentification. In regular cases of *de re* misidentification, like Sam the gun, the grounds do not involve a singling-out identification and are hence immune to which-object misidentification.

identity-neutral grounds, but grounds that allow for a *de re* judgment. Hence contrary to Pryor's claim, immunity to which-object misidentification does not imply immunity to *de re* misidentification.

# 4 Implications

I argued that Pryor's notion of which-object misidentification, according to which *de re* misidentification and which-object misidentification are mutually exclusive categories, is inconsistent with his claim that which-immunity implies *de re* immunity. I finish by pointing out the broader philosophical import of my discussion.

To begin with, let me quickly set an exegetical issue straight. Pryor argues that both Wittgenstein and Shoemaker are best interpreted as having had immunity to which-object misidentification in mind rather than immunity to *de re* misidentification (cf. 286–288). I think Pryor is misled in this point by his claim (whide). Given that the two kinds of immunity are logically independent, it is quite clear that Shoemaker had immunity to *de re* misidentification in mind. First, Shoemaker's definition of immunity explicitly requires that "the speaker knows *some particular thing* to be φ" and not just that the speaker knows something or other to be φ (1968: 557; my emphasis). But more importantly, Shoemaker has in fact considered a case of which-object misidentification long before Pryor introduced these cases to the debate. However, Shoemaker's stance was to *not* count such cases as in error through misidentification (cf. 1970: 270, fn. 4).

Moving on to the substantial points, consider how (whide) figures in Pryor's discussion of immunity to error through misidentification. Pryor's main strategy is to focus on immunity to which-object misidentification, because this, he claims, is the more basic and hence more interesting epistemic status (cf. 272, 286, 287). Pryor's idea that which-immunity is more fundamental is based precisely on the claim that which-immunity implies *de re* immunity: "immunity to *wh*-misidentification entails, but is not entailed by, immunity to *de re* misidentification. In that sense, immunity to *wh*-misidentification is a more basic and more rare epistemic status" (286). Pryor's strategy to focus on which-immunity plays an important role for instance in his discussion of quasi-memory. Pryor defends (against Evans) Shoemaker's view that quasi-memory undermines the immunity of memory-based self-ascriptions. In one of his objections, Pryor claims that even if Evans has shown that memory-based judgments are immune to *de re* misidentification, they are not immune to which-object misidentification (cf. 293). The strength of this objection depends crucially on the question whether which-immunity really is the more interesting phenomenon.

Moreover, Pryor's claim that memory judgments are vulnerable to which-object misidentification presupposes that memory-judgments involve identity-neutral grounds. As Coliva points out: "which-misidentification [applies] to memory-based judgments only if their grounds are identity-*neutral*" (2006: 411; Coliva's emphasis). But many authors find that memory necessarily presents remembered experiences as one's own.[6] More generally, many authors find that introspection in general and also proprioception necessarily present one's states as one's own. A claim that figures centrally in the debate on immunity is precisely the idea that introspection and proprioception do not leave open or even allow for the question whose states

[6] Pryor addresses this worry by claiming that it is possible to partially undercut memory judgments in a way that leaves intact grounds for an existential claim (cf. 296 f.). However, it has been objected that the kind of defeater suggested by Pryor, namely telling the subject "that some of his memories are quasi-memories of events in someone else's past life" (295), brings in new grounds for an existential claim rather than leaving intact the original grounds (cf. Smith 2006, see also Coliva 2006: 409).

one is being aware of (see e. g. Shoemaker 1968; Evans 1982; Coliva 2002; Wright 2012). Given that the immunity thesis is a thesis about judgments based on introspection (or, more generally, based on first-personal awareness, including bodily awareness), and hence about judgments that do not involve identity-neutral grounds, it may be worried that Pryor's notion of which-object misidentification simply does not apply to the interesting cases.

Of course, things aren't quite that simple. Although *normally* introspection doesn't leave open any question of identification, two kinds of unusual cases have come to dominate the debate. First, we have cross-wiring cases, that is cases in which one subject's perceptual, proprioceptive, or mnemonic experiences derive in a causally deviant way from another subject (see e. g. Shoemaker 1970, Evans 1982: 184–189, Smith 2006: 278, Chen 2009: 29ff., and Langland-Hassan (2014)). Second, we have pathological cases such as thought insertion, in which a subject is introspectively aware of a mental state but claims of that state that it is not his own (see e. g. Campbell 1999, Gallagher 2000, Coliva 2002, Lane & Liang 2011, de Vignemont 2012). There is a lot of controversy surrounding the question whether subjects in cross-wiring cases or in pathological cases have grounds for a *de re* ascription, an existential claim, or any grounds at all. No matter how this issue is resolved, my main point will stand: given that *de re* misidentification and which-object misidentification are mutually exclusive, which-immunity does not imply *de re* immunity and the discussion of immunity to error through misidentification cannot be reduced to the discussion of which-immunity.

## References

Campbell, J. (1999), 'Schizophrenia, the Space of Reasons, and Thinking as a Motor Process', *Monist* **82** (4), 609–625.

Chen, C. K. (2011), 'Bodily Awareness and Immunity to Error through Misidentification', *European Journal of Philosophy* **19** (1), 21–38.

Coliva, A. (2002), 'Thought Insertion and Immunity to Error Through Misidentification', *Philosophy, Psychiatry, & Psychology* **9** (1), 27–34.

Coliva, A. (2006), 'Error through Misidentification: Some Varieties', *Journal of Philosophy* **103** (8), 403–425.

Evans, G. (1982), *The Varieties of Reference*, Oxford University Press, Oxford.

Gallagher, S. (2000), Self-Reference and Schizophrenia: A Cognitive Model of Immunity to Error through Misidentification, *in* D. Zahavi, ed., 'Exploring the Self: Philosophical and Psychopathological Perspectives on Self-Experience', John Benjamins, Amsterdam, pp. 203–239.

Lafraire, J. (2013), 'Two Notions of (Mis)-Identification', *Philosophical Inquiries* **1**, 39–53.

Lane, T. & Liang, C. (2011), 'Self-Consciousness and Immunity', *Journal of Philosophy* **108**, 78–99.

Langland-Hassan, P. (2014), 'Introspective Misidentification', *Philosophical Studies*. doi.org/10.1007/s11098-014-0393-x.

Prosser, S. (2012), Sources of immunity to error through misidentification, *in* Prosser & Recanati, eds. (2012), pp. 158–179.

Prosser, S. & Recanati, F., eds. (2012), *Immunity to Error Through Misidentification: New Essays*, Cambridge University Press, New York.

Pryor, J. (1999), 'Immunity to Error through Misidentification', *Philosophical Topics* **26** (1 & 2), 271–303.

Recanati, F. (2012), Immunity to error through misidentification: what it is and where it comes from, *in* Prosser & Recanati, eds. (2012), pp. 180–201.

Shoemaker, S. (1968), 'Self-Reference and Self-Awareness', *Journal of Philosophy* **65** (19), 555–567.

Shoemaker, S. (1970), 'Persons and Their Pasts', *American Philosophical Quarterly* **7** (4), 269–285.

Smith, J. (2006), 'Which Immunity to Error?', *Philosophical Studies* **130** (2), 273–283.

de Vignemont, F. (2012), Bodily Immunity to Error, *in* Prosser & Recanati, eds. (2012), pp. 224–246.

Wittgenstein, L. (1958), *The Blue and Brown Books*, Blackwell, Oxford.

Wright, C. (2012), Reflections on Recanati's 'Immunity to error through misidentification: what it is and where it comes from', *in* Prosser & Recanati, eds. (2012), pp. 247–280.