Proceedings of the Amsterdam Graduate Philosophy Conference - Truth, Meaning, and Normativity

Amsterdam, September 30-October 2, 2010

Crespo, M.I.; Gakis, D.; Weidman-Sassoon, G.

Publication date
2011

Document Version
Final published version

Proceedings of the

Amsterdam Graduate Philosophy Conference

—Truth, Meaning, and Normativity—

Amsterdam, September 30 – October 2, 2010

Proceedings of the Amsterdam Graduate Philosophy Conference —Truth, Meaning, and Normativity—

Amsterdam, September 30 – October 2, 2010

María Inés Crespo, Dimitris Gakis,
and Galit Weidman-Sassoon (eds.)

Department of Philosophy/ILLC
Universiteit van Amsterdam

Preface

The 3^{rd} Amsterdam Graduate Philosophy Conference on "Truth, Meaning, and Normativity" was organised by the Department of Philosophy and the Institute for Logic, Language and Computation of the Universiteit van Amsterdam. The conference invited submissions from graduate researchers conducting novel philosophical research into any of the three conference topics. Some of the papers in this volume inform the discussion about truth, meaning, and/or normativity by offering a philosophical interpretation of results from other fields such as logic, cognitive psychology and formal semantics. A typical example for this is Cova and Égré's paper, including experimental results about the semantics of 'many' as a gradable adjective and their variety of philosophical implications.

Another area of interest is semantic normativity with respect to meaning, use, content, and context. This topic was taken up by Belleri's work on predicates of personal taste. Other dominant topics included formal theories of truth and deflationism, dealt with in the majority of papers in this volume including those by Gruber, McKinnon and Speck.

Since the topics of truth, meaning, and normativity naturally feed into each other, some contributions explore several of the intricate ways in which these notions relate to one another. We include here Wieland and Turbanti as representative authors.

The organisers were María Inés Crespo, Dimitris Gakis, and Galit Weidman-Sassoon, and they consulted: Dr. Maria Aloni, Dr. Paul Dekker, Dr. Catarina Dutilh-Novaes, Prof. Dr. Jeroen Groenendijk, Prof. Dr. Michiel van Lambalgen, Dr. Benedikt Löwe, Dr. Robert van Rooij, Prof. Dr. Martin Stokhof, and Prof. Dr. Frank Veltman. Our Programme Committee included: Denis Bonnay, Filip Buekens, Fabrice Correia, Paul Égré, Henri Galinon, Manuel García Carpintero, Jussi Haukioja, Michael Hegarty, Wolfram Hinzen, Ole

Hjortland, David Hunter, Vasso Kindi, Mikhail Kissine, Max Kölbel, Kepa Korta, Michiel van Lambalgen, Daniel Lassiter, Hannes Leitgeb, Reinhard Muskens, Daniele Porello, François Recanati, David Ripley, Olivier Roy, Sebastian Sequoiah-Grayson, Isidora Stojanovic, Martin Stokhof, Peter Pagin, Luis Urtubey, Stelios Virvidakis, and Åsa Wikforss. We are most grateful to the following sponsors for their support: Institute for Logic, Language, and Computation, Leerstoelgroep Logica en Taalfilosofie, Afdeling Wijsbegeerte, NWO Projects: "The Origins of Truth, and the Origins of the Sentence", "Indefinites and beyond: Evolutionary pragmatics and typological semantics", "The Inquisitive Turn: A New Perspective on Semantics, Logic, and Pragmatics", NWO & LogiCCC ESF Project "Vagueness, Approximation, and Granularity", Allard Pierson Museum, and Gemeente Amsterdam.

We appreciate and acknowledge local support by the conference's Steering Committee: Theodora Achourioti, Edgar Andrade-Lotero, and Marc Staudacher. On finances and administration, we thank the ILLC buro, in particular Peter van Ormondt, Ingrid van Loon, and Marco Vervoort. We also wish to thank Stéphane Airiau and Joel Uckelman for their technical assistance in the edition of the present volume.

Finally, we want to thank the speakers for their contributions.

The Editors,
Amsterdam, March 2011

Contents

Crespo, M.I., Gakis, D., Sassoon, G. W. (eds.),
Proceedings of the Amsterdam Graduate Philosophy Conference
— Truth, Meaning, and Normativity —
ILLC Publications X-2011-01, 1–18

Moral asymmetries and the semantics of *many*

Florian Cova and Paul Égré [1]
(Institut Jean-Nicod (CNRS-ENS-EHESS))

paulegre@gmail.com and florian.cova@gmail.com

We present the results of two experiments concerning the evaluation people make of sentences involving "many", showing that two sentences of the form "many As are Bs" vs. "many As are Cs" need not be equivalent when evaluated relative to a background in which B and C have the same cardinality and proportion to A, but in which B and C are predicates with opposite semantic and affective value. The data provide evidence that subjects lower the standard relevant to ascribe "many" for the more morally negative predicates. We relate the results to similar semantic asymmetries discussed in the psychological literature, in particular to the Knobe effect and to framing effects.

1 Introduction

The aim of this paper is to investigate the semantics of the quantifier "many" in relation to a family of moral asymmetries that have been documented in various places in the literature.

[1] Corresponding author: Paul Égré. Thanks to audiences in Paris (EALING) and Amsterdam (AGPC) for questions and comments. We are grateful to A. Bachrach, E. Chemla, M. Cozic, B. Geurts, N. Hansen, M. Kneer, J. Knobe, H. Leitgeb, E. Machery, S. Pighin, D. Ripley, R. van Rooij, G. W. Sassoon, P. Schlenker, B. Spector, F. Veltman and S. Yalcin for various helpful exchanges, suggestions or comments, as well as to two reviewers of previous work. This work was done with the support of the ANR project "Cognitive origins of vagueness" (ANR-07-JCJC-0070).

The first of those is an asymmetry evidenced by J. Knobe (2003b) regarding people's ordinary judgments about intentional action. Knobe presented the following scenario to two groups of subjects. One group read the scenario with the word "harm", the other group with "help" uniformly in place of "harm":

> 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/help] the environment.' The chairman of the board answered, 'I don't care at all about [harming/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 [harmed/helped].

Subjects in each group then had to respond by yes or no to the following question:

(1) Did the chairman intentionally [harm/help] the environment?

In the harm condition, a large majority of subjects agreed that the chairman intentionally harmed the environment. In the help condition, by contrast, most subjects denied that the chairman intentionally helped the environment. This effect is surprising, since in each scenario, the chairman exerts the same influence on the outcome, and is described as equally informed and equally indifferent toward the side-effect.

In a recent paper, Pettit and Knobe (2009) have outlined a convincing explanation for this asymmetry, based on an analogy with the semantics of gradable adjectives. What Pettit and Knobe point out is that two liquids, coffee and beer, can be at the same temperature, of say 20°C, but be such that one would judge the first to be cold and the second not to be cold. To judge that the coffee is cold is to judge that it is colder than it should be, given the expected temperature for coffee; to deny that the beer is cold is to judge that the beer is not as cold as it should be, given the expected temperature for beer. Phrased in terms of degrees, this is equivalent to saying that the degree to which coffee is cold exceeds the norm or standard relevant to ascribe coldness for coffee; to deny that beer is cold is to judge that the degree to which beer is cold is below the norm relevant for beer. By analogy, to judge that an action type is done intentionally is to judge that the degree of intention attached to the action is above the normative threshold relevant for that kind of action. In

the same way in which the threshold for "cold" can vary from beer to coffee, the threshold for "intentional" can thus vary from "harm" to "help" along the dimensions relevant in Knobe's scenario. Thus, although the chairman's internal properties and causal influence are the same in each condition, whether an action is described as harm or help makes different standards of comparison salient in order to judge whether that action was done intentionally.

Further evidence was proposed in Égré (2010a), (2010b) to articulate and substantiate Pettit and Knobe's explanation. Basically, the suggestion is that the Knobe effect might be an instance of a more general asymmetry concerning our expectations between negatively valued vs. positively valued outcomes. This asymmetry, in particular, has been documented in the psychological literature concerning people's perception of risk, starting with Tversky and Kahneman's experiments on framing effects (Kahneman and Tversky (1979), Tversky and Kahneman (1981)), and including what Weber and Hilton (1990) describe as a "worry effect". In a recent study on risk communication, Pighin et al. (2009) compared the rankings given by four groups of pregnant women concerning the probability of [1 in 307/1 in 28] that a particular will have [Insomnia/Down syndrome], on a 7 point scale ranging from "extremely low" to "extremely high". What they found is that subjects ranked significantly higher the lower probability of 1 in 307 for the child to have Down syndrome, in comparison to the probability of 1 in 28 for the child to have Insomnia. Those answers were found to correlate with how severe they judged each disease to be. This phenomenon, also known as the *severity bias* (see S. Pighin (2009), Bonnefon and Villejoubert (2006), suggests that the same mechanism operates in judgments about whether an action is intentional and in judgments about whether a probability is high. In the latter case, two identical probability values on the scale from 0 to 1 can be such that the first will be judged to be high in comparison to the standard relevant for a severe disease, while the other will be judged not to be high in comparison to the standard relevant for a non-severe disease. Contrasts in judgments correspond to shifts of the standard relevant in each domain.

In Égré (2010a), the hypothesis was formulated that one should observe essentially the same kind of asymmetry in judgments about quantities expressed in terms of the vague quantifier "many". That

is, the prediction was made that judgments should be found to differ in pairs of the form "many As are Bs" vs "many As are Cs" for B and C with identical proportion to A, depending on how the B and C outcomes are valued. In what follows, we present experimental confirmation of this prediction. The way we tested for the prediction involves two steps. In a first experiment, we simply asked people to assent or dissent to such a pair of sentences for a specific scenario in order to probe their truth-conditional intuitions. In a second experiment, we looked for information about people's positioning of the threshold in relation to their judgments about "many". One advantage of this methodology is that it fits the way in which judgments involving gradable expressions can be modeled, essentially in terms of a comparison to an implicit normative standard (see Sapir (1944), Bartsch and Vennemann (1972), Fara (2000), Lappin (2000), Kennedy (2007)). Furthermore, because the scale associated to "many" is more transparent than the one for "intentional", the data give us insight into the way in which norms and expectations determine our judgments involving vague predicates.

In the first part of this paper, we start with some background on the semantic analysis of the severity bias and the Knobe effect, in terms of shifting standards of comparison, and extend this analysis to sentences of the form "many As are Bs". In the second part, we present our experiment and show that the data comport with this semantic model. In the last part, we conclude with some considerations about the relation of our data with Kahneman and Tversky's prospect theory on the one hand, and about the link between Knobe-type asymmetries and framing effects on the other.

II Shifting standards and the semantics of *many*

II.1 The severity bias and the Knobe effect

Pighin et al.'s (2009) data on judgments about probabilities in relation to the scale involving the predicates "high" and "low" indicate that (2-a) can be judged true and (2-b) false in the same context without inconsistency:

(2) a. A probability of 1/307 for a child to have Down syndrome is high

b. A probability of 1/28 for a child to have Insomnia is high

The contrast between the two judgments can be represented by means of a semantics *à la* Bartsch and Vennemann, assuming that "high" is a predicate that maps individuals to degrees, and for a probability to be high is for that probability to be higher than the norm for highness in relation to the kind of event under consideration. Let p denote the probability 1/307 to have Down Syndrom and p' denote the probability 1/28 to have Insomnia and compare:

(3) a. $[\![high]\!]_w(p) \succ \mathbf{norm}_w(DownSyndrom)(high)$
b. $[\![high]\!]_w(p') \succ \mathbf{norm}_w(Insomnia)(high)$

Assume for simplicity that

$$[\![high]\!]_w(p) = 1/307 \text{ and } [\![high]\!]_w(p') = 1/28,$$

namely that the degrees of highness in the context w are identical to the numerical probabilities, but that

$$\mathbf{norm}_w(DownSyndrom)(high) = 1/1000$$
and that
$$\mathbf{norm}_w(Insomnia)(high) = 1/15.$$

This is a situation in which (3-a) is true and (3-b), false.

In agreement with Pettit and Knobe's remarks, essentially the same analysis can be given for Knobe's examples based on the adjective "intentional". Thus, one can consistently judge that:

(4) a. The harm brought about to the environment by the chairman was intentional
b. The help brought about to the environment by the chairman was not intentional

assuming the standard of comparison for whether an action type is "intentional" is set lower for "harming the environment" than for "helping the environment" on the relevant scale of comparison. Let h stand for "the harm brought about to the environment" and h' for "the help brought about to the environment", and assume that $[\![intentional]\!]_w(h) = [\![intentional]\!]_w(h')$, namely that the degrees of intention attached to each action type are identical. Letting the standard shift from one case to the other, it is possible to have:

(5) a. $[\![intentional]\!]_w(h) \succ \mathbf{norm}_w(Harm)(intentional)$
b. $[\![intentional]\!]_w(h') \preceq \mathbf{norm}_w(Help)(intentional)$

Prima facie, the Pighin pair and the Knobe pair suggest that the more detrimental an event type is perceived to be, the lower the standard will tend to be positioned for interest-relative predicates such as "high" or "intentional". This hypothesis calls for some qualifications, however.

A first caveat concerns the fact that the validity of the hypothesis depends on the polarity of the adjective under consideration. Obviously, for the negative adjective "unintentional", one would expect the threshold to be lower for "help" than for "harm", namely for the beneficial outcome in this case (similarly, *mutatis mutandis*, if "low" were used instead of "high" to qualify probabilities). A second issue, of more methodological nature, concerns the fact that neither Knobe's experiment, nor Pighin's experiment provide us with much information about how subjects locate the thresholds relative to each other in either condition. In the case of the adjective "intentional" as applied to action types, the problem is intrinsically more complex than for "high" as applied to probabilities, since the structure of the associated scale of comparison is not transparent in this case. But even for "high", Pighin et al.'s study does not allow us to see how far the thresholds will be located apart from each other depending on the kind of disease under consideration.

II.2 *Many*

To get information of that kind, we selected a pair of sentences involving the vague quantifier "many" as applied to the count noun "children", in order to have a simple and discrete scale of comparison, that is the scale of natural numbers with their usual ordering. Secondly, we designed the experiment so as to get both within-subject and between-subject information about the relative position of the threshold for "many" in relation to two distinct predicates, the predicate "survive", and the predicate "die", one denoting a positively valued event type, the other a negatively valued event type. Finally, as pointed out of "low" vs. "high" or "unintentional" vs. " intentional", we note that we do expect the main prediction to be reversed if we had picked "few" or "not many" instead of "many", that is, the threshold to be lower for the less negative predicate. We did

not control for that prediction, however, and chose to focus only on "many", rather than its antonym.

Before getting to the details, we first rehearse a few basic facts about the semantics of "many" in sentences of the form "many As are Bs". For the most part, the semantics of the quantifier expression "many" obeys the same pattern as that already introduced for the gradable adjective "high". Basically, to say that "many As are Bs" is to consider that there are more ABs than what is expected or normal in a given context (see Sapir (1944), Keenan and Stavi (1986), Lappin (2000)). As in the case of gradable adjectives like "high", the threshold for "many" is vague and context-dependent. Moreover, "many" is not purely extensional, but is intensional (see Lappin (2000)). This means that this threshold is similarly sensitive to the meaning of its arguments, namely to which comparison class is specified by the restrictor of the quantifier as well as by its nuclear scope. To take an example given by Lappin, (6-a) can be judged false and (6-b) true in a situation in which there are as many violinists as musicians, and as many women as Italians:

(6) a. Many musicians at the concert are women.

a. Many violinists at the concert are Italian.

Suppose that there are 100 musicians and violinists at the concert, including 30 women, and 30 Italians. Then (6-a) may be judged false if the normative threshold for women to count as many musicians when there are 100 musicians is 50, and (6-b) may be judged true if the normative threshold for Italians to count as many violinists when there are 100 violinists is 20.

By analogy with the truth-conditions given earlier for "high", we propose that "many As are Bs" is true in context w provided:

(7) $|A|_w \cap |B|_w \succ \mathbf{norm}_w(A, B, |A|_w)$

This says that many As are Bs provided the actual number of ABs is greater than the norm or expected value for As and Bs relative to the actual cardinality specified by the restrictor A. As assumed for the semantics of gradable adjectives given above, this normative value varies depending on further contextual elements relevant in w besides the three main arguments (see the Concluding remarks

in section 5 below). Also, the reason we single out the cardinality of the restrictor, rather than of the nuclear scope or both, is because in the examples we will focus on, subjects base their judgments foremost on information they receive about the cardinality of the restrictor, making it an essential parameter to how they ascribe "many". The truth-conditions laid out here agree with those stated by Lappin (2000), in particular they take account of the intensionality of "many" and of the fact that "many" is nonsymmetric with regard to its arguments ("many As are Bs" does not entail "many Bs are As").[2]

In the study we conducted, we asked subjects to evaluate two sentences of the form:

(9) a. Many children died

b. Many children survived

by setting a scenario in which the number of dead children and of children surviving were identical. In line with the severity bias, the prediction formulated in Égré (2010a) was that subjects would more readily assent to the first sentence than to the second, thus establishing a shift in the threshold for "many" depending on the predicate under consideration. Like "harm the environment" and "help the environment" in Knobe's scenario, or "getting Down Syndrom" and "getting Insomnia" in Pighin's scenario, the two predicates "die" and "survive" denote event types with opposite affective values. Moreover, "survive" and "die" are arguably contradictories, assuming "survive" is semantically analyzable as "not die". As applied to the predicate "children", finally, they produce a high contrast in expectations.

[2] Lappin's parametric semantics for "many" is given by the following clause, where $N(w)$ denotes the set of normative situations that are relevant relative to the context w:

(8) $[\![many]\!]_w = \lambda P \lambda Q \forall w' \in N(w)(|P|_w \cap |Q|_w \succ |P|_{w'} \cap |Q|_{w'})$

That is, many As are Bs provided the actual number of ABs is above the number of ABs in each normative alternative to the actual world. (7) is one way of rewriting (8), letting $\mathbf{norm}_w(A, B, |A|_w) := min\{|A|_{w'} \cap |B|_{w'}; w' \in N(w)\}$. Note that (8) provides symmetric truth-conditions for "many As are Bs" and "many Bs are As", but Lappin retrieves nonsymmetry by imposing appropriate constraints on $N(w)$.

To compare those predictions to the semantics laid out above for "many", we designed two experiments. In the first experiment, we merely probed for subjects truth-conditional intuitions in relation to the two target sentences involving "many", in order to test for the occurrence of a contrast between the two sentences. In the second experiment, run on a different population, we asked subjects to provide explicit information about the numerical threshold relevant to ascribe "many" for appropriate counterparts of the target sentences of Experiment 1.

III Experiments and results

The two experiments were run on French speaking subjects and French text used in place of the English translations provided here. "Many children died" and "many children survived" in particular translate "beaucoup d'enfants sont morts" and "beaucoup d'enfants ont survécu" respectively (literally: *beaucoup de = many of*).

III.1 Experiment 1

III.1.a Method

In this experiment, we used the following statement:

> 10 children were present in a school when a fire broke out. 5 children survived, the other 5 died.

50 participants were recruited in the Laboratoire de Sciences Cognitives et Psycholinguistique in Paris. 32 were women and the age mean was 23.8. Half of the participants first were given the following question:

> Would you say that many children survived? ("YES" or "NO")

Then they got a second question:

> Would you say that many children died? ("YES" or "NO")

The other half got the same two questions, but in reverse order. So, we had one variable (the answer) and two factors: the type of *predicate* ("SURVIVED" and "DIED") and the *order* ("FIRST" or "SECOND").

III.1.b Results

The percentages of positive answers by condition are summarized in Table 1. We used a two-factor ANOVA with repeated measures.

There was a main effect of *predicate* ($F(1, 48) = 153.3$, $p < .001$) but no main effect of *order* ($F(1, 48) = 0.4$, $p = .52$). There was also a marginally significant interaction between the two factors ($F(1, 48) = 3.4$, $p = .07$).

Order	Predicate: Died	Predicate: Survived
'Died' First	100%	28%
'Died' Second	92%	12%

Table 1: Percentage of positive answers by condition for Experiment 1

III.2 Experiment 2

III.2.a Method

In this experiment, we used the first part of the statement used in Experiment 1, namely "10 children were present in a school when a fire broke out". The subjects were then given the following questions:

1. From which number of children being dead would you say that many children died?

2. From which number of children having survived would you say that many children survived?

40 participants were recruited in the Laboratoire de Sciences Cognitives et Psycholinguistique in Paris. 34 were women and the age mean was 22.8. Half of participants received both questions in one order and the other half in the reverse order.

III.2.b Results

As in Experiment 1, we had one variable (the answer) and two factors: the type of *predicate* ("SURVIVED" and "DIED") and the *order* ("FIRST" or "SECOND"). The mean answers by condition are summarized in Table 2. We used a two-factor ANOVA with repeated measures. There was a main effect of *predicate* ($F(1, 38) = 51.2$, $p < .001$) but no main effect of *order* ($F(1, 38) = 0.1$, $p = .90$) and no interaction effect ($F(1, 38) = 0.1$, $p = .70$).

Order	Predicate: Died	Predicate: Survived
'Died' First	$3,5$	$7,0$
'Died' Second	$3,6$	$7,2$

Table 2: Mean answers by condition for Experiment 2

III.3 Interpretation

In Experiment 1 we set equal cardinalities for the number of children dying and of children surviving, namely 5, and we ensured that each would correspond to a ratio of 1/2 in proportion to the total number of children. As discussed in particular by Partee (1989), "many" is possibly ambiguous between a cardinal reading and a proportional reading. We selected figures so as to make the difference neutral with regard to the prediction we wanted to test.

The first observation to make about Experiment 1 is that it confirms the prediction at issue, that is subjects were much more willing to use "many" in relation to the most negatively loaded of the two sentences, despite the fact that in the context under discussion the two sentences express the same proposition. From a semantic point of view, the results therefore confirm the fact that "many" does not behave purely as an extensional quantifier: by this we mean that the evaluation of "many As are Bs" is not sensitive merely to the cardinality of As, Bs or to the ratio of ABs to As.

Secondly, we can see only a slight tendency for subjects to be more willing to say that "many children survived" when the question comes second. The lack of significant order effect indicates that subjects are little prone to readjusting the respective threshold they associate to each predicate depending on their previous answer. This observation is more amply confirmed by the results of Experiment 2, where the interaction between the two conditions disappears.

In Experiment 2, subjects generally diverged in how they positioned the thresholds for each predicate between 1 and 10, consistently with the fact that "many" is a vague quantifier. Few subjects, however, picked identical numbers for the two predicates "died" and "survived" (4 subjects out of 40), and few set the standard higher

for "died" than for "survived" (3 out of 40). That is, most subjects (the remaining 33, viz. 82.5%) introduced a gap between the two thresholds and selected a lower threshold for the more negatively valued predicate "die".

Taken together, the data of Experiments 1 and 2 are consistent with the truth-conditions laid out in (7). In particular, assuming a sufficiently large sample of subjects in experiment 1 can be described as subjects for whom, in the fire and school context w under discussion,

$$\mathbf{norm}_w(children, die, |children|_w = 10) < 4$$
$$\text{and}$$
$$\mathbf{norm}_w(children, survive, |children|_w = 10) \geq 7$$

then the contrast in truth-values between (9-a) and (9-b) follows when

$$|children|_w \cap |die|_w = |children|_w \cap |survive|_w = 5.$$

For those subjects, given the actual number of dead children specified in Experiment 1, this means that they would have expected more children to survive and fewer children to die.

IV Comparison with prospect theory

The data we obtained about "many" bear a striking connection with empirical evidence at the origin of Kahneman and Tversky's prospect theory (see Kahneman and Tversky (1979)). At the bottom of prospect theory is the observation of an asymmetry between losses and gains, namely the observation that "losses loom larger than gains" (Kahneman and Tversky, 1979, p. 297). This connection is not entirely fortuitous, since Weber and Hilton's work on the so-called severity bias, as well as Pighin's subsequent work on it, are both antedated by Tversky and Kahneman's findings about framing effects. One of the interests of the data here presented concerning "many", however, is that they do not involve the representation of probabilistic uncertainty, but only of quantities presented as certain outcomes (viewed as what Kahneman and Tversky call prospects, they are therefore reducible to their utility components).

Another element of convergence, moreover, concerns the relativity of losses and gains to what Kahneman and Tversky call the *reference point* in their theory. On Kahneman and Tversky's theory, the value of prospects is relative to a zero position on the scale, with regard to which negative or positive deviations are evaluated as gains or losses. It is possible to envisage the results of our experiment 1 in terms of that specific notion of reference point, considering 0 children dying or surviving (0 children involved) to be the reference point, and that 5 dead children is seen as a relative loss, and 5 children surviving as a relative gain. The maximal gain, in that perspective, is 10 children surviving, and the maximal loss 10 children dying. A further ingredient of Kahneman and Tversky's theory is the idea that the value function associating positive and negative utilities to numerical gains and losses is steeper for losses than for gains. Under those assumptions, it is possible to conceive of people's judgment about "many" in relation to that value function v. In other words, it would mean that $v(5) < -v(-5)$. However, to account for the data, one needs to consider that ascriptions of "many" will depend on a common threshold t on the scale of absolute values, such that $v(5) < t < -v(-5)$.

This representation of the situation is amply motivated, but it strikes us as more cumbersome from a semantic point of view than the one we used, in which we used a single axis to represent quantities, and simply postulated different normative standards for the ascription of "many" along that single axis, depending on the predicates. Furthermore, although Kahneman and Tversky's notion of reference point bears some affinity with the notion of standard of comparison in play in the semantics of gradable expressions, it is not exactly the same notion. Prima facie, Tversky and Kahneman motivate the notion in ways that fit the very example Pettit and Knobe use to account for the Knobe asymmetry about ascriptions of "intentional" to action types, that is, they write ((1979), p. 277):

> When we respond to attributes such as brightness, loudness, or temperature, the past and present context of experience defines an adaptation level, or reference point, and stimuli are perceived in relation to this reference point. Thus, an object at a given temperature may be experienced as hot or cold to the touch depending on the temperature to which one has adapted. The same principle applies to non-sensory attributes such as health, prestige, and wealth. The same level of wealth, for example, may imply abject

> poverty for one person and great riches for another-depending on their current assets.

As Kahneman and Tversky acknowledge later in that paper, however, the reference point is initially seen by them as a status quo position, but they point out that "there are situations in which gains and losses are coded relative to an expectation or aspiration level that differs from the status quo" ((1979), p. 286). This notion of expectation level is in fact the right counterpart to the notion of standard of comparison, or normative threshold, relevant for the semantics of gradable expressions and that we used to account for the semantics of "many". Strictly speaking, Kahneman and Tversky's reference point should be viewed primarily as a neutral, zero value on the scale, distinct from the notion of normative threshold, which can vary depending on the property ascribed along that scale.

Setting aside these theoretical differences, the asymmetry we found in moral judgments about "many" appears to bear more than a family resemblance with the asymmetries uncovered by Knobe in judgments about intentional action. Each time, opposite judgments can be derived from a shift in expectations that depends on the property of which the predicate is predicated ("help" vs "harm" in Knobe's scenarios, "survive" vs "die" in ours). This connection is valuable, since it sets the Knobe effect on a continuum with so-called framing effects, and it shows that both kinds of effect are susceptible of semantic analysis using familiar semantic tools.

V Concluding remarks

The asymmetry we examined in judgments about "many" confirms that judgments about whether "many As are Bs" depend on more than the cardinality of As and Bs and their ratios to one another. They depend on expectations that are sensitive to the meaning of A and B.

An important qualification to make is that these expectations in turn will vary depending on the context and the practical interests of the speakers. For instance, it would be easy enough to manipulate the context so as to obtain different judgments from the ones we obtained. Suppose we inform subjects that fires in schools are very frequent in a particular region of the world, and that on average, only 3 children in 10 survive in case of a school fire. In a context in which 5 children died and 5 survived out of 10, subjects would most

likely deny that many children survived, but should be less prone to accepting that many children died. This agrees with the idea that expectations are not purely based on moral considerations, but that frequency facts also have an impact (compare with Mandelbaum and Ripley (2010), who make this kind of objection to Knobe's emphasis on moral norms). Moreover, in some cases the practical interests of the evaluator of a sentence can be at odds with the default moral expectation. For a dangerous pyromaniac expecting to destroy at least 9 children out of 10, it would be true that "many children survived", and false that "many children died" when only 5 children died in the fire.

These two examples, however, only show that the standards for whether "many As are Bs" can easily be manipulated. The important point for us, however, is the fact that even as these expectations shift, they remain systematically sensitive to semantic properties of scales as associated with the predicates used, and to what we may call default expectations, shared by a community, on what counts as morally positive or negative. As pointed out by B. Spector, the contrast we found between "many children died" and "many children survived" should be linked with the one we can feel between:

(10) a. *only 5 children died (out of 10)

b. only 5 children survived (out of 10)

"Only 5 As are Bs" can only be used if the speaker expected more As to be B than actually happened. Here (10-a) is marked because it can only be uttered by someone who expected more children to die, against the default moral norm.[3] Geurts (2009) discusses similar contrasts based on the operator "it is good that", which can be adapted to the same example:

(11) a. *It is good that 5 children died (out of 10)

b. It is good that 5 children survived (out of 10)

Here again, (11-a) is the marked case, since uttering it implies that, had more than 5 children died, it would still have been good

[3] See already Sapir (1944) for observations about "only" and other adverbial expressions that emphasize the interplay between grading and "affect" in the interpretation of number sentences.

(see Sanford A. J. and Moxey (2007), Geurts (2009) and Nouwen (2011) for more on monotonicity constraints in relation to framing effects). Or consider the following pair, originally presented in Zuber (1983) to illustrate a different point:

(12) a. *Bill regrets that the glass is half-full
b. Bill regrets that the glass is half-empty

(12-a) is marked here since presumably, "X regrets that $P(y)$" implies that X would have liked y not to be at least as much P. Hence (12-a) implies that Bill would have liked the glass to be less than full, which goes against the default expectation in a context in which no specific information is given about what kind of liquid is in the glass.

In all these examples, although the position of the relevant standard on the scale for antonym pairs can be manipulated depending on the context, we observe a systematic asymmetry, suggesting that the marked case is always evaluated against a default affective or moral norm. This whole range of data, in our view, gives ample confirmation of the view articulated by Pettit and Knobe (2009) and Knobe (2010) about the Knobe effect, but it also suggests that the latter belongs to the same family of semantic asymmetries as we find instantiated in framing effects more generally, even though the case rests on further specifics about the semantics of the adjective "intentional".

References

Bartsch, R., Vennemann, T., 1972. The grammar of relative adjectives and comparison. Linguistische Berichte 2-, 19–32.

Bonnefon, J.-F., Villejoubert, G., 2006. Tactful or doubtful? expectations of politeness explain the severity bias in the interpretation of probability phrases. Psychological Science 17, 747–751.

Égré, P., 2010a. Intentional action and the semantics of gradable expressions (forthcoming). In: in B. Copley, Martin, F. (Eds.), Forces in Grammatical Structures (provisional title).

Égré, P., 2010b. Qualitative judgments, quantitative judgments and norm-sensitivity. (open peer commentary on knobe 2010). Behavioral and Brain Sciences 33 (4), 335–336.

Fara, D., 2000. Shifting sands: an interest-relative theory of vagueness. Philosophical Topics 28 (1).

Geurts, B., 2009. Goodness. In: M. Aloni, H. Bastiaanse, T. d. J. P. v. O., Schulz, K. (Eds.), Proceedings of the 17th Amsterdam Colloquium. pp. 277–285.

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica 47 (2), 263–292.

Keenan, E., Stavi, 1986. A semantic characterization of natural language determiners. Linguistics and Philosophy 9, 253–326.

Kennedy, C., 2007. Vagueness and grammar: the semantics of absolute and relative gradable adjectives. Linguistics and Philosophy 30, 1–45.

Knobe, J., 2003b. Intentional action and side-effects in ordinary language. Analysis 63, 190–193.

Knobe, J., 2010. Person as scientist, person as moralist. Behavioral and Brain Sciences 33, 315–365.

Lappin, S., 2000. An intensional parametric semantics for vague quantifiers. Linguistics and Philosophy 23, 599–620.

Mandelbaum, E., Ripley, D., 2010. Expectations and morality: a dilemma. (open peer commentary on knobe 2010). Behavioral and Brain Sciences 33 (4), 346.

Nouwen, R., 2011. Degree modifiers and monotonicity. In: Égré, P., Klinedinst, N. (Eds.), Vagueness and Language Use. Palgrave McMillan, pp. 146–164.

Partee, B., 1989. Many quantifiers. repr. in. In: Compositionality in Formal Semantics, Selected Papers by Barbara Partee. pp. 241–258.

Pettit, D., Knobe, J., 2009. The pervasive impact of moral judgments. Mind and Language 24, 568–604.

S. Pighin, J-F. Bonnefon, L. S., 2009. Overcoming number numbness in prenatal risk communication, university of Toulouse and

Department of Cognitive Science and Education, University of Trento.

Sanford A. J., D. E., Moxey, L., 2007. A unified account of quantifer perspective effects in discourse. Discourse Processes 44, 1–32.

Sapir, E., 1944. Grading, a study in semantics. Philosophy of Science 11 (2), 93–116.

Tversky, A., Kahneman, D., 1981. The framing of decisions and the psychology of choice. Science 211 (4481), 453–458.

Weber, E., Hilton, D., 1990. Contextual effects in the interpretation of probability words: perceived base rate and severity of events. Journal of Experimental Psychology: Human Perception and Performance 16 (4), 781–789.

Zuber, R., 1983. Semantic restrictions on certain complementizers. In: Proceedings of the 13th International Congress of Linguists. Tokyo.

Crespo, M.I., Gakis, D., Sassoon, G. W. (eds.),
Proceedings of the Amsterdam Graduate Philosophy Conference
— Truth, Meaning, and Normativity —
ILLC Publications X-2011-01, 19–30

Relative Truth, Lost Disagreement and Invariantism on Predicates of Personal Taste

Delia Belleri
(University of Bologna)

deliabelleri@gmail.com

One of the advertised advantages of Relativism on predicates of personal taste is that it manages to capture those elements of contradiction and faultlessness that characterise disagreement in the personal taste area of discourse. My aim in this paper is twofold: first, I wish to show that Relativism fails this task, in that it fails to capture any interesting notion of disagreement; second, I shall suggest that the only way to preserve a notion of "faultless disagreement" is to opt for an Invariantism on predicates of personal taste, accompanied by an epistemic notion of faultlessness.

I Taste Disputes and the "Ordinary View"

Suppose that Alice and Grace find themselves involved in a dispute over whether guacamole is tasty, and discover that they disagree on that matter. This is how their dialogue could look:

(1) a. Alice: "Guacamole is tasty"

b. Grace: "No, you're wrong! Guacamole is not tasty"

Here the use of expressions like "No" and "You're wrong" is a sign that the two speakers at least take themselves as disagreeing, and that they behave accordingly. All appearances suggest that what is going on between Alice and Grace is a genuine disagreement. But isn't such a disagreement special in some way? The feeling is that

neither Alice nor Grace is really wrong or at fault. Let us then say that, when it comes to matters of personal preference, the following "Ordinary View"[1] applies to disagreement:

The Ordinary View

(OV1) Contradiction: A asserts that P and B asserts that not-P;

(OV2) Faultlessness: Neither A nor B is at fault;

In this paper, the role of the Ordinary View shall be that of representing a piece of "common sense", though one to be taken into consideration by any theorist aiming to provide a semantic account of predicates of personal taste. It is therefore important to bear in mind that the Ordinary View is neutral as to which theory is the correct one for capturing the semantic profile of predicates like "tasty".

II Contextualism

One semantic account that seems at first sight to predict the Ordinary View is Contextualism. According to this view, the predicate *tasty* contains an implicit argument-place which takes as its value a standard of taste, s. Utterances of sentences of the form "X is tasty" really express the content that X is tasty by standard s. This seems to capture well the faultlessness element (OV2), for it may well be true that guacamole is tasty by Alice's standard, as well as that guacamole is not tasty by Grace's standard. However, if this is so, the two utterances express different contents, and the contradiction element (OV1) is lost: Alice and Grace are simply talking past each other (see MacFarlane (2007)).

Contextualism can predict that (1) is an example of contradiction only on the assumption that the value of the taste-standard parameter s does not vary in each utterer's mouth. So, as long as, e.g., Alice's assertion expresses the content that guacamole is tasty

[1] I here basically follow the line of Wright (2006): there the Ordinary View combines three elements rather than two: *Contradiction, Faultlessness and Sustainability*, where Sustainability consists in the parties' being rational in sticking to their respective views "even after the disagreement comes to light and impresses as intractable" (Wright, 2006, p. 38). For ease of exposition, I here choose to omit the element of Sustainability, though all I am going to say can be extended also to a version of the Ordinary View that includes Sustainability.

by the standard of community c, and Grace's assertion is the denial of exactly that content, a contradiction is obtained and point (OV1) is met. As a consequence, though, the faultlessness element of the dispute (OV2) gets neglected for, since it is either the case or not the case that guacamole is tasty by the standards of c, either Alice or Grace is in error about what the facts are.

III Realism

A possible alternative to Contextualism is Realism, of which one can distinguish two varieties: (i) what Wright (2001) calls Rampant Realism, according to which the word "tasty" denotes the objective property of tastiness, which is either possessed or not possessed by objects; and (ii) a Moderate Realism, which treats the property of tastiness as response-dependent (see Wright (2006)), in such a way that it counts as tasty only what is attributed this property by the majority of a group of designated judges or experts. Both Rampant and Moderate Realism see a genuine contradiction in disputes like (1), and hence preserve point (OV1); unfortunately, though, they both fail to capture the speakers' faultlessness — hence giving up (OV2) — for, if tastiness is to be an objective property, then either A or B must be mistaken.

IV Relativism

Another, more widely chosen option is Relativism. In general, what characterises a relativist position — in some domain of discourse D — is that "the relativist about a given domain, D, purports to have discovered that the truths of D involve an unexpected relation to a parameter" (Boghossian, 2006, p. 13).

Applied to the semantics of predicates like "tasty", the relativist claims that the truth of any utterance of the form "X is tasty" is relative to a taste-dedicated parameter. The truth-conditions of a sentence like "Guacamole is tasty" will thus be spelled out as follows:

(2) "Guacamole is tasty" is true in circumstances of evaluation $\langle w, s\rangle$ if, and only if, guacamole is tasty in $\langle w, s\rangle$;

where the truth of the sentence is relative to both a parameter w on possible worlds and a parameter s on standards of taste.

Though many authors have characterised Relativism in this or in equivalent manners (see Lasersohn (2005), Kölbel (2009)), such

characterisation is still under-determined, for there is still a way for (2) to express a form of Contextualism, namely in the case in which the value of the taste-parameter s is systematically fixed by the context of utterance. If this were the case, then it would be impossible for a speaker A to (i) retract on her previous assertion (or belief) as to the tastiness of guacamole, by saying that what she asserted (or believed) is *false*; and (ii) to disagree with another speaker B, by saying that what B asserts is *false*.

In view of such considerations, authors such as MacFarlane (2005), (2007) have claimed that the distinctive feature of Relativism is not just the relativity of truth to an extra-parameter (like s), but the fact that the value of such extra-parameter is fixed by a *context of assessment*, i.e. not a context in which a proposition is uttered, but one in which it is evaluated as true; for the Relativist, such context is independent of, and hence needs not be identified with, the context of use. The truth-conditions of "Guacamole is tasty" are thus to be formulated as follows:

(2') "Guacamole is tasty" is true in circumstances of evaluation $\langle w_u, s_a \rangle$ if, and only if, guacamole is tasty in $\langle w_u, s_a \rangle$;

where w_u is the world of the context of utterance and s_a is the taste-standard of the context of assessment. For ease of exposition, I shall simply call "Relativism" the approach from sensitivity to contexts of assessment just presented.[2]

One of the implications of Relativism is that cases of faultless disagreement like that in (1) are perfectly dealt with (for this point, see particularly Kölbel (2004), (2009)). First of all, assessment-sensitivity captures the faultlessness element (OV2): if truth is assessment-sensitive, then Alice and Grace qualify as faultless to the extent that they are evaluating the proposition that guacamole is tasty from two different contexts of assessment. Secondly, assessment-sensitivity allows the theorist to say that Alice and grace are using the same prejacent proposition (namely, "guacamole is tasty"), whose truth is relative to a world plus a taste-standard parameter s. Since Alice is affirming that proposition while Grace is denying it, the propositions they assert are contradictory: so (OV1) is satisfied, too.

[2] Though all I will say will apply also to the other versions, their under-determination suitably adjusted.

But is that really the case? Are Alice and Grace contradicting each other in any *interesting* way? Before moving on to what the problem for Relativism is with respect to disagreement, let us pause for a moment and consider what an interesting disagreement in general may amount to.

V Disagreement as Open-issue

A case of interesting disagreement may be described as an "open-issue" situation that the world is required to settle. Suppose that Alice and Grace disagree over whether a certain surface is red or white-but-red-looking because of the lightning conditions: Alice utters "That surface is red" and Grace utters "That surface is not red". The situation seems to be such that there is an open issue between them on what the facts in the world are. The motivation for Grace to disagree with Alice is that Grace believes that the world is not arranged the way Alice describes it. Conversely, if Alice does not change her mind and instead sticks to her guns, that will be because she thinks that the world is arranged the way she describes it. Both of them will contend that the facts in the world support their respective assertions, but finally only one of them can be right.

Generalising from this example — on the assumption that what usually interests speakers in a dispute is the fact of the matter in the relevant area of discourse — what makes a subject A disagree with another subject B is plausibly the idea that the world is not as A describes it, but it is rather as B claims it to be. What the fact of the matter is, is the "open issue"; he world should be the basis for settling the issue.

My contention here is that a fact of the matter (or at least its possibility) is necessary for there to be a disagreement. If it were discovered that no fact of the matter could (even possibly) settle a dispute in a certain area of discourse, then any basis for an interesting disagreement would disappear — even though the activity of disputing could still have a point for other reasons. Suppose it were discovered that no moral fact of the matter is ever possible. Then Alice and Grace could still engage in the activity of disputing over whether it is morally right to e.g. eat meat: Alice could utter "Eating meat is right" while Grace could utter "Eating meat is not right"; however, since this activity could not possibly be aimed at establishing what the fact of the matter is, it would not be an

interesting disagreement, even though it could be something else: for example, a harmless exchange of views between the speakers, in which each expresses what she believes; alternatively, it could be an attempt from each speaker to persuade her interlocutor, just for the sake of prevailing against one another.

Again generalising from the example, I shall say that for there to be an interesting disagreement, there has to be a fact of the matter in the world, or at least a fact of the matter has to be possible. The role of the world is that of (potentially) settling the dispute, i.e. that of deciding who among the parties in the disagreement is telling the truth. If this condition is not satisfied, then there shall be only a dispute involving utterance of contradictory contents, that though doesn't amount to any interesting, substantive disagreement.

VI Relativism fails to capture disagreement

Suppose now that the following Relativist assumptions are true:

1. The truth-value of a proposition P is settled in a (relevant) context of assessment;

2. It is reasonable to evaluate the same proposition P from different contexts of assessment;

3. No context of assessment is privileged with respect to another;

Suppose also that:

4. P is evaluated as true in A's context of assessment;

5. P is evaluated as false in B's context of assessment;

[6] below is consistent with [4]-[5] and [2]:

6. A's and B's assessments are both reasonable;

Plus, from [4]-[5] and [3] it follows that:

7. Neither A's nor B's assessment is privileged with respect to the other.

As one can see, to the extent that the truth of each utterance is fixed in a (relevant) context of assessment, A's and B's assessments

are fully compatible. This means that nothing can settle the issue between the two parties the way a fact of the matter in the world would settle the issue in an "open-issue" situation. So, if contexts of assessment settle the truth of taste-assertions, there is no room for the world's settling the dispute. However, that the world settles the dispute is necessary for there being an interesting, substantive disagreement. Therefore, there is no interesting disagreement between A and B; more generally, there is no interesting disagreement in a Relativist framework.

Suppose, now, we drop Relativism altogether and start viewing a predicate like "tasty" as an un-relativised predicate altogether. The truth or falsity of utterances of "X is tasty" becomes absolute. If the truth of taste-assertions is conceived of as absolute, then a sense of disagreement as "open-issue" can be regained. But why would one want to see "tasty" as an un-relativised predicate?

VII Going Invariantist on "Tasty"

Endorsing a picture in which "tasty" is an un-relativised predicate means defending an Invariantist view of the semantics of this expression, according to which: (i) the word "tasty" corresponds to the one-place predicate *tasty* and it denotes the monadic property of tastiness; (ii) utterances of sentences of the form "X is tasty" express the un-relativised proposition *that X is tasty tout court* and are true if, and only if, X is tasty *tout court*. What evidence can be exhibited in favour of this view?

A first piece of evidence comes from comparison of "tasty" with other expressions, whose semantics may be plausibly cashed out in relativistic terms. If "tasty" had a relativistic semantics just like these expressions, some utterances that seem perfectly OK would be predicted as false. Let's first consider a predicate like "local". The truth conditions of "X is local" may be represented in a relativistic fashion, as in:

(3) "X is local" is true in a circumstance of evaluation $\langle w, l \rangle$ iff X is local in $\langle w, l \rangle$;

Suppose Alice is in Paris and Grace in London; Alice utters:

(4) Grace and I went to a local bar together.

It's not difficult to see that, in the situation envisaged, (4) is false, because the locational parameter associated with the occurrence of "local" can take just one value, while the referent of "I" and the referent of "Grace" are in two different locations. If "tasty" were like "local", then supposing Alice and Grace have two sufficiently different taste-standards s_1 and s_2, the following would be false, too:

(5) Grace and I believe that guacamole is tasty.

However, utterances of (5) seem perfectly fine, for it seems perfectly fine to conceive both Grace and Alice as attributing to guacamole the simple, un-relativised property of being tasty.

A second source of evidence for Invariantism on "tasty" lies in the fact that accounting for agreement on taste does not require resorting to any relativisation of the predicate. Consider the following exchange:

(6) a. Grace: "Guacamole is tasty";
b. Alice: "I agree. Guacamole is really tasty";

Here it seems that there's no need to say that Alice and Grace both believe that guacamole is tasty with respect to a standard s. No relativisation of the predicate need be invoked to account for agreement, for it seems OK to see Alice and Grace as just attributing to guacamole the property of being tasty *tout court*.

Thirdly, a supporter of Invariantism on "tasty" might urge that, if we really disagree on taste, then "tasty" must be an un-relativised predicate. Why so? Because it seems that sensibly disagreeing with someone's claim to the effect that X is F means being ready to engage in the task of establishing what the fact of the matter about X is in the world. Suppose Alice and Grace were disagreeing about Peg's age. The following would be a correct way for them to conduct their disagreement:

(7) a. Alice: "Peg is thirty-five years old";
b. Grace: "No she's not. I saw her ID once and the birth-date was 1965";

Otherwise, apparent disagreement may simply be the expression of each party's personal opinion, as if Grace, rather than uttering (7b), uttered (7b'):

(7) b'. Grace: "No she's not. At least so I *believe*".

Or again, the apparent disagreement may instead be just an attempt to persuade the opponent, as if Grace responded to Alice by uttering (7b"):

(7) b". Grace: "No she's not. You just have to listen to what *I* say".

If the previous evidence is sound, and if there are strong enough reasons to believe that Invariantism about "tasty" is true, then a distinction is in order, between the *truth-conditions* of an utterance of "X is tasty" and the *reasons* a subject has to assert (or believe) that X is tasty. What is the import of such a distinction? Let me illustrate with an example.

Suppose Alice has a certain gustatory experience E as of the tastiness of guacamole, which Alice takes as evidence in favour of the proposition that guacamole is tasty. This (putative) evidence E is a reason for Alice to assert/ believe that guacamole is tasty. However, in the Invariantist framework, it is not the case that, if Alice ends up asserting/believing that guacamole is tasty on the basis of some experience E as of its tastiness, then the proposition that guacamole is tasty is true "relative to Alice's experience". If Invariantism is true, then the proposition that guacamole is tasty is true iff guacamole is tasty, period. This implies that, no matter what Alice's experience E is, the proposition she comes to believe on the basis of E is true or false independently of E.

VIII Rescuing Faultless Disagreement

Let us now return to Alice's disagreement with Grace in (1). In light of our Invariantist approach, we might then say that Alice and Grace are respectively affirming and denying the same proposition that guacamole is tasty *tout court*. Since it is either the case or not the case that guacamole is tasty *tout court*, Alice's and Grace's utterances are really contradicting each other, because the truth of the former implies the falsity of the latter, and vice-versa. One can therefore see how the contradiction element (OV1) is captured by the Invariantist approach.

But what of the faultlessness element, (OV2)? If guacamole is

either tasty or not tasty, then either Alice or Grace is making a mistake. The faultlessness element seems to be gone in the Invarantist framework. Is Invariantism then as defective as the views surveyed at the outset? Not quite so. Recall the distinction made previously between the reasons for asserting (or believing) that X is F and the truth-conditions of "X is F": In this picture, both Alice and Grace appear as faultless in their taking their own experience as of the tastiness (or non-tastiness) of guacamole as evidence and, therefore, as a reason to assert (or believe) that guacamole is tasty or not tasty. The reason they are faultless is that they have no other way to ascertain (in the first person) the tastiness of guacamole but to taste it. The absence of fault here relates to their being *epistemically impeccable* with respect to the proposition that guacamole is tasty. The faultlessness element (OV2) is then restored by Invariantism on "tasty", though with the important qualification that absence of error in taste-disputes boils down to absence of *epistemic* fault.

The Ordinary View is therefore predicted by Invariantism on taste-predicates, though with an important distinction: (i) that contradiction pertains to the semantics of taste-expressions (and relates to the metaphysics of the properties denoted); (ii) while faultlessness pertains to the epistemology of taste.

Before closing, two notes are in order. First, admittedly, the realist claim I have endorsed to the effect that the world decides whether X is tasty or not sounds bad. As Egan observes, when matters of taste are at issue, "the idea that there is any crucial evidence to be found, that there are any objective facts in this domain to be discovered, seems deeply suspect." (Egan, 2010, p. 14).

As understandable as these worries are, let me try to provide some reassuring considerations. Accepting Realism on the metaphysics of taste properties doesn't entail making any predictions on speakers' behaviour such that they regard speakers as likely to behave differently from how they are *actually* likely to behave. For, even if guacamole is objectively tasty (or not tasty), one can predict that Grace and Alice are likely to disagree over its tastiness, on the account that they are likely to have different experiences of the way guacamole tastes. Despite its suspicious appearance, then, Realism doesn't ultimately have any suspicious consequences, since the predictions yielded by Realism track actual patterns of speakers' behaviour.

Secondly, it could be pointed out that my approach entails that, even though a certain item X is objectively either tasty or not tasty, no subject could possibly come to know that, because no subject is in a position to know whether her gustatory experience is a reliable guide to the tastiness (or non-tastiness) of X. I am not sure that this claim is true: let us suppose that knowledge is justified true belief and that, moreover, whether S's beliefs about the tastiness of items qualify as knowledge is not accessible to S's awareness. If guacamole is objectively tasty, this means that if S believes, on the basis of her gustatory experience, that guacamole is tasty, then she also *knows* that; only, S doesn't know that she knows that. If this is the case, then my approach is compatible with the claim that subjects can have knowledge about the tastiness of items, even though this knowledge is not *transparent* to them.[3] Further issues that would be interesting to investigate are whether this lack of transparency is necessary or merely contingent and, in case it is contingent, whether it is remediable or not. However, I have to reserve consideration of these matters for another paper.

Acknowledgments

I would like to thank, for their helpful feedback, the participants to the Amsterdam Graduate Philosophy Conference, especially Max Kölbel, Floris Roelofsen, Martin Stokhof, Åsa Wikforss and Lucian Zagan. Many thanks also to Annalisa Coliva, Paolo Leonardi, Sebastiano Moruzzi, Marco Santambrogio, and to all the members of the COGITO research group, for discussion on the paper and on the topic in general.

References

Boghossian, P., 2006. What is relativism? In: Greenough, P., Lynch, M. (Eds.), Truth and Realism. Oxford University Press.

Egan, A., 2010. Disputing about taste. In: Feldman, R., Warfield, T. (Eds.), Disagreement. Oxford University Press.

Kölbel, M., 2004. Faultless disagreement. In: Proceedings of the Aristotelian Society. Vol. 104. pp. 53–73.

[3] Thanks to Annalisa Coliva for discussion and helpful suggestions on this point.

Kölbel, M., 2009. The evidence for relativism. Synthese 166, 375–395.

Lasersohn, P., 2005. Context dependence, disagreement, and predicates of personal taste. Linguistics and Philosophy 28 (4), 643–686.

MacFarlane, J., 2005. Making sense of relative truth. In: Proceedings of the Aristotelian Society. Vol. 105. pp. 321–339.

MacFarlane, J., 2007. Relativism and disagreement. Philosophical Studies 132, 17–31.

Wright, C., 2001. On being a quandary. Mind 110 (437), 45–98.

Wright, C., 2006. Intuitionism, realism, relativism and rhubarb. In: Greenough, P., Lynch, M. (Eds.), Truth and Realism. Oxford University Press.

Crespo, M.I., Gakis, D., Sassoon, G. W. (eds.),
Proceedings of the Amsterdam Graduate Philosophy Conference
— Truth, Meaning, and Normativity —
ILLC Publications X-2011-01, 31–39

Does Tarski's critique of the Redundancy Theory apply to all Deflationary Theories of Truth?

Monika Gruber
(Universität Salzburg)

monika.gruber3@gmail.com

The concept of truth has always been one of the most important and widely debated topics in the history of philosophy. One of the most popular approaches to truth in the twentieth century is presented by the deflationsts. Their theories originate from Ramsey's revolutionary statement made in *Facts and Propositions* (1927). There, he holds first of all that truth and falsity are primarily ascribed to propositions. Furthermore, he holds that the proposition that 'Caesar was murdered is true' means the same as 'Caesar was murdered'.[1] This statement provided an inspiration for all deflationary theories of truth.

However, the real turning point in the development of the theories of truth was made by Tarski.[2] In 1933 he presented an impeccable definition of truth which gave truth a central role in philosophical thought. All further theories of truth use as basis Tarski's equivalence schema. Ironically, the deflationary theories of truth which deny that truth is a substantial property also use Tarski's equivalence schema as basis.

[1] Cf. Ramsey (1994), pp. 34-39

[2] As basis for this article I use the original polish version of Tarski's paper: *Pojęcie prawdy w językach nauk dedukcyjnych* (1933), edited by J. Zygmunt in (1995). The only exception is made for direct quotations of the English translation from (1956) titled *The Concept of Truth in Formalized Languages* for which I use the newest edition (2006). For detailed information see the references.

After stating the impossibility of constructing a materially adequate and formally correct definition of the notion of truth within the colloquial languages, Tarski proves his attempts successful on the grounds of a formalized language. He emphasizes the importance of distinguishing between the language about which we speak (the object-language) and the language in which we speak (the metalanguage), as well as between the science which is the object of our investigation and the science in which the examination is performed. He delivers the necessary terms, axioms and definitions and arrives at the famous Convention T, where the symbol 'Tr' denotes the class of all true sentences.

> CONVENTION **T**. A formally correct definition of the symbol 'Tr', formulated in the metalanguage, will be called an *adequate definition of truth* if it has the following consequences:
>
> (α) all sentences which are obtained from the expression '$x \in Tr$ if and only if p' by substituting for the symbol 'x' a structural-descriptive name of any sentence of the language in question and for the symbol 'p' the expression which forms the translation of this sentence into the metalanguage ;
>
> (β) the sentence 'for any x, if $x \in Tr$ then $x \in$ S' (in other words '$Tr \subseteq$ S').' (Tarski, 2006, pp. 187-188)

From that we arrive at the famous equivalence scheme:

(T) X is true, if and only if, p.

It has ever since been used by philosophers and logicians in order to formulate their theories of truth, including the deflationary theories of truth. The deflationists hold that all that can be meaningfully said about truth can be said by the means of the equivalence schema. Tarski believes the matter to be much more complicated. If the investigated language contained a finite number of sentences, and if we could enumerate all these sentences, then the construction of a correct definition of truth would not be a problem. However, since this is not the case, since languages contain infinitely many sentences, the definition constructed according to the above schema would also have to consist of infinitely many words. Such sentences cannot be formulated either in the metalanguage or in any other language. Hence, Tarski introduces the notion of *satisfaction of a given sentential function by given objects*, in this case by a given class of

individuals. The way Tarski explains the notion of satisfaction reflects the natural generalization of the method used for the concept of truth. The intuitively simplest case is that in which the given sentential function contains only one free variable. We can then significantly say of every single object that it either does or does not satisfy the given function. In the following scheme:

> For all a - a satisfies the sentential function x if and only if p,

we replace the free variable by *'a'*, and then we substitute for *'p'* the given sentential function and for *'x'* some individual name of this function. The situation is more complicated when the given sentential function contains an arbitrary number of free variables. In this case it has to be said that:

> A given infinite sequence of objects satisfies a given sentential function.

Tarski defines the notion of satisfaction in Definition 22 and emphasizes its importance for the construction of the definition of a true sentence, which he presents in Definition 23:

> Def. 23: x is a *true sentence*— in symbols $x \in Tr$— if and only if $x \in$ S and every infinite sequence of classes satisfies x.

Tarski's conception of truth consists in regarding the sentence *'X is true'* as equivalent to the sentence denoted by 'p'. Therefore, the term 'true', whether occurring in a simple sentence or in a complex one as a part of the expression *'X is true'*, can be removed, and the sentence of the metalanguage can be substituted by an equivalent sentence in the object language. However, the term 'true' cannot be eliminated in all cases. While discussing the redundancy of semantic terms, and their possible elimination, Tarski names two instances which require the use of the predicate 'is true'. In a simple sentence, where the name of the sentence which is said to be true is not in a form enabling us to reconstruct the sentence itself, the discussed elimination is impossible, e.g. *The first sentence written by Plato is true.* Perhaps, the most important example of the sentences where the predicate 'is true' cannot be eliminated in the simple manner

contemplated is that of universal statements, e.g., *All consequences of true sentences are true.*[3]

Furthermore, Tarski makes an important remark, which has often gone unnoticed.

> It should be emphasized that neither the expression (T) itself (which is not a sentence, but only a schema of a sentence) nor any particular instance of the form (T) can be regarded as a definition of truth. We can only say that every equivalence of the form (T) obtained by replacing 'p' by a particular sentence, and 'X' by a name of this sentence, may be considered a partial definition of truth, which explains wherein the truth of this one individual sentence consists. (Tarski, 1986, p.668)

Deflationary theories differ in their approaches to truth, however they all claim in unison that truth has no underlying nature and, therefore, plays no substantial role in philosophical thought. Besides any particular weaknesses each deflationary theory might have, there is one major problem which concerns every deflationary theory of truth, and which neither of them has been able to overcome. It is the generalization problem.

We can generalize such sentences as 'Paul is mortal', 'Hartry is mortal', etc., without applying the truth predicate to the sentences, and thus say 'All men are mortal'. Similarly we can generalize on 'Paul is Paul', 'Hartry is Hartry' and say 'Everything is itself' omitting the truth predicate. However, when we want to generalize 'Hartry is mortal or Hartry is not mortal', or 'Snow is white or snow is not white' we have to use the truth predicate and talk about sentences: 'Every sentence of the form *'p* or not *p'* is true'. We are able to generalize over the sentences given in the first two examples because the changes that occur are simply changes in names. Therefore, we can read this generalization 'x is mortal for all *men* x' – all things x of the sort that 'Paul' is a name of. However, if we want to generalize 'Hartry is mortal or Hartry is not mortal' without using the truth predicate, we come up with 'p or not p for all things p of the sort that sentences are names of'. The problem is that sentences are not names and this reading is incoherent because the variable *'p'* is pronominal and occupies name positions, thus, it cannot meaningfully be put in sentence positions.[4]

[3] Cf. Tarski (1986), pp. 665-698.
[4] Cf. Quine (1986), pp. 11-12.

Deflationists misinterpreted Tarski's schema and obviously also Quine's interpretation of it.

> By calling the sentence ['snow is white'] true, we call snow white. The truth predicate is a device of disquotation. We may affirm the single sentence by just uttering it, unaided by quotation or by the truth predicate; but if we want to affirm some infinite lot of sentences that we can demarcate only by talking about the sentences, then the truth predicate has its use. We need it to restore the effect of objective reference when for the sake of some generalization we have resorted to semantic ascent." (Quine, 1986, p. 12)

What is meant by 'wanting to affirm some infinite lot of sentences that we can demarcate only by talking about the sentences' is that we want to affirm a generalization. The deflationists confuse generalization with the infinitary conjunction, which allows them to formulate their deflationary thesis. Gupta points out that the deflationists make some very strong claims about the meaning of 'true', which when closely examined prove very problematic. Their accounts appear plausible only when they are read in a weaker way. However, the weaker readings do not yield their deflationary conclusions. According to the deflationary account, the function of the truth predicate is to express certain infinite conjunctions and disjunctions. The truth predicate serves these functions in virtue of its disquotational character, by undoing the effect of quotation marks.[5] For example in:

(1) 'snow is white' is true

the truth predicate cancels the quotation marks allowing us to arrive at the sentence:

snow is white

which yield the same sense. If we want to 'affirm some infinite lot of sentences' as Quine puts it, we wish to affirm all sentences of the form:

_____ & snow is white [= A]

This means that we want to affirm the conjunctions of all sentences obtained by filling the blank in A with sentences of English:

[5] For the following argument cf. Gupta (2005a), pp. 203-205.

(2) [Sky is blue & snow is white] & [Chicago is blue & snow is white] & ...

As Gupta notes, we lack explicit and direct means of formulating the infinite conjunction, since it is infinite and we will never be able to fill in all the possible combinations of sentences. Nevertheless, according to Quine and the deflationists, the truth predicate provides us with an indirect means. However, we cannot generalize A by saying

For all x: x & snow is white

since the variable 'x' is pronominal and can only represent names, not sentences. According to the disquotational account, the disquotational feature of truth makes (2) equivalent to:

(3) ['Sky is blue' is true & snow is white] & ['Chicago is blue' is true & snow is white] & ...

But, the position '_____' in

_____ is true & snow is white

is nominal and can be quantified using the pronominal variable 'x'. Therefore, we can say

(4) For all sentences x: [x is true & snow is white]

Since (4) is equivalent to (3) it is also equivalent, in virtue of disquotation, to (2). The truth predicate enables us to express the infinite conjunction (2). On the disquotational account, truth is a logical device, enabling us to generalize over sentence positions while using pronominal variables, and thus delivers the additional expressive power. The deflationists hold that the equivalence schema explains the meaning of 'true' and that it issues from our understanding of it.

Deflationists claim that the truth predicate is a device for expressing certain infinite conjunctions and disjunctions. However, they do not specify their usage of the ambiguous term 'express'. It is not clear if their thesis means that (4) and (2) are materially equivalent, necessarily equivalent or if they have the same sense. But, the

way they use their Infinite Conjunction Thesis requires that 'express' be read in a strong way.

The function of (4) is to express (2). This however, is only possible if (2) and (3) are equivalent. Therefore, deflationists hold that (2) and (3) *need* to be equivalent, since only then the truth predicate can play its expressive role. However, the equivalence of (2) and (3) has to be understood as the sameness of sense, any weaker reading will not yield the disquotational thesis. And this is where deflationism fails. Universal statements like (4) do not have the same sense as the infinite conjunction of its instances (3). As Gupta points out, they do not even imply the same things, they are equivalent only in a much weaker sense. Its proponents have ignored the difference between affirming the generalization and affirming each of its instances. Perhaps the most important reason why generalizations involving 'true' are so useful is that they do not mean the same as their instances analysed separately. [6] Generalizations are logically stronger than the conjunctions of their instances because they imply these conjunctions, whereas the conjunctions do no imply the generalizations.

In the *Postscript* to his *Truth*, Paul Horwich replays to Anil Gupta's critique. He weakens his minimal theory and admits that a theory claiming that 'p' and 'The statement (belief, ...) *that* p is true' is implausibly strong. Instead, he holds that the function of truth "requires merely that the generalizations permit us to *derive* the statements to be generalized–which requires merely that the truth schemata provide material equivalences. This isn't to deny that the instances so understood are not only true but necessarily true (and a priori). The point is that their mere truth is enough to account for the generalizing function of truth." (Horwich, 1998, p. 124). (Italics added by the author.) Furthermore, he claims that "there is a truth-preserving rule of inference that will take us from a set of premises attributing to each proposition some property, F, to the conclusion that all propositions have F."(Horwich, 1998, p. 137) Therefore, from 'x is F' we come to the conclusion that 'All propositions are F'. Horwich's truth-preserving rule says:

(R) That p_1 is F, that p_2 is F, that p_3 is F, ...; therefore, all propositions are F.

[6] Cf. Gupta (2005a), p. 207.

However, as Gupta points out in his replay, the premisses of this rule form an infinite totality. For each proposition p, the totality contains the premiss that p is F. From this infinite totality the rule allows us to derive the conclusion that all propositions are F.[7] It is clear that any theory of truth which does not have to resort to an additional rule in order to explain the generalizations about truth is a better and a simpler theory than Horwich's minimal theory. And that is precisely what Tarski's theory of truth does. It provides explanations of a range of generalizations about truth without invoking any infinitary rule.

Therefore, Tarski's critique regarding the redundancy theory applies to all deflationary theories of truth.

References

Gupta, A., 2005a. A critique of deflationism. In: Armour-Garb, B. P., Beall, J. (Eds.), Deflationary Truth. Open Court, pp. 199–226.

Gupta, A., 2005b. Postscript to 'a critique of deflationism'. In: Armour-Garb, B. P., Beall, J. (Eds.), Deflationary Truth. Open Court, pp. 227–233.

Horwich, P., 1998. Truth, 2nd Edition. Clarendon Press.

Quine, W. V., 1986. Philosophy of Logic, 2nd Edition. Harvard University Press.

Ramsey, F. P., 1994. Facts and propositions (1927). In: Mellor, D. H. (Ed.), Philosophical Papers. Cambridge University Press.

Tarski, A., 1986. The semantic conception of truth and the foundations of semantics (1944). In: Givant, S. R., McKenzie, R. N. (Eds.), Alfred Tarski. Collected Papers. Vol. 2. 1935-1944. Birkhäuser, Basel, Boston, Stuttgart, pp. 665–699.

Tarski, A., 1995. Pojęcie prawdy w językach nauk dedukcyjnych (1933). In: Zygmunt, J. (Ed.), Pisma Logiczno-Filozoficzne. Vol. 1. Prawda. Wydawnictwo Naukowe PWN, Warszawa, pp. 13–172.

[7] Cf. Gupta (2005b) pp. 228-231.

Tarski, A., 2006. The concept of truth in formalized languages. In: Corcoran, J. (Ed.), Logic, Semantics, Metamathematics. Papers from 1923 to 1938, 2nd Edition. Hackett Publishing Company, pp. 152–278.

Crespo, M.I., Gakis, D., Sassoon, G. W. (eds.),
Proceedings of the Amsterdam Graduate Philosophy Conference
— Truth, Meaning, and Normativity —
ILLC Publications X-2011-01, 40–49

Giving Warrant Credit in Warranted Assertibility: Against Wright's Inflationary Argument

Rhys McKinnon
(University of Waterloo)

rhys.mckinnon@gmail.com

Crispin Wright has famously argued that a deflationary theory of truth cannot account for truth's role in norms of warranted assertibility. Truth and warranted assertibility are normatively coincident but extensionally divergent and the only explanation of this is some property of truth. Thus, since truth has a property above what can be accounted for by the disquotational schema, truth amounts to something more substantial than the deflationist can allow. Thus, deflationism fails. Or, so goes the argument. In this paper I will argue that Wright mistakenly attempts to explain the normative coincidence but extensional divergence of truth and warranted assertibility as a property of truth. Instead, I will argue that it is a fundamental property of warrant that explains this difference. Consequently, the deflationist position can adequately account for truth's role in warranted assertibility and Wright's inflationary argument fails.

I Introduction

Deflationism about truth is, roughly speaking, the claim that the Disquotational Schema (DS) — that "p" is true iff p — sufficiently captures everything that needs to be captured about truth. Thus, to assert that a proposition is true is merely to repeat the assertion

of the proposition. "'Today is Wednesday' is true" is equivalent in meaning to "Today is Wednesday." Furthermore, it is the denial that truth amounts to anything more than the DS in so far as truth is not "analyzable" beyond the DS. For example, deflationists sometimes claim that many of the considerations that motivate theories of truth such as correspondence are claimed to not be the "job" of truth. How true statements may "connect" with the world or be "caused by" features of the world is not something for which a theory of truth must account. Instead, it is suggested, this may be the purview of theories of language or metaphysics but not of "truth" *per se*.

Crispin Wright has famously argued against deflationism in an attempt to demonstrate that a deflationary theory of truth is not adequate for some roles of truth. Specifically, he argues that the DS cannot adequately account for truth's role in norms of assertion. The role of truth in topics such as norms of assertion is a contemporary debate that spans metaphysics, philosophy of language, and epistemology. We may define assertion as a speech act whereby an agent expresses a commitment to the truth of a proposition (often characterized as a belief). Wright's strategy is to argue that the deflationist is committed to truth playing a particular role in norms of assertion and that since deflationary truth cannot adequately fill this role, we should reject deflationism as a theory of truth.

In this paper I will argue that Wright's "inflationary" argument against deflationism is ultimately unsuccessful. I will argue that Wright does not demonstrate that the deflationist is committed to the role of truth in norms of warranted assertion that Wright claims they are. In particular, I focus on the claim that truth and warranted assertibility are normatively coincident but extensionally divergent. Wright argues that the only explanation for this, by the deflationist's own light, is that truth is somehow normative and thus a real property.[1] I will argue that it is not a property of truth that explains this difference, but rather a defining property of *warrant*. Thus, while Wright's argument may be successful against some particular deflationists, it does not cut against all forms of deflationism. However, this paper is not a defence of deflationism; rather, it is merely a defence of deflationism against a particular attack.

[1] I wish to avoid discussion of what it means for truth to be, or have, a "property" as it is a beyond the scope of this paper.

II Wright's Inflationary Argument

Wright argues that truth plays a central role in norms of warranted assertion and that a deflationary theory of truth cannot adequately account for this role. Moreover, he argues that the deflationist is committed to truth having such a role in norms of assertion since this role follows from the DS. The implication being that such a role is more "substantial" than what a deflationary theory of truth can allow and remain deflationary. Substantial is meant in the sense that truth has a property (*viz.* normativity) beyond that for which the deflationist can account. Specifically, Wright argues that truth is both a prescriptive and descriptive norm of assertion. He defines "a *predicate*, F, is (positively) descriptively normative just in case participants' selection, endorsement and so on of a move is as a matter of fact guided by whether or not they judge that move is F. [...] Likewise a predicate is prescriptively normative just in case the selection, or endorsement, of a move *ought* to be so guided within the practice concerned." (Wright, 1992, pg. 16, emphasis in the original.) He then argues that "deflationism is committed to the thesis that the T-predicate is positively normative, both descriptively and prescriptively, *of any assertoric practice*." (Wright 1992, my emphasis.)[2]

Wright thinks that the deflationist is committed to this position since such a position follows from the DS: "p" is true iff p implies that "any reason to think that a sentence is T may be transferred, across the biconditional, into reason to make or allow the assertoric move to assert that p". (Wright, 1992, p.18) Thus, we may represent this norm as follows.

WA1 Any reason to believe that p is true is a reason to assert that p.[3]

For the sake of argument, I will accept this as at least *prima facie* plausible.[4] The reason for this is that it is not the focus of my criticism.

[2] The T-predicate referring to the truth predicate.

[3] Strictly speaking the text involves "think" rather than "believe." There may be an issue with my choice to use "believe", but I do not think that it makes a significant difference here.

[4] However, I do not think that the DS implies that a deflationist is committed to the position that any reason to believe that p is true is also a reason to assert that p. Elsewhere, I take a position where the norms of epistemic belief and

Due to the relationship between warranted assertion and truth in WA1 Wright argues that the norms of warranted assertion and truth are "normatively coincident." By this he means that if one side of the biconditional has normative force, then so does the other (at least in a defeasible sense). However, he argues that "although coincident in normative force in the senses indicated, 'T' and 'is warrantedly assertible' *have* to be regarded as registering distinct norms — distinct in the precise sense that although aiming at one is, necessarily, aiming at the other, success in the one aim need not be success in the other."(Wright, 1992, p.19) That is, truth and warranted assertibility are normatively coincident but extensionally divergent. So, although Wright recognizes that "p is true" and "p is warrantedly assertible" are necessarily coincident norms of assertion in that any reason to believe that p is true (epistemic justification) is a reason to think that p is warrantedly assertible (warranted assertibility), which follows from the DS, he also recognizes that these norms may have divergent extensions. That is, "while 'is T' and 'is warrantedly assertible' are normatively coincident, satisfaction of the one norm need not entail satisfaction of the other."(Wright, 1992, p. 21)

Wright thinks that it is clear that the norms must be extensionally divergent since it isn't the case that if a proposition, p, is true then it is necessarily warrantedly assertible. An obvious example would be a lucky guess.[5] If Mike were to form the belief that it is raining in London without any evidence concerning the weather in London, then although it may actually be raining in London (his belief is true) Mike does not have warrant to assert it. The key to Wright's inflationary argument is that what explains the normative coincidence but extensional divergence of "is warrantedly assertible" and "is true" is some property of truth. Thus, since there is some

warranted assertion may come apart such that any reason to believe that p is true may not also be a reason to assert that p. Wright suggests that we should only interpret this as a "defeasible" reason to assert that p, but this does not make much difference. It is plausible to believe that there could be a reason to believe that p is true whereby such a reason is sufficient for epistemic justification but falls short of warranted assertibility because such a reason does not sufficiently meet norms of assertion which are not present in norms of justified belief.

[5] While this is not Wright's strategy for establishing the extensional divergence of truth and warranted assertibility, I use it because it is a simple and accessible alternative. I do not think that any significant presuppositions are introduced in making this substitution.

property of truth that must explain this difference, and the deflationist is committed to truth not having any properties other than that characterized by the DS, then the deflationist cannot account for this property of truth. Wright attempts to argue that the deflationist is committed to truth having this property and that, therefore, their position is inconsistent.

III Analysis and Criticism of Wright's Argument

Crucial to Wright's inflationary argument is the specific role of truth in norms of assertion. The deflationist, Wright argues, is committed to truth having a central role in warranted assertibility which follows from the DS. That is, any reason to think that p is true is a reason to assert that p. This is WA1. However, I will argue that one may grant Wright's WA1, but that it is not some property of truth that explains the normative coincidence and extensional divergence of truth and warranted assertibility; instead, it is a defining property of *warrant*. Thus, the deflationist may avoid the inflationary argument because there's nothing demanding explanation *qua* truth.

What is unclear in Wright's argument is exactly why he thinks that truth has the special role in norms of assertion that he thinks it does. It is clear that he thinks that the deflationist is committed to truth and warranted assertibility being normatively coincident because of the DS: any reason to believe that p is true can be transferred across the DS biconditional ("p" is true iff p) as a reason to assert that p. However, the question remains of what exactly is the appropriate norm of warranted assertion and how exactly truth plays its role. For example, at one point Wright suggests that truth is a *constitutive* norm of assertion, but I will argue that this interpretation would be problematic given Wright's other comments.

I suggest that Wright's use of "constitutive" is sufficiently close to Williamson's whereby "if it is a constitutive rule that one must φ, then it is necessary that one must φ. More precisely, a rule will count as constitutive of an act only if it is essential to that act: necessarily, the rule governs every performance of the act."[6] Thus, if truth is

[6] Williamson (2000), p. 239. This is because Williamson's definition that "a rule will count as constitutive of an act only if it is essential to that act: necessarily, the rule governs every performance of the act" appears sufficiently consonant with Wright's claim that "the T-predicate is positively normative, both descrip-

a constitutive norm of assertion, then every instance of warranted assertion must be true (*viz.* warranted assertion must be "factive"). This appears to be a charitable interpretation of Wright's description of truth as a norm of assertion, since Wright argues that truth is essential to the act of warranted assertion. Thus, it is reasonable to interpret this to be the claim that truth is a constitutive norm of assertion.

Although Wright appears to suggest that truth is a constitutive norm of assertion, he gives us reason to think that truth can't be a constitutive norm of assertion. Specifically, consider his claim that "while 'is T' and 'is warrantedly assertible' are normatively coincident, satisfaction of the one norm need not entail satisfaction of the other." (Wright, 1992, p. 21) From this, it is not the case that the norm of assertion would look like "If p is warrantedly assertible, then p is true" since success in one norm (warranted assertion) need not entail success in the other norm (truth). However, if truth is a constitutive norm of assertion, then truth is a necessary condition for warranted assertion, since it governs the selection and making of moves in assertoric practices. Since truth "is positively normative, both descriptively and prescriptively, of any assertoric practice" it follows that if p is warrantedly assertible, then p must be true. However, we have just seen how Wright explicitly denies that truth is a constitutive norm of assertion. Thus, either Wright contradicts himself or we should not interpret his arguments as being towards truth as the constitutive norm of assertion. I suggest the latter.

Given this tension, perhaps it would be objected that we should not interpret Wright as arguing for truth as a constitutive norm of assertion; instead, a better interpretation could be that Wright may merely mean by the normative coincidence of truth and warranted assertion that "aiming at one is, necessarily, aiming at the other."

---

tively and prescriptively, of any assertoric practice" (Williamson (2000); Wright (1992), p.16). Cf. Wright (1992), pp. 15-6: "Each type of norm may further be regarded as *constitutive* of a practice, or not, depending on whether its being largely observed (if it is a descriptive norm) or its supplying defeasible reason for the making, refusal and so on of moves (if it is a prescriptive norm) enters constitutively into the identity of the practice concerned." He subsequently discusses what sort of norms of assertion the deflationist is committed to, *viz.* being positively descriptively and prescriptively normative of *any* assertoric practice. This appears sufficiently close to Williamson's definition of "constitutive" and Wright himself mentions but then doesn't use the term.

(Wright, 1992, p. 19) One could interpret this broadly to mean that an agent aiming to have a warrantedly assertible belief must, necessarily, be aiming at having a true belief. But what could this mean and what does it mean for the deflationist? I will argue that if we interpret Wright's argument for the role of truth in norms of assertion this way, then his inflationary argument will be unsuccessful.

As mentioned previously, critical to Wright's inflationary argument is that the only factor that explains the normative coincidence but extensional divergence between truth and warranted assertion is some property of truth. However, in what follows I will argue that it need not be a property of truth that explains this divergence, but a defining property of *warrant*. For this discussion I will borrow from the rich literature of epistemic justification.[7] Epistemic externalists — specifically, reliabilists — take the position that a belief p is justified iff it was formed through a reliable (cognitive) belief forming process.[8] What is important for my purposes is how "reliable" is understood; namely, that a process is reliable iff it produces a sufficiently high proportion of *true* beliefs to false beliefs. What value suffices for "reliable" is not my present concern. What is important is that a fundamental property of epistemic justification is that a belief is justified iff it properly "aims" at being true (*viz.* was formed by a reliable belief forming process).[9] I suggest that the picture is similar, for a fundamental property of an assertion being warranted is that it also "properly aims at truth." Aiming at warranted assertion really is aiming at truth. However, this is a defining property of *warrant* rather than of truth. The fact that warranted assertion is connected to truth is that warrant must properly aim at truth; that is, it is not in virtue of some property of truth that norms of warranted assertion necessarily involve an important role for truth.

[7] For example see Alston (1989) and (2005), Goldman (1976b) and (1986), and Plantinga (1993a) and (1993b). Since there is a very large body of support for the position that norms of belief and norms of assertion are very closely linked, I find it largely uncontroversial to discuss epistemic justification as an analogue to warrant in warranted assertion. I am referring to the "belief assertion parallel." See Dummett (1981), Williamson (2000), Adler (2002), and Douven (2006).

[8] Goldman (1976a)

[9] In fact, using a particular theory of justification may not be required for my argument. All that is required is that there can be justified false beliefs and unjustified true beliefs. Most theories of justification (internalism, reliabilism, virtue epistemology, etc.) allow for these.

I suggest that we are now in a much better position to explain the observation that truth and warranted assertion are normatively coincident but extensionally divergent. Let us begin with the latter. It is clear that there are true beliefs (or propositions) which an agent is not necessarily warranted in asserting. The easy case is lucky guesses. The truth of a lucky guess is not sufficient to provide warrant for assertion. Thus, the extensions diverge, because there are some instances of true propositions which are not warrantedly assertible. But, what explains this divergence in extension between truth and warranted assertibility is a lack of warrant and not some property of truth.[10] With respect to the observation of the normative coincidence of truth and warranted assertibility I believe that we can find the explanation in the analogous case of truth's role in epistemic justification. That is, a belief is justified iff it "properly aims" at truth in so far as it is the product of a reliable belief forming process. Thus, necessarily, truth plays an important role in norms of epistemic belief (justification), but this is not a property of truth: it is a defining property of justification. Analogously, it is an apparent defining property of warranted assertion that the proposition asserted properly aims at truth (perhaps even that it must be true if truth is a constitutive norm of assertion).

The upshot of this is that what explains the normative coincidence but extensional divergence of truth and warranted assertion is a property of *warrant* and not truth. Thus, I have argued that it is not a property of truth that explains this difference. Since Wright's inflationary argument critically depends on the requirement that it is some property of truth (and not warrant) that explains this difference, I have argued that Wright's argument fails. Truth plays an important role in warranted assertion, but what explains the fact that truth and warranted assertion are normatively coincident but extensionally divergent is a defining property of warrant rather than

[10] There are two ways in which truth and warranted assertion can diverge in extension. The first is for some proposition to be true but not warrantedly assertible (which is the case discussed). The other is for a proposition to be warrantedly assertible but false. This latter case is often taken to be impossible: no warrantedly assertible false propositions/beliefs. I take exception to this, but it is not critical for my purposes here. If it is possible for a warrantedly assertible proposition to be false then this is merely further support for my argument that the explanation of the divergence in extension is due to some property of warrant and not truth.

truth.

IV Conclusion

In this paper I have suggested that we could interpret Wright's arguments for truth's role in norms of warranted assertion in two ways. First, that truth is a constitutive norm of assertion. Thus, if p is warrantedly assertible, then p is true. However, Wright offers arguments against this, since success in one norm need not be success in the other. Thus, Wright would appear to contradict himself. In order to avoid this contradiction I have suggested that we could interpret Wright in another way such that truth's role in norms of assertion is that aiming at one norm is to aim at the other. So, to aim at satisfying norms of warranted assertion is to also aim at truth. This may be uncontroversial, but I have argued that if we interpret Wright in this way then his inflationary argument fails. This is because properly "aiming at truth" is a defining property of warrant. So, what explains the normative coincidence and extensional divergence of truth and warranted assertion is a property of *warrant* rather than truth. Since Wright's argument depends on some property of *truth* required to explain this relation, and no such property of truth is required for this explanation (since it's a property of warrant which does the work), then Wright's inflationary argument fails.

References

Adler, J., 2002. Belief's Own Ethics. MIT Press, Cambridge, MA.

Alston, W. P., 1989. Epistemic Justification: Essays in the Theory of Knowledge. Cornell University Press, Ithaca, NY.

Alston, W. P., 2005. Beyond "Justification": Dimensions of Epistemic Evaluation. Cornell University Press, Ithaca, NY.

Douven, I., 2006. Assertion, knowledge, and rational credibility. Philosophical Review 115, 449–485.

Dummett, M., 1981. The Interpretation of Frege's Philosophy. Harvard University Press, Cambridge, MA.

Goldman, A., 1976a. Discrimination and perceptual knowledge. Journal of Philosophy 73 (20), 771–791.

Goldman, A., 1976b. What is justified belief? In: Pappas, G. (Ed.), Justification and Knowledge. D. Reidel, Dordrecht, Holland, pp. 1–23.

Goldman, A., 1986. Epistemology and Cognition. Harvard University Press, Cambridge, MA.

Plantinga, A., 1993a. Warrant: The Current Debate. Oxford University Press, New York, NY.

Plantinga, A., 1993b. Warrant and Proper Function. Oxford University Press, New York, NY.

Williamson, T., 2000. Knowledge and its Limits. Oxford University Press, New York, NY.

Wright, C., 1992. Truth and Objectivity. Harvard University Press, Cambridge, MA.

Crespo, M.I., Gakis, D., Sassoon, G. W. (eds.),
Proceedings of the Amsterdam Graduate Philosophy Conference
— Truth, Meaning, and Normativity —
ILLC Publications X-2011-01, 50–64

Note on Horsten's *Inferentialist Deflationism*

Jönne Speck
(University of St Andrews and Birkbeck, University of London)

jspeck01@mail.bbk.ac.uk

Introduction

According to the deflationist about truth, the English expression '... is true' (the 'truth predicate') does not stand for a property. To say that 'Snow is white' is true is just saying that snow is white.

However, little agreement has been achieved how this 'just' is to be understood. Leon Horsten (2009) has now set out to exploit the resources of another, more definite programme: inferentialism. Horsten argues for *inferentialist deflationism* about truth:[1] there is nothing to truth but a set of inference rules that govern the truth predicate.

Horsten derives his inferentialist deflationism from a specific reading of formal truth theory. In the following, I will argue that this approach fails. Firstly, I summarize Horsten's argument. Then, I will level an objection on the basis of Hartry Field's recent achievements in formal truth theory (Field 2003, 2007, 2008). The third section will develop possible responses on Horsten's behalf, none of which, however, I shall find conclusive.

I Horsten's Argument

Horsten's argument has three premises.

[1] Horsten speaks of 'inferential deflationism'. I hope that my terminology clarifies where the proposal is located in conceptual space.

P1 Deflationists need to explain truth by that formal theory that currently proves the most principles of truth.

P2 The theory that currently proves the most principles is PKF.[2]

P3 PKF does not prove universal quantifications into the truth predicate over all sentences of the language with truth predicate, but is closed under *inference rules* for the truth predicate.

The first and second premises together require the deflationist to base her philosophy of truth on PKF. Given the third premise, it follows that the deflationist better be inferentialist about truth, too. Surely, this reasoning is no valid deduction by itself, Horsten modestly calls it an 'inference to the best explanation' (Horsten, 2009, p. 578). Nonetheless, it has philosophical force and would open up a new, promising route for the deflationist, if Horsten's premises were well-founded. That this is not the case I argue in the following.

However, I will not question the first premise. Horsten argues quite convincingly that this is a lesson we need to draw from the failure of earlier formulations of deflationism. If I say that 'Snow is white and grass is green' is true, then I am committed to accept also that 'Snow is white' is true and 'Grass is green' is true.

To accommodate this intuition, the deflationist needs a formal theory that proves

$$\forall x \forall y (\dot{Sent}_a(x) \wedge \dot{Sent}_a(y) \rightarrow (Tx \dot{\vee} y \leftrightarrow Tx \vee Ty)) \qquad (1)$$

I.1 Why Deflationism Should be Based Upon PKF

It is Horsten's second premise that I will challenge. First, however, I summarize what Horsten says in justification of it.

I.1.a The Failure of the disquotational Theory of Truth

Traditionally, deflationists have championed the T-Schema. Due to the paradoxes, though, it need be restricted. A consistent theory is obtained if one extends the arithmetical base theory by the Tarski biconditionals for arithmetical sentences. The result is called the 'disquotational' theory of truth.

[2] PKF is a non-classical axiomatization of Kripke's fixed point models (see section I.1.d). I have to assume the reader to be familiar with Kripke's work.

This disquotational theory, however, does not prove the universal quantification (1). Since it cannot account for the compositional intuition, it is not good enough a theory for deflationism.

I.1.b The Failure of the Compositional Theory of Truth

A better choice would be what Horsten calls the *compositional theory* ('TC') (Horsten, 2009, p. 5). It simply takes the universally quantified principles as axioms.

Nonetheless, TC still is not good enough (Horsten, 2009, p. 16f). If I say that 'Snow is white' is true, then I am committed to accept that "Snow is white' is true' is true. TC does not prove that the truth predicate can be iterated:

$$\forall x(\dot{Sent}_a(x) \rightarrow (Tx \rightarrow T\dot{T}x)) \tag{2}$$

.

For this reason, Horsten dismisses the compositional theory as well, and turns to Kripke's theory of truth (Kripke, 1975).

I.1.c Kripke's Theory

The theory of Kripke's fixed point models is much stronger than TC. It does not only contain the principles of compositionality and iteration for arithmetical sentences. For every sentence that has a classical value in the minimal fixed point the principles of compositionality and iteration also hold in the form of object-linguistic conditionals. Especially, Kripke's theory contains (1) and (2) from above.

One would therefore expect Horsten to take Kripke's theory as the basis for deflationism. But he does not. Only an axiomatization of Kripke's theory, he argues, could serve the deflationist's purpose. In fact, Horsten has in mind a specific axiomatization of Kripke's fixed point models: the theory PKF as developed in Halbach and Horsten (2006) and (Horsten, 2009, p. 19).

I.1.d Axiomatizing Kripke's Theory

Clearly, Kripke's theory cannot be axiomatized in the strict sense of the word. In the end, it contains the theory of the standard model. What can be done instead, though, is defining rules for the truth predicate that correspond to the definitional clauses of the 'Kripke jump', the monotone operator that generates the fixed point models.

These rules, added to a sufficiently strong theory of arithmetic and closed under Strong Kleene logic, now capture desirable features of Kripke's theory in a proof-theoretic setting. Especially, it proves the compositionality as well as the iteration principles for *ramified truth* up to an ordinal below ω^ω (Halbach and Horsten, 2006, p. 705). It is for this reason that Horsten wants PKF to be the basis of deflationist theory building.

I.1.e Horsten's Third Premise: A Fact About PKF

Horsten's third premise, finally, is simply a fact about the theory PKF. Indeed, it's fundamental to the Kripkean approach that *ungrounded* sentences are not ascribed a classical truth value. Consequently, for some ϕ, T of $\ulcorner\phi\urcorner$ lacks a classical value, too. Therefore, no universal quantification into the predicate 'T' over every sentence of the language can be true in the fixed point models. So, since PKF is sound with respect to the Kripkean fixed point models, it cannot prove any such universal quantification.

On the other hand, PKF is closed under inference rules. Horsten's third premise thus is just a mathematical fact about the formal theory PKF.

I.2 A Tension in Horsten's Argument

However, there is a tension in Horsten's position, indeed a fatal one, as I now turn to argue. On one hand, Horsten measures the quality of a formal truth theory by its strength, more precisely, by the range of universal quantifications it proves: the more the better. On the other hand, Horsten's argument rests on the best such theory *not* proving quantifications over every sentence. In other words, Horsten's case for inferentialist deflationism relies on there being an upper bound to the strength of formal truth theory.

In consequence, Horsten's argument goes through only if no sound theory proves unrestricted universal quantifications into the truth predicate. This I take to be an overly contentious assumption. In fact, I think it is false.

II Objection

In this section I will argue that Horsten's case for inferentialist deflationism is ill-founded. Contrary to his assumption, the best formal truth theory available today does prove 'unrestricted generalities

about truth'.

Field extends the language by a binary operator '$\rightsquigarrow$'. This new conditional is now defined quasi-inductively, that is by revision-theoretic means. In a nutshell, Field starts with a valuation c that ascribes all the new sentences '$\phi \rightsquigarrow \psi$' the non-classical third truth value of the Kleene value space. Note that trivially, any ϕ has the same value as the sentence that results from ϕ by replacing some sub-sentence ψ by $T^{\ulcorner}\psi^{\urcorner}$: the null valuation obeys *intersubstitutivity*.

On this basis, the truth predicate is interpreted, as in Kripke, by a minimal fixed point valuation v_f^c. Now, a new valuation $F(c)$ is defined as follows:

$$F(c)(\phi \rightsquigarrow \psi) = \begin{cases} 1 & \text{iff } v_f^c(\phi) \leq v_f^c(\psi) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Given this valuation, a new fixed point model is constructed that again leads to a new interpretation of $\rightsquigarrow$, and so on. Since the value of some formulae may change from 1 to 0 or back, the operator F is not monotone. Hence, different to Kripke's operator, it does not provide fixed points. Instead, it generates a *revision sequence* (Gupta and Belnap, 1993, 5C.3). The general idea is to apply the revision rule again and again, transfinitely many times (for details, see appendix I).

During this revision sequence, the value of some of the sentences '$\phi \rightsquigarrow \psi$' stabilizes. Now, Field defines the 'ultimate value' of a sentence as 1 or 0 if its value stabilizes at 1 respectively 0. Otherwise, the sentence is ultimately ascribed the third, non-classical value $\mathfrak{u}$. The general theory of revision sequences implies that some fixed point model in the sequence coincides with this ultimate valuation (appendix I.1). It is the theory of this model Δ that Field endorses, and which I think disproves Horsten's second assumption.

Since Δ is just another fixed point model Field's theory contains Kripke's and PKF. Due to the revision theoretic definition of $\phi \rightsquigarrow \psi$, however, it is based on a much stronger logic. For example, Field's theory includes every instance of $\phi \rightsquigarrow \phi$, simply because every sentence has always a self-identical value. Especially, it contains, for every term t of $\mathcal{L}_{at\rightsquigarrow}$, $Tt \rightsquigarrow Tt$.

Now, Δ obeys the following attractive feature: ϕ has the same value as any sentence $\phi(T)$ that results from ϕ by replacing some

subsentence ψ by one or more occurrences of $T^{\ulcorner}\psi^{\urcorner}$ (appendix I.2). Therefore, Field's theory also contains $Tt \rightsquigarrow T\dot{T}t$. Since the fixed point value of universal quantifications $\forall x\phi(x)$ is determined substitutionally, the ultimate value of

$$\forall x(Sent_{at\rightsquigarrow}(x) \rightsquigarrow (Tx \rightsquigarrow T\dot{T}t)) \tag{4}$$

is 1, too. Generally, Field's theory contains the principle of iteration as a universal quantification over every sentence of the extended language.

Horsten rejects the compositional theory, because it does not contain the principle of iteration as universal quantification over every (arithmetical) sentence. Instead, he champions PKF. This axiomatization of Kripke's theory, however, does not prove the stronger

$$\forall x(Sent_{at}(x) \rightarrow (Tx \rightarrow T\dot{T}x)) \tag{5}$$

i.e. the universal quantification over *every* $\mathcal{L}_{at}$-sentence.

Field's theory now proves this principle, which just is a special case of (4). Therefore, if Horsten wants his argument against DT and TC to hold, he has to accept that Field's theory outruns PKF.

Now, Horsten's argument for inferentialist deflationism looks much less convincing. The model Δ validates (4). Thus, Field's theory contains universal quantifications into the truth predicate, over every sentence of the extended language. Thus, Horsten's assumption that the best formal truth theory does not prove such principles is refuted. His argument for inferentialist deflationism breaks down.

In the remainder of this paper I will discuss possible objections on Horsten's behalf.

III Discussion

III.1 Horsten On Field

In his forthcoming Horsten (to appear), Horsten rejects Field's work (Horsten, to appear, §10.2.2). His reason is the following. The schema $\phi \wedge (\phi \rightsquigarrow \psi) \rightsquigarrow \psi$ is not valid in Field's logic (Field, 2008, p. 269).[3] Any acceptable formalization of the natural language indicative conditional, however, satisfies this object-linguistic schema of *modus ponens*, Horsten assumes. Therefore, Field's theory cannot account for the real truth predicate.

[3] Let ϕ be the Curry sentence whose ultimate value is $\mathfrak{u}$, and let ψ be $\overline{0} = \overline{0}$.

Unfortunately, this argument does not square well with Horsten's project as a whole. The Strong Kleene logic of PKF, too, does not validate the modus ponens schema.[4] Moreover, even if Horsten can coherently establish the inadequacy of Field's conditional, I would still not see how this saves Horsten's argument. Assume that $\rightsquigarrow$ is no conditional but some other connective. This does not alter the fact that (4) is a universal quantification over every $\mathcal{L}_{at}$-sentence.

Maybe, however, what Horsten means is the following. Field's theory disproves Horsten's assumption only if it is the best theory currently available. Above, I argued that it is, because it proves many principles such as (4). This presupposes, though, these principles to capture the compositionality and iteration of the truth predicate of *ordinary discourse*. Maybe it is this assumption that Horsten challenges. Since $\phi \rightsquigarrow \psi$ is no adequate conditional, Field's theory does not prove the *real* principles of compositionality and iteration.

Whether Horsten does this move or not, it would not succeed anyway. The reason is simple: Field's theory contains PKF, so it is at least as good as Horsten's preferred theory. Recall that Field merely *extends* the language $\mathcal{L}_{at}$, and that the $\rightsquigarrow$-fragment of his theory is just a Kripke fixed point theory. Thus, the theory contains also every principle proved by PKF, free of the supposedly dubious new operator but just with the material conditional defined in terms of the strong Kleene operators $\neg$ and $\vee$. Hence, even if (4) and the others do not capture the ordinary discourse principles of truth, Field's theory is at least as good as PKF, that is by Horsten's assumption, the best formal theory of truth currently available, and Horsten's reasoning fails.

III.2 Deflationism and Model-Theory Revisited

Horsten has a better argument at his disposal. Recall that he dismissed Kripke's theory because it is defined semantically, as the set of sentences that receive designated value in the minimal fixed point model. This model again cannot be defined in the language of the truth predicate, but only in a meta-theory. Since the deflationist, however, needs a formal theory that captures ordinary truth talk, no semantical theory can serve the deflationist purpose. For this reason, only the axiomatic theory PKF can be interpreted deflationistically.

[4] Again, let ϕ be $\mathfrak{u}$ and ψ 0.

In a similar manner, Horsten could respond to my objection from above.[5] Field's theory is again defined meta-theoretically. More precisely, he develops his semantics in classical set theory (ZFC) (Field, 2003, p. 166). Only if axiomatized, it could rival PKF and serve as a counterexample to Horsten's argument.

At this point, Horsten could refer to a result by Welch (2008): Field's theory cannot be axiomatized. Hence, Horsten may argue, it cannot serve to explicate the meaning of the natural language truth predicate. Therefore, it is irrelevant that Field's construction validates unrestricted universal quantifications into 'T'. PKF does not contain any general principle, and since PKF is still the best theory for the deflationist purpose, Horsten's argument for inferentialist deflationism is saved.

Above, I did not question Horsten's rejection of Kripke's model-theoretic construction. But now I ask: why precisely do only the results of axiomatic theories matter for deflationism? Horsten's reason seems to be this. Since deflationists aim for an account of the *real* truth predicate, they need formal theories that can be applied to ordinary discourse. Theories such as Kripke's or Field's, that are obtained only by meta-theoretic model-theory do not fit this bill, since

> (...) we do not have a metalanguage for English. (Horsten, 2009, p. 571)

It is a reasonable assumption that deflationists aim for an account of ordinary truth talk. But how does that exclude theories that are defined in a meta-theory?

The reason Horsten suggests is that the meta-theory must be stronger than the truth theory for, as he puts it, 'familiar Gödelian reasons' (Horsten, 2009, p. 17). If this is his argument, however, then an implicit assumption is involved too, namely that no theory is stronger than ordinary discourse. This idea is not uncommon. It may be traced back to Tarski himself, who held that 'if we can speak meaningfully about anything at all, we can speak about it in colloquial language' (Tarski, 1956, p. 164).

But, as Belnap and Gupta argued some time ago, this supposed *universality* of natural language taken at face value is plainly false:

[5] In fact, this is what Horsten showed inclination to in private communication.

The theory PKF in the formal language of arithmetic plus truth is not part of ordinary discourse in English, Dutch or Chinese (Gupta and Belnap, 1993, p. 257)

III.2.a The Indefinite Extensibility of Ordinary Discourse

Fairer, maybe, it is to take Tarski to claim an indefinite extensibility of natural languages: 'anything whatsoever can be expressed in them once suitable resources are added' (ibid.). This again is certainly right, but holds just as much of any language; especially of all the formal theories for which we have set up well functioning meta-theories.

III.2.b Semantic Self-Sufficiency

What then do we make of the claim that ordinary discourse does not need a meta-theory? The very least that a meta-theory is needed for is to do semantics. Horsten's claim thus becomes that of the *semantic self-sufficiency* of ordinary discourse: Ordinary discourse is capable of doing its own semantics, especially of deriving every semantic fact about itself. This is why it doesn't need a meta-theory. And this is also why any theory that needs a meta-theory cannot serve as a formalization of ordinary discourse.

III.3 The Supposed Semantic Self-Sufficiency of Ordinary Discourse

Horsten argues for inferentialist deflationism about truth on the assumption that PKF is the strongest formal theory currently available. Against this, I pointed out that Field's theory is stronger, and doesn't support Horsten's argument for inferentialist deflationism. However, since Field's theory is not axiomatizable, Horsten can respond to my objection and reject the relevance of Field's theory for deflationism about truth, *if* ordinary discourse is semantically self-sufficient.

However, this is quite a contentious assumption. For one, it presupposes that the semantics of a natural language can be formulated in this very language. Different to the formal languages considered above, however, a natural language is an essentially *indeterminate* object. Natural languages gain and lose vocabulary and also their grammar changes continuously. Is ordinary discourse supposed to provide the semantics of every such stage? This is absurd. Not only it is impossible to determine even the syntax of future stages of En-

glish. Also, its features become less and less known the further one looks into the past.

The only reasonable approach, therefore, is to consider the current stage of its language. This again commits Horsten to the semantic self-sufficiency of our discourse *today*. If only speaking the English of the year 2099 we could interpret it fully, what would be the difference to meta-theoretical reasoning?[6]

Now we need to ask: what justifies the assumption that today we have a complete semantics of our own reasoning? Horsten does not hint at why he thinks so, but elsewhere, an interesting argument is found. Vann McGee justifies the assumption of semantic self-sufficiency from a broadly naturalist stance (McGee, 1994, p. 628). Whoever subscribes to the view that human life is '(. . .) amenable to scientific understanding (. . .)' (ibid.) must especially hold that the semantics of our common reasoning is comprehensible to us.

In order to reject meta-theoretical truth theory, however, this line of thought presupposes that ordinary speakers *now* have this understanding. Now, Gupta distinguishes between two ways this may be meant (Gupta, 1997, pp. 441n).

> In one sense, it means simply the ability to understand and use the language. In this sense it is tautological that English is comprehensible by English speakers. And nothing much follows from this triviality. In the other sense, 'comprehensibility' means the ability to give a systematic theory of English.

Therefore, the thesis of semantic self-sufficiency can hold only if ordinary speakers are capable, today, of providing a complete, scientific semantics of their reasoning. This, however, seems plainly false.[7] I can only agree with Gupta when he concludes

> The philosophical underpinnings of semantic self-sufficiency need to be carefully considered before it is used as a criterion of adequacy on theories of truth. (Gupta, 1997, p. 422).

[6] Some authors do understand semantic self-sufficiency in this weak sense, e.g. Peter Simmons in his 1993 book [p. 13n]. Maybe they do not consider that as the common reasoning of 2099, a meta-theory may be nothing more than an extension, or development, of the object-theory.

[7] Matti Eklund recently has provided a helpful overview on the positions on this question, and concluded that '(...) the arguments for semantic self-sufficiency are unpersuasive'(Eklund, 2007, p. 59)

Moreover, it is not wise anyway for Horsten to commit himself to semantic self-sufficiency as a necessary requirement on formal truth theory. PKF, namely, is not semantically self-sufficient, either.

Being a sub-theory of Kripke's, it is likewise not able to express that the liar sentence has value $\mathfrak{u}$. In fact, since it is a subtheory of Field's (section III.1) it is at least as expressively limited. Horsten claims that PKF avoids revenge, because it

> makes no claim concerning the truth value of the liar sentence. (Horsten, 2009, p. 21)

What solution, however, is made up by such quietism? Either, he means that revenge is a problem only for meta-theoretically determined theories. Then, however, Horsten would beg the question against Field. Or, he frankly admits that PKF is likewise not semantically self-sufficient. In this case, however, Horsten could not reject Field's theory because of its expressive limits, on pain of losing the basis of his own argument for inferentialist deflationism.

In the end, therefore, I do not see a way for Horsten to respond to my objection from §II. His argument for inferentialist deflationism fails in view of Field's recent achievements.

Conclusion

Pairing deflationism about truth with an inferentialist account of the truth predicate provides an attractive opportunity to specify and strengthen the deflationist position. Horsten has derived *inferentialist deflationism* from an interpretation of formal truth theory. In the present paper I argued that this approach does not succeed.

Horsten's argument presupposes the prospects of formal truth theory to be limited. The theory that proves the most universal quantifications into the truth predicate does not prove *unrestricted* quantifications. To this claim I advanced a counterexample. Field's theory proves more principles of truth then Horsten's favourite PKF. But it also proves unrestricted universal quantifications, disproving Horsten's assumption.

I then turned to discuss possible responses to my objection. First, I considered a worry about Field's proposal that Horsten raises elsewhere (Horsten, to appear). He argues that sentences $\phi \rightsquigarrow \psi$ cannot be regarded as adequate formalizations of natural language conditionals. This reasoning, however, cannot rule out Field's theory as

a counterexample to Horsten's assumption, for two reasons. First, the criticism equally applies to the conditionals of PKF and second, Horsten's classification of truth theories does not require them to prove conditional principles.

Consequently, I focused on a different response which I found motivated by his treatment of Kripke's fixed point model theory. In §III.2 I argued on Horsten's behalf that Field's work is irrelevant for the deflationist, because the theory is not axiomatizable.

This line of reasoning, however, presupposes the *semantic self-sufficiency* of ordinary discourse. I explained why this assumption is a contentious empirical claim, as well as at odds with Horsten's preference for PKF. I concluded that Horsten's argument for inferentialist deflationism fails.

References

Eklund, M., 2007. The Liar Paradox, Expressibility, Possible Languages. In: Beall, J. (Ed.), Revenge of the Liar. Oxford University Press, Oxford, pp. 53–77.

Field, H., 2003. A Revenge-Immune Solution to the Semantic Paradoxes. Journal of Philosophical Logic 32, 139–17.

Field, H., 2007. Solving the Paradoxes, Escaping Revenge. In: Beall, J. (Ed.), Revenge of the Liar: New Essays on the Paradox. Oxford University Press, Oxford, pp. 78–144.

Field, H., 2008. Saving Truth from Paradox. Oxford University Press, New York.

Field, H., 2010. Precis of *Saving Truth from Paradox*. Philosophical Studies 147, 415–420.

Gupta, A., 1997. Definition and Revision: A Reponse to McGee and Martin. Philosophical Issues 8: *Truth*, 419–443.

Gupta, A., Belnap, N., 1993. The Revision Theory of Truth. MIT Press, Cambridge, MA.

Halbach, V., Horsten, L., 2006. Axiomatizing Kripke's Theory of Truth. The Journal of Symbolic Logic 71 (2), 677–712.

Horsten, L., 2009. Levity. Mind, 555–581.

Horsten, L., to appear. The Tarskian Turn: Deflationism and Axiomatic Truth.

Kripke, S., 1975. Outline of a Theory of Truth. The Journal of Philosophy 72 (19), 690–716, seventy Second Annual Meeting Americal Philosophical Association.

McGee, V., 1994. Afterword: Truth and Paradox. In: Harnish, R. M. (Ed.), Basic Topics in the Philosophy of Language. Harvester Wheatsheaf, pp. 615–633.

Simmons, K., 1993. Universality and the Liar: An Essay on Truth and the Diagonal Argument. Cambridge University Press, Cambridge, New York.

Tarski, A., 1956. The concept of truth in formalized languages. In: Logic, Semantics, Metamathematics. Hackett Publishing Company (1983 ed.).

Visser, A., 2004. Semantics and the Liar Paradox. In: Gabay, D., Günther, F. (Eds.), Handbook of Philosophical Logic, 2nd Edition. Kluwer, pp. 617–706.

Welch, P. D., 2008. Ultimate truth vis-a-vis stable truth. Review of Symbolic Logic 1, 126–142.

Appendix

I Field's Revision Theory of $\rightsquigarrow$

Given the null valuation c, F yields a revision sequence, the following transfinite sequence of valuations $(c_0)_\alpha$.

$$(c_0)_0 = c_0 \tag{6}$$
$$(c_0)_{\alpha+1} = F((c_0)_\alpha) \tag{7}$$
$$(c_0)_\lambda = \lim_{\alpha\to\lambda} \inf (c_0)_\alpha \tag{8}$$

I.1 The Existence of Field's Δ

Revision sequences such as $(c_0)_\alpha$ eventually enter a *cycle*: there is an *initial ordinal* α_0 such that for every $\beta \geq \alpha_0$ it is the case that for any γ there is a $\delta \geq \gamma$ such that $(c_0)_\delta = (c_0)_\beta$. In other words, from $(c_0)_{\alpha_0}$ onwards, the valuations *recur* infinitely often (Gupta and Belnap, 1993, 5C.7).[8] Further, it can be proved that if the value of a sentence ever stabilizes, then it has done so at the initial ordinal (ibid., 5C8).

The existence of Δ follows from the more general 'Reflection Theorem' of Gupta and Belnap (1993, 5C.10) who ascribe it to Herzberger. Apparently, Field has recognized this connection only recently (Field, 2010, fn. 6).

Let δ be a *reflection* ordinal iff δ is $\geq$ the initial ordinal α_0 and whenever $\chi \in Sent_{\rightsquigarrow}$ stabilizes at $w \in \{0, \mathfrak{u}, 1\}$ then $(c_0)_\delta = w$. The Reflection Theorem says now that the class R of reflection ordinals is closed and unbound.

Clearly, for any reflection ordinal δ, $c_u(\chi) = (c_0)_\delta(\chi)$ for those χ that stabilize at 0 or 1. The challenge is to find one such that this holds also for all the χ that do not stabilize. Recall, however, that for any limit ordinal λ, $(c_0)_\lambda(\chi) = \mathfrak{u}$ iff χ does not stabilize below λ.

Now for arbitrary θ, the Reflection Theorem ensures the existence of the least limit ordinal in R above θ. Let Δ be this ordinal. Since Δ is $\geq \alpha_0$, any χ unstable in $(c_0)_\Delta$ never reaches a stable value. Therefore, $(c_0)_\Delta(\chi) = c_u(\chi)$ for every χ.

[8] Consult also theorem 56 of the helpful Visser (2004).

I.2 Intersubstitutivity

The intersubstitutivity of c_0 is inherited by all the $(c_0)_\alpha$ (cf. the 'substitutivity' lemma in Field (2003, p. 144). $(c_0)_\alpha(\phi) = (c_0)_\alpha(\phi(T^\ulcorner\psi^\urcorner/\psi))$ is shown by transfinite induction on α with side-inductions on the complexity of ϕ. The base is trivial (see above). Let $\alpha = \beta + 1$ and ϕ be $\chi \rightsquigarrow \xi$ for some atomic χ, ξ. Now either $\psi = \chi$ or $\psi = \xi$, assume $\psi = \chi$.

$$(c_0)_\alpha(\phi) = (c_0)_\alpha(\psi \rightsquigarrow \xi) = F((c_0)_\beta)(\psi \rightsquigarrow \xi)$$

Now since ψ atomic

$$v_f^{(c_0)_\beta}(\psi) = v_f(\psi) = v_f(T^\ulcorner\psi^\urcorner) = v_f^{(c_0)_\beta}(T^\ulcorner\psi^\urcorner) \tag{9}$$

and

$$v_f^{(c_0)_\beta}(\psi) \leq v_f^{(c_0)_\beta}(\xi) \text{ iff } v_f^{(c_0)_\beta}(T^\ulcorner\psi^\urcorner) \leq v_f^{(c_0)_\beta}(\xi)$$

we have

$$\begin{aligned} F((c_0)_\beta)(\psi \rightsquigarrow \xi) = \begin{Bmatrix} 1 & \text{iff} & 1 \\ 0 & \text{iff} & 0 \end{Bmatrix} &= F((c_0)_\beta)(T^\ulcorner\psi^\urcorner \rightsquigarrow \xi) \\ &= (c_0)_\alpha(T^\ulcorner\psi^\urcorner \rightsquigarrow \xi) \\ &= (c_0)_\alpha(\phi(T^\ulcorner\psi^\urcorner/\psi)) \end{aligned}$$

For $\psi = \xi$ proceed analogously.

Now let χ and ξ be complex of degree n and assume the claim holds for every $\zeta \in Sent_{\rightsquigarrow}$ of complexity $\leq n$. Again, we can focus on the case that $\phi(T^\ulcorner\psi^\urcorner/\psi)$ is $\chi(T^\ulcorner\psi^\urcorner/\psi) \rightsquigarrow \xi$, the other case is shown in exact analogy. We have

$$v_f^{(c_0)_\beta}(\psi) = v_f^{(c_0)_\beta}(T^\ulcorner\psi^\urcorner) \tag{10}$$

since if, on one hand, $\psi \in Sent_{at}$, then (9) holds as above, and if, on the other hand, $\psi \in Sent_{at\rightsquigarrow}$, then (10) follows from the induction assumption.

$(c_0)_\alpha(\phi) = (c_0)_\alpha(\phi(T^\ulcorner\psi^\urcorner/\psi))$ is now shown as in the induction base.

Finally, assume that α is a limit ordinal. However, since

$$(c_0)_\alpha = \lim_{\beta\to\alpha} inf\, (c_0)_\beta = (c_0)_{\gamma+1}, \quad \gamma + 1 < \alpha$$

the claim follows by analogous reasoning.

Crespo, M.I., Gakis, D., Sassoon, G. W. (eds.),
Proceedings of the Amsterdam Graduate Philosophy Conference
— Truth, Meaning, and Normativity —
ILLC Publications X-2011-01, 65–78

Modality in Brandom's Incompatibility Semantics

Giacomo Turbanti
(Scuola Normale Superiore di Pisa)

giacomo.turbanti@sns.it

In the fifth of his *John Locke Lectures*, Robert Brandom takes up the challenge to define a formal semantics for modelling conceptual contents according to his normative analysis of linguistic practices. The project is to exploit the notion of *incompatibility* in order to directly define a modally robust relation of entailment. Unfortunately, it can be proved that, in the original definition, the modal system represented by *Incompatibility Semantics* (*IS*) collapses into propositional calculus. In this paper I show how *IS* can be technically amended so to overcome this failure: the required modifications are already known and consist in adapting and including the main notions of Kripke's standard framework of possible worlds. I also show that the modifications do not jeopardize Brandom's original project.

1 Introduction

One of Wilfrid Sellars's characteristic seminal claims was that *Truth* is not a relation holding between linguistic and non linguistic items. Many fruits of this thought can be found in Robert Brandom's normative analysis of linguistic practices. In Brandom (1994), he describes *sapient* beings as engaging in practices of *giving and asking for reasons*, whose contents are defined by what speakers are *entitled* and *committed* to endorse: the commitment to one *reason* might rule out the entitlement to others, in the sense that it is *incompatible* with them. *Incompatibility Semantics* (*IS*) is Brandom's attempt to define a formal semantics as a model for those contents: his basic

idea is to define the semantic interpretant of a sentence p as the set of sentences which are incompatible with it. But *IS* is also part of a wider project. Brandom declares that with his semantics he aims to

Claim 1.

> explore the relations between normative and modal vocabulary [...], showing how normative vocabulary can serve both as a pragmatic metavocabulary for modal vocabulary and as the basis for a directly modal formal semantics for ordinary empirical vocabulary that does not appeal in any way to a notion of truth. (Brandom, 2008, p. 116)

Unfortunately, the original definitions of *IS* fail the representation of modality, but *IS* can be modified to overcome this failure, by applying some results from Göcke et al. (2008) and Peregrin (2010). These modifications do not jeopardize Brandom's project as expressed in Claim 1.

II Definitions for IS

Let me recall the essential definitions of *IS*. [1] Consider a language $\mathcal{L}$ as a set of sentences. Let an *incoherence relation*, Inc, be defined over $\mathcal{L}$: $X \cup Y \in Inc$ is to be construed as "one can't commit both to X and to Y". Let Inc obey just to the following property:

(Persistence): $X \subseteq Y \Rightarrow X \in Inc \Rightarrow Y \in Inc$.

This means that the only way to solve an incoherence is to discard some commitment. Then let an *incompatibility* function $I : \wp(\mathcal{L}) \to \wp(\wp(\mathcal{L}))$ be related to Inc as follows:

(Partition): $X \cup Y \in Inc \Leftrightarrow X \in I(Y)$.

Now, entailment is defined by exploiting the idea that X incompatibility-entails Y if and only if everything that is incompatible with Y is also incompatible with X:

($\vDash_I$): $X \vDash_I Y$ iff $\bigcap_{p \in Y} I(\{p\}) \subseteq I(X)$.

Eventually, connectives are introduced. Since the semantic interpretant of any $p \in \mathcal{L}$ is the set of sentences incompatible with p, the questions to be asked are "What is it to be incompatible with *not* p?" and "What is it to be incompatible with p *and* q?". Thus:

($\neg$): $X \in I(\neg p)$ iff $X \vDash p$;

($\wedge$): $X \in I(p \wedge q)$ iff $X \in I(\{p, q\})$.

[1] For further details see Lecture V of Brandom (2008).

III Modality

III.1 A first failure

What is it to be incompatible with *necessarily p*? It turns out it is not so obvious to express that in *IS*. Rather than simply stating a definition, I am going to tell a story about how to establish it. There are mainly two reasonings one can follow. The fist one starts from necessary cases and moves forward. Thus, to begin with:

(A): Everything that is self-incompatible is incompatible with *necessarily p*.

but also:

(B): Everything that is incompatible with p is incompatible with *necessarily p*.

What else? It is tempting to borrow from common knowledge about modality the idea that *not p* rules out *necessarily p*. Then, given the definition of negation in *IS*, the suggestion is that something is incompatible with *necessarily p* if it *does not entail p*. Thus, to put it straightly according to the definition of $(\vDash_I)$,

(C): Everything that is compatible with something incompatible with p is incompatible with *necessarily p*.

Unfortunately, this is a wrong suggestion, for the technical reason that this definition would validate both the *S5*-axiom and the converses of the Brouwerian axioms. And this situation, as it is well known, produces a collapse of modality, in the sense that $p \equiv \Box p$ turns out to be valid. In the standard framework of possible worlds, the only models that satisfy both *S5*-axiom and the converses of the Brouwerian axioms are those which contain just one single world. In the case of *IS*, it is the very semantical definition of necessity that picks up the collapsed case by simply ignoring what differentiates it from the others. To understand why, begin with noticing that, in *IS*, two incompatible sentences behave like *contraries*: one cannot commit to both, but one can just take no commitment at all. This, conversely, generates inside a language *families of compatible sentences* which do not rule out each other, in the sense that they can in principle be endorsed all together. So, to define what is incompatible with *necessarily p* as what is compatible with *not p* –

i.e. what is compatible with something incompatible with p – is to narrow the application of modal vocabulary within one single family of compatibles: that makes modal vocabulary superfluous.

But what is the alternative? The solution is to try to go beyond the boundaries of one single family of compatibles. An obvious way to do that is to require, for something to be incompatible with *necessarily* p, not only that it *does not entail* p, but that it *is compatible with something that does not entail* p. Thus, formally

$(\Box\mathrm{I})$: $X \in I(\Box p) \Leftrightarrow X \in Inc$ *or* $\exists Y(Y \cup X \notin Inc \wedge Y \nvDash p)$.

This is the definition eventually adopted in *IS* for the modal operator.

Here is where it is important to consider the second reasoning. It starts from sufficient cases and moves backwards. In an intuitive interpretation of necessity, one may say that something is necessary if nothing would prevent it. Thus, something is incompatible with *necessarily* p if something is incompatible with p. Here we meet again the suggestion that leads to the collapse of modality, but we have just analyzed it and we know how to avoid the pitfall: it is not enough to take something that is incompatible with p, we have to consider what does not imply p, since the defeasor of p might be in another *family* of compatibles. This establishes:

$(\Box\mathrm{I}')$: $X \in I(\Box p) \Leftrightarrow X \in Inc$ *or* $\exists Y(Y \notin Inc \wedge Y \nvDash p)$.

Time to take stock. We followed two reasonings that led us to two different definitions for the introduction of the necessity operator. Now the crucial question is: how are they different? In point of fact, it can be proved that, contrary to the appearances, they are equivalent in *IS*. And this becomes "the basic observation about modal formulae":

Proposition 2. $X, \Box p \vDash \varnothing \Leftrightarrow X \vDash \varnothing$ *or* $\Box p \vDash \varnothing$.

It basically says that what is incompatible with $\Box p$ has nothing to do with X: either p is necessary or $\Box p$ is self-incompatible. This is what establishes the simplest kind of necessity as represented by **S5** system.

It is worth pausing here to take a deeper look at the proof of this theorem. [2] All the trick is in the $(\Rightarrow)$ direction. It says that if some-

[2] See Proposition 3.3 in Brandom (2008, p. 144).

thing (self-compatible) does not imply p then $\Box p$ is self-incompatible. It does it by showing that

$$X \cup \Box p \in Inc \Rightarrow \nvDash p \Rightarrow \Box p \in Inc.$$

The X simply disappears. The proof is established by applying one main observation:

$$X \cup Y \notin Inc \Rightarrow X \notin Inc.$$

In fact, this is why $\exists Y(X, Y \nvDash \varnothing \wedge Y \nvDash p)$ implies $\exists Y(Y \nvDash \varnothing \wedge Y \nvDash p)$, which is equivalent to $\nvDash p$. But the real magic is in the ($\Leftarrow$) direction which 'simply' follows by *Persistence*. Notice that *Persistence* amounts to the contrapositive of the above principle:

$$X \in Inc \Rightarrow X \cup Y \in Inc.$$

The crucial point is that once X is vanished, *Persistence* makes it never come back. In this sense, what Proposition 2 shows is that the particular families of compatibles are *irrelevant*, because *any* proposition may be the defeasor of $\vDash p$. Thus, *a fortiori*, it does not matter if what invalidates $\vDash p$ is somehow indirectly, i.e. transitively, compatible with X. And this is why *S4*-axiom cannot fail.

But now one, solicited by the previous discussion, may wonder whether the irrelevance of families of compatibiles has any consequence on the problem of the collapse of modality to prospositional calculus: what does make the difference between the semantic interpretant of p and that of $\Box p$? Unfortunately, this irrelevance has the expected very bad consequence on modality: ($\Box$I) is actually equivalent to principle (C). The basic reason is that *Persistence* allows the following equivalences:

$$\exists Y(X, Y \nvDash \varnothing \wedge Y \nvDash p) \Leftrightarrow \nvDash p \Leftrightarrow \exists Y(X, Y \nvDash \varnothing \wedge Y \in I(p)).$$

Is this the tragic wreck of the whole enterprise? Hopefully not.

III.2 To persist is diabolical

Hitherto we have followed Brandom (2008). Let me now try to tell a story about why this version of *IS* fails and about how to amend it. So Proposition 2 establishes the appearance of an **S5**-sort of modality by concealing the collapse of modality to propositional

calculus. And the examination of the proof of Proposition 2 detected the axiom of *Persistence* as the main suspect for the collapse of modality in *IS*. In the previous section I suggested that the problem with *Persistence* is a problem of *relevance*. This remark helped me to qualify the problem, but now I have to admit I used it also as a bait. Those who might have swallen it, probably resonate to a certain way to construe the logical representation of necessity which has been put forward in *Relevance Logic*. At the opening of Anderson and Belnap (1975), Anderson and Belnap present the motivating reasons of the whole enterprise of relevance logic as in a par with C. I. Lewis's complaints for the so called "paradoxes of material implication" in Russell's *Principia Mathematica*, in particular,

$$p \rightarrow q \rightarrow p.$$

Here material implication only represents purely extensional relations between propositional contents and this makes any other relation *irrelevant* for the implication of a true proposition. [3] This is what Lewis avoided with his strict implication:

> In terms of *material* implication, if $pq. \supset .r$ and p is true then $q \supset r$, since $pq. \supset .r :=: p. \supset .q \supset r$. But in terms of *strict* implication, if two premises, p and q, together imply r, and p is true, it does not follow in general that $q \dashv r$; since $pq. \dashv .r$ is not equivalent to $p. \dashv .q \dashv r$." (Lewis and Langford, 1932, p. 165)

Now, it takes but a moment to realize that Lewis and Brandom work with very akin intuitions about entailment. For instance, compare Lewis's definition of strict implication upon the binary operator "$\circ$" for consistency,

$$p \dashv q =_{Def} \neg(p \circ \neg q),$$

with Brandom's definition of entailment which can be equivalently expressed as

$$p \models_I q =_{Def} \neg\exists X(X \notin I(p) \wedge X \in I(q)).$$

And yet, while Lewis construes strict implication as the proper representation for the necessary character of entailment and then

[3] See Lewis and Langford (1932, p. 85), and Anderson and Belnap (1975, pp. 3-5).

he procedes to define material implication in a different way, Brandom treats his definition as of *the* only notion of implication in his system and then he procedes to define modal operators to express counterfactually robust conditionals.

The first crucial point to notice is that, in spite of the idea of incompatibility as a directly modal notion, in this sense *IS* is a system of material implication:[4] in fact it is trivial to prove $\vDash_I p \to q \to p$. But it is important to see why. Now $\vDash_I p \to q \to p$ follows from $p, q \vDash_I p$ – which is valid in *IS* – because a standard form of deduction theorem is valid. [5] This is a typical situation you want to avoid if you care about the issue of *relevance*, but the temptation to see it here as a stark choice between two obvious principles of implication – deduction theorem and reflexivity – should be resisted, because there is more than meets the eye. A quick look to algebras for substructural logics could help. [6] Let me borrow just the essential to make my point. Consider a lattice ordered groupoid $\langle S, \leq, \circ \rangle$ and introduce a binary operation "$\to$" such that it satisfies the following property, usually named *left-residuation*:

$$a \circ b \leq c \text{ iff } a \leq b \to c.$$

Now, this property is important precisely because it shows the relations holding between operations on algebras. For what concerns us here, it enables us to see that deduction theorem does nothing but display the relation between "$\to$" and that particular sort of conjunction which is ",":

$$a, b \vDash c \text{ iff } a \vDash b \to c.$$

In other words, material implication residuates extensional conjunction. Notice there is nothing wrong with this. What we want to avoid is that material implication residuates also intensional conjunction, or fusion, "$\circ$". That would force us to accept *Augmentation*, i.e. $p \circ q \vDash p$, which is unwanted for fusion – compare with "if p is compatible with q then p is true".

[4] Here and in what follows, I rely on Brandom's representation theorem for *IS*. Brandom (2008)

[5] See Theorem 3.3 in Brandom (2008, p. 159).

[6] I suggest Dunn (1991); Dunn and Hardegree (2001), which are directly connected with the topic.

To sum up, there is a teanable position in between the two options of the troubling choice we faced above: to require both that strict implication does not residuate extensional conjunction and that intensional conjunction does not validate lower bounds, i.e. $p \circ q \leq p$. This is what both the systems of strict implication and system **R** of relevance logic require, by imposing fusion not to be idempotent.[7] Brandom, instead, does not prevent that in *IS*. He allows his conditional to be the left-residual of compatibility, but with the axiom of *Partition* he forces compatibility to validate *Augmentation*: as a result his implication behaves materially.[8]

The second crucial point to notice, then, is that Brandom's definition of the necessity operator lies directly against this material implication and does not impose any other level for the evaluation of another sort of implication. Recall again definition ($\Box$I): it can be also rephrased by saying that something is incompatible with $\Box p$ if it can be conjoined with something that does not imply p. Thus what is at play is just conjuntion and implication which, as we have just seen, are respectively extensional and material. The idea that necessity should emerge from nothing but the logic of incompatibility relations was certainly one of Brandom's *desiderata*, but since this logic in *IS* is classical and necessity cannot *technically* bootstrap out of material implication, the result is the collapse of modality.

III.3 Towards a stable system

If all the problems come from the axiom of *Persistence*, why do not we just drop it? Indeed, there are encouraging reasons to believe that this would be a good idea. Among these, there is the quite promising fact that all the characterizing formulas of normal modal

[7] Probably neither Lewis nor Belnap and Anderson ever wrote that fusion is not idempotent, and yet to require that is enough both for strict implication and for system **R** to avoid the collapse into material implication *as described above*. Since I refer to his work below, I have to note here that Read (1988, p. 128) explicitly contrasts **R**'s fusion with Lewis's consistency operator and claims that the latter validates *Augmentation*. This however, as far as I can see, needs some clarification. Lewis does define $\Diamond(pq) \dashv \Diamond p$, but this amounts to $p \circ q \dashv p \circ p$, which does not imply *Augmentation* for the consistency operator if it is not idempotent, but $p \circ p \dashv p$ is not valid. What is valid, as Read remarks, is that "any impossible proposition is inconsistent with any other proposition whatever", that is to say $\neg(p \circ p) \dashv \neg(p \circ q)$: this might be bad for relevance logic but does not affect modal logic.

[8] An extended story about this is told in Read (1988, pp. 36-50).

systems would be easily provable anyway: from rule of *Necessitation* to K-axiom, and so on. But what is best is that the validity of each single theorem would depend on the expected properties of compatibility relations: T-axiom would be valid if and only if compatibility is *reflexive* (which is the case), B-axiom would be valid if and only if compatibility is *symmetric* (which is the case), *S4*-axiom would be valid if and only if compatibility is *transitive* (which, presumably, is not the case).

But there are some discouraging facts as well. Suppose in fact we could really drop the axiom of *Persistence* without any unacceptable loss. Well, the bad news are that that would not be enough to avoid its effects in *IS*. Consider the most inescapable and apparently innocuous principle for an entailment relation, *Reflexivity*. The problem is that where, as in *IS*, entailment is defined as a relation between *sets* of sentences, *Reflexivity* becomes: $X \vDash a$ iff $a \in X$. This immediately gives a form of *Augmentation* since it cannot be denied that $X, a \vDash a$. But things are even worse. Once this is acknowledged, *Weakening* on the left can be re-established in its full generality:

Lemma 1. $X \vDash p \Rightarrow X, Y \vDash p$[9]

Proof. Assume $X \models_I p$. We show $X, Y \models_I p$ for arbitrary Y.
Suppose $Z \in I(p)$. But, as a consequence of *Reflexivity*,
$I(p) \subseteq I(Y, p)$.
Thus $Z \in I(Y, p)$ by *Transitivity* of "$\subseteq$".
Then by *Partition*, $Z \cup Y \in I(p)$. Then $Z \cup Y \in I(X)$ by $(\vDash_I)$.
Then $Z \in I(X, Y)$ by *Partition* again. Thus $X, Y \models_I p$. □

The moral to be drawn is that *IS* is too deeply entrenched in a set theoretic extensional fremework.

IV Possible worlds in IS

Until now I have talked loosely about 'families of contraries' and correspondent 'families of compatibles'. Let me now formally qualify my loose talk. Fortunately, I do not have to look far: if there is an idea deeply entrenched in the whole modern reflection on modality since Leibniz, it is the notion of *compossibility*. As Leibniz himself explains to Bourguet:

[9] A correspondent proof was originally provided by Alp Aker.

> "[N]ot all possibles are compossible. Thus, the universe is only a certain collection of compossibles, and the actual universe is the collection of all existing possibles, that is to say, those which form the richest composite. And since there are different combinations of possibilities, some of them better than others, there are many possible universes, each collection of compossibles making up one of them." (Leibniz, 1875-90, vol. III, p. 573, L. 662)

This naturally leads to the standard definition of possible worlds as *maximally consistent sets of propositions*. This idea can be easily adopted in Brandom's framework:

(PW): $PW_{Inc} =_{Def} \{S \mid S \notin Inc \text{ and } \forall X(X \cup S \notin Inc \Rightarrow X \subseteq S)\}$.

Peregrin, in Peregrin (2010), notices that a useful fact immediately follows. One of the reasons of discontent with incompatibility - entailment is that it seems to drop, together with the notion of *Truth*, also the idea that one main feature of a consequence relation is to represent the preservation of a certain semantically relevant status. But now, consider what it means to be true in a possible world in the framework just defined. Given the definition of possible worlds as maximally coherent sets of propositions, for a proposition to be true in a possible world is for it to be part of that world, in the standard sense that it is compatible with it. Notice then that it is equivalent to say that a sentence p is true in a possible world w, that $p \in w$, that everything compatible with w is compatible with p, and that $w \models_I p$.

IV.1 How to Kripke *IS*

Now that a definition of possible worlds and a notion of *Truth* have been derived, it is obviously tempting to try to do better than Brandom in *IS* by following Kripke's well-trodden path. That this can be done has already been shown by Göcke et al. (2008) and Peregrin (2010).

To begin with, recall that, metaphysical issues apart, the main problem with the reception of this idea of possible worlds as maximally coherent sets of sentences inside the standard truth-functional semantics was that to treat compossibility as consistency in a strictly bipartite evaluation of semantic contents is to crush necessity on

logical validity. Kripke's relational semantics, by pivoting on the primitive notion of *accessibility*, disentangled modal possibility from logical possibility and opened the doors to the modern analysis of modality.

Our next goal then is to define something like the *accessibility* relation with the resources of *IS* plus the standard definition of possible worlds. Fortunately, the trick to obtain *accessibility* is common knowledge. Suppose you have a space of possibilities already defined in terms of possible worlds, then a binary accessibility relation R between worlds can be introduced simply by reversing the basic definition of necessity as truth in any accessible world: just let a world w_1 be accessible from world w_2 if and only if everything which is necessary in w_1 is true in w_2:

$$w_2 R w_1 \text{ iff } \{p \mid \Box p \in w_1\} \subseteq w_2.$$

In terms of Brandom's definition of the necessity operator, that is to say that w_1 is accessible by w_2 if and only if for any $p \in w_2$ there is a subset $X \subseteq w_1$ such that $X \cup p \notin Inc$. [10] Formally,

(Compossibility): $w_2 R w_1$ iff $\forall p(w_2 \vDash p \Rightarrow \exists X(X \subseteq w_1 \wedge X \cup p \notin Inc))$

As Peregrin notices, in *IS* this amounts to treat *compossibility* as a second-level weaker compatibility: while any two possible worlds are incompatible as a whole, it might well be that any piece of the one is compatible with *some* piece of the other. This definition of accessibility very aptly fits with Brandom's own treatment of modal operators. We can simply adapt this idea here by saying that something is incompatible with *necessarily* p if and only if any possible world which contains p is *compossible* with a possible world which contains *not* p. Formally,

(PW-$\Box$I): $X \in I(\Box p)$ iff $\forall w_1(w_1 \vDash X \Rightarrow \exists w_2(w_2 R w_1 \wedge w_2 \nvDash p))$

Notice that *compossibility* inherits all the properties of compatibility. In particular, for what concerns us here, it is *reflexive* and

[10] Notice that version (C) of the introduction of necessity is adopted here. This is acceptable now, since with the *accessibility* relation we gain another parameter to play with in order to evaluate compatibility and avoid the collapse of modality.

symmetrical. Thus, it is easy to show that *IS* with (PW-□I) validates *T*-axiom and *B*-axiom. [11]. Instead *S4*-axiom fails because, in general, compossibility is *not transitive*

By mimicking Peregrin (2010)'s labels I will call this semantics which implants the Kripkean framework inside Brandom's *IS*, *Extended Incompatibility Semantics* (*EIS*).

IV.2 What it means to Kripke *IS*

In this last section I want to claim that the application of Kripke's relational semantics to *IS* does not pull any rabbit out of a hat, rather it simply makes explicit some features of modality that remain implicit in *IS*. Does that mean that Kripke's framework provide a better *semantic* metavocabulary for incompatibility? Let us see.

Before we even begin with the analysis, it is crucial to ask whether, even with this Kripkean implant, modality still collapses in *EIS*. The answer is negative. First, *EIS* does not verify *S5*-axiom, and that is enough to prevent the collapse. Second, *EIS* does not even verify the converses of the Brouwerian axioms. These results were expected: the relation of *compossibility* produces this second-level compatibility that blocks *Persistence*. This however might raise serious worries about the fulfillment of Brandom's purposes as stated in Claim 1. Thus, one may wonder whether the implant of possible worlds, while convenient from a logical point of view, is a step back from the expressive results of *IS* itself. The intended benefit of *IS* would have been the possibility to deploy a directly modal notion of entailment and to substitute the metaphysically laden semantic primitive of *truth in a world* with the pragmatically entrenched one of *incompatibility*. If, instead, it would be shown that *accessibility* is nonetheless required to obtain the same expressive results of Kripkean modal logic, then *IS* would need one more primitive and its value would quickly get lost. In other words, one may wonder whether the indirect path of *compossibility* amounts to declaring, after all, the failure of the Brandom's project with *IS*. But the answer, again, is negative. To see why, it is enough to get clear about what "directly" means in Brandom's Claim 1: while the middle step through possible worlds' vocabulary complicates the elaboration of the practices required to deploy *EIS*, the pragmatic metavocabulary

[11] For the detailed proofs, I refer to Göcke et al. (2008).

of *incompatibility* is still *sufficient* to express them.

But if this is true, then the expressive advantage of *EIS* over Kripke's relational semantics is patent. Let me try to illustrate this. In the previous section it was claimed that the sort of modality of *EIS* is the Brouwerian one of system **B**. Does that mean that according to *EIS*, or in general according to Brandom, system **B** represents the *real* modality? This question might be tricky. *EIS*, as a formal semantics, is a semantic modal metavocabulary for making explicit normative contents implicit in linguistic practices. In this sense, Kripke's relational semantics for modal logic can be construed as a similar metavocabulary. The decisive advantage of *EIS* over Kripke's relational semantics is that it is based on an independent normative analysis of linguistic practices, which provides the pragmatic metavocabulary to express it. And the expressively direct connection with such a normative analysis still holds for *EIS*. This advantage pays back not only because it cuts off metaphysical issues about possible worlds – which is not a faint result, by the way –, but also because it puts some normative flesh on the algebraic bones of the accessibility relation. So, is **B** the *real* modality? In Kripke's relational semantics the answer would be: "Well, let me check if *accessibility* is reflexive and symmetric but not transitive." But how can you tell that? In *EIS* the answer is: "Well, let me check if *compossibility* is reflexive and symmetric but not transitive." How can you tell that? Look at normative linguistic practices.

References

Anderson, A. R., Belnap, N., 1975. Entailment: The Logic of Relevance and Necessity. Vol. I. Princeton University Press, Princeton.

Brandom, R., 1994. Making It Explicit: Reasoning, Representing, and Discursive Commitment. Harvard University Press.

Brandom, R., 2008. Between Saying and Doing: Towards an Analytic Pragmatism. Oxford University Press.

Dunn, M. J., 1991. Gaggle theory: An abstraction of galois connections and residuation with applications to negation and various logical operations. Logics in AI, Proceedings European Workshop JELIA 1990, Lecture notes in Computer Science. 476, 31–51.

Dunn, M. J., Hardegree, G., August 2001. Algebraic Methods in Philosophical Logic. Vol. 41 of Oxford Logic Guides. Clarendon Press, Oxford.

Göcke, B. G., Pleitz, M., von Wulfen, H., 2008. How to kripke brandom's notion of necessity. In: Prien, B., Schweikard, D. P. (Eds.), Robert Brandom. Analytic Pragmatist. Ontos, pp. 135–147.

Leibniz, G., 1875-90. Die philosophischen Schriften von Gottfried Wilhelm Leibniz, Gerhardt, C.I. (ed.). Weidmann, Berlin, repr. Hildesheim: Georg Olms, 1978.

Lewis, C. I., Langford, H. C., 1932. Symbolic Logic. Century Company, New York, reprinted by Dover Publications (New York) in a 2nd edition, 1959, with a new Appendix III by Lewis, "Final Note on System S2".

Peregrin, J., 2010. Brandom's incompatibility semantics. Philosophical Topics 2 (36), 99–122.

Read, S., 1988. Relevant logic: a philosophical examination of inference. Blackwell, Oxford.

Crespo, M.I., Gakis, D., Sassoon, G. W. (eds.),
Proceedings of the Amsterdam Graduate Philosophy Conference
— Truth, Meaning, and Normativity —
ILLC Publications X-2011-01, 79–92

Rules Regresses

Jan Willem Wieland
(Ghent University)[1]

Jan.Wieland@UGent.be

Is the content of our thoughts determined by norms such as 'if I know that p, then I ought to believe that p'? Glüer and Wikforss (2009a) set forth a regress argument for a negative answer. The aim of this paper is to clarify and evaluate this argument. In the first part I show how it (just like an argument from Wittgenstein (1953)) can be taken as an instance of an argument schema. In the second part, I evaluate the relevant premises in some detail, and argue that the dialectical situation is slightly more complicated than a 'dilemma of regress and idleness', as Glüer and Wikforss have dubbed it.

I Introduction

Content Determining Normativism is the following thesis:

CD The content of a subject S's thoughts is determined by the norms governing S's reasoning. (Glüer and Wikforss, 2009a, p. 54)

Glüer and Wikforss (henceforth G&W) point out that CD Normativism is to be distinguished from Content Engendered Normativism on the one hand, i.e. the thesis that the content of our thoughts engenders certain norms, and from meaning Determining/Engendered Normativism on the other, i.e. the same thesis in terms of meaning rather than content. Yet, in the following I shall

[1] I am grateful to Marc Staudacher, Åsa Wikforss and the participants of AGPC10 for their comments. I am PhD fellow of the Research Foundation Flanders at Ghent University.

focus on CD Normativism only. Also, there is a strong and a weak version of CD depending on whether the determination by norms is all there is to content or whether this plays only a partial role. Following Glüer and Wikforss (2009a, p. 54), I shall consider CD in general.

CD is about what norms? Here are two candidates from Glüer (2009b, §3.2); note that I have put the obligations in the consequent, and that throughout the paper I assume that p and q are to be substituted for sentences):

- If I know that p, then I ought to believe that p.
- If I believe that p and that if p then q, then I ought to believe that q.

One of the main aims of G&W's 'Against Content Normativity' (2009a) is to disprove CD. Their strategy is as follows: "We are going to suggest that there cannot be such rules." (2009a, p. 54) In particular, some regress arguments for this position are on offer: one concerning a regress of motivations, one concerning a regress of contents, and one concerning a regress of implicit norms. In the following, I will focus on the second case, viz. the regress argument of contents, and set the others aside. I have selected this case, because it is immediately directed against CD Normativism (cf. Glüer and Wikforss, 2009a, p. 56). By contrast, the two other regress arguments are directed against slightly different claims (e.g. that belief formation is motivated by rules), and it remains to be seen how CD Normativism and possibly other positions are exactly committed to these.

Here is the relevant text at length:

> As we said, all CD Normativists are committed to the following: [CD, cited above]. This holds for S's beliefs quite as well as for any other of S's intentional states, including S's intentions and other pro-attitudes. Thus, already the requirement of a pro-attitude toward what is in accordance with a rule R clearly leads into a rule-regress for CD Normativism. Let us call this the *regress of contents*. Its moral is the following: CD Normativism cannot, on pain of vicious regress, construe any kind of intentional mental state as a condition on rule-following. (Glüer and Wikforss, 2009a, p.57)[2]

[2] For a version of this argument cf. (Boghossian, 2008, p.487).

The central aim of this paper is to clarify this argument. What exactly is its conclusion? What premises are responsible for it? As it is a regress argument, it is likely that it shares the same kind of premises and inferences with a group of other regress arguments. So in the first part of this paper I set forth an argument schema, and show how G&W's argument can be taken as an instance of that schema (§II). In the second part, I evaluate the relevant premises in some detail and see how the argument can be used against CD Normativism (§III). (Note that any other argument for or against CD Normativism will be left unaddressed.)

II Reconstruction

Consider the following argument schema.

Regress Schema

1. For any item x of type i, S can φ x only if S can ψ x.

2. For any item x of type i, S can ψ x only if there is a new item y of type i and S can φ y.

3. For any item x of type i, S can φ x only if S can φ an infinity of items of type i. (1, 2; TRA, ICI)

4. S cannot φ an infinity of items of type i.

5. For any item x of type i, S cannot φ x. (3, 4; MT)

Throughout this paper, 'S' is to be replaced with an arbitrary subject, 'items of type i' with a specific domain, and the Greek letters φ, ψ with predicates which express actions involving the items in that domain. The inference rules are abbreviated as follows: TRA = Transitivity, ICI = Conjunction Introduction in the Implicatum, MT = Modus Tollens. There are three premises, i.e. lines (1), (2), (4), and two inferences, i.e. lines (3) and (5). Line (3) is the infinite regress. An alternative for this would be

3*. For any item x of type i, S can φ x only if [S can φ another item y, and S can φ yet another item z, and S can φ yet another item v, etc.].

It might be disputable whether you can reach infinity by Conjunction, but important for the argument is that the number of items exceeds S's capacity.

I would like to stress that I do not think that this schema is the most basic argument schema for regress arguments, because there are at least two others (see Wieland, In preparation). The reason why I have chosen for the above schema in this case is that G&W seem to have a conclusion of the form 'S cannot φ any item x of type i' in mind. This is explicit in the motivations case: "Belief formation motivated by rules turns out to be impossible." (Glüer and Wikforss, 2009a, p. 56)

There are many regress arguments in philosophy (ranging from epistemology to ethics), and it would be worth exploring which of them can be stated in terms of the above schema. Compare some well-known sceptical conclusions: S cannot justify any proposition or norm; S cannot resolve the liar paradox; S cannot demonstrate that B follows from A and if A then B; S cannot fix the reference of 'rabbit'. In the following I provide an instance of the schema from Wittgenstein (1953, §§185-6).

Instance 1: Rules

1. For any linguistic item x, S can fix the correct use of x only if S can use a rule to fix the correct use of x.

2. For any linguistic item x, S can use a rule y to fix the correct use of x only if S can fix the correct use of y.

3. For any linguistic item x, S can fix the correct use of x only if S can fix the correct use of an infinity of rules. (1, 2)

4. S cannot fix the correct use of an infinity of rules.

5. For any linguistic item x, S cannot fix the correct use of x. (3, 4)

Here is an example of the regress in line (3):

- S can fix the correct use of '+2' only if S can appeal to a rule such as 'for all numbers n, one ought to write n+2'.

- S can use 'for all numbers n, one ought to write n+2' to fix the correct use of '+2' only if S can fix the correct use 'for all numbers n, one ought to write n+2'.
- S can fix the correct use of 'for all numbers n, one ought to write n+2' only if S can appeal to a rule such as 'for any occurrence of 'all', the meaning of 'all' does not shift after 1000'.

etc.

This regress, or at least a version of it, is sometimes called a regress of interpretations (e.g. Glüer and Wikforss, 2009a, p. 58). The reason seems to be that each rule can be seen as an interpretation of previous rule, and not that fixing the correct use of something would be a form of interpretation. In particular, it is Wittgenstein's pupil who has to interpret the expression '+2', yet the argument above is about the teacher's abilities.

In the following, I use the argument schema to reconstruct G&W's regress argument against CD Normativism. (Note that I will use an extra premise, but as this premise just states one extra necessary condition, this does not affect the general form of the argument.)

Instance 2: Contents

1. For any thought x, S can think x only if S can be guided by a rule.
2. For any rule x, S can be guided by x only if S can have a pro-attitude towards what is in accordance with x.
3. For any rule x, S can have a pro-attitude towards what is in accordance with x only if S can think that p is in accordance with x.
4. For any thought x, S can think x only if S can think an infinity of thoughts. (1, 2, 3)
5. S cannot think an infinity of thoughts.
6. S cannot think any thought. (4, 5)

Alternatives for lines (1), (2) and (3) can be obtained via

- For any item x of type i, S can $\varphi\ x$ only if S can $\psi\ x =$ in order to $\varphi\ x$, S has to $\psi\ x$.

This would give us the following (which are universally quantified versions of the premises suggested to me by Åsa Wikforss):

1*. For any thought x, in order to think x, S has to be guided by a rule.

2*. For any rule x, in order to be guided by x, S has to have a pro-attitude towards what is in accordance with x.

3*. For any rule x, in order to have a pro-attitude towards what is in accordance with x, S has to think that p is in accordance with x.

Now the overall dialectic of the argument is as follows. CD Normativism is to be committed to premise (1), and if the rest of the premises is equally in place, then that view would entail that we cannot think any thought. As this is an absurd result, CD Normativism has to go.

III Evaluation

The reconstructed regress of contents argument from the previous section has four premises. If we assume that all inferences are valid, then there are four options to resist it, viz. by denying one of the premises. This is interesting because G&W suggest that there is only one option (viz. idleness) next to the regress. I will turn to this at the end of this section. First I go through the premises one by one.

Premise (1): For any thought x, S can think x only if S can be guided by a rule.

Here, the issue is not whether this is plausible in general, but only whether CD Normativism is committed to it (rather than any other position). It seems clear that this premise follows from CD (see §I) as long as CD is read fully unrestricted: the content of *all of* a subject S's thoughts is determined by the norms governing S's reasoning.

As a consequence, CD Normativism may resist the premise by holding that the content of many, but not all, of our thoughts is

determined by rules. Yet, this restriction strategy would need proper motivation (just like restriction strategies to resolve paradoxes, for example). In this case it is to be shown why there would be two sorts of thoughts, viz. those for which the content is determined by rules, and those where this is not the case.

Premise (2):: For any rule x, S can be guided by x only if S can have a pro-attitude towards what is in accordance with x.

If any rule is to determine the content of my thoughts, then the idea of this premise is not that I am required to hold firm or even true beliefs about what is in accordance with the rule and what is not, but I minimally need to have a pro-attitude towards that. This means, simply put, that I should want what is in accordance with the rule. Compare the actions case. If the rule 'for any number n, I ought to write n+2' plays a role in the course of my actions, then I at least want what is in accordance with this rule. Yet, why not suppose, as some readers of Wittgenstein have suggested (e.g. (Wright, 2007, pp. 496-8)), that the rules might remain implicit and that we may follow them 'blindly' without such pro-attitudes?

G&W's argument here is that pro-attitudes are needed to distinguish rule-determined content from content which is merely in accordance with a rule (Glüer and Wikforss (2009a, pp. 57-9), cf. Glüer and Pagin (1999, p. 208), Boghossian (2008, pp. 480ff).) Take the actions case again. If I have not at least a pro-attitude towards what is in accordance with 'for any number n, I ought to write n+2', then on what grounds can it be said that this rule guides me whatever I do? Even in the case where I write the right series of numbers, then my pro-attitude is needed to distinguish my rule-guided behaviour from behaviour which is merely in accordance with the rule, i.e. from regular, mechanical behaviour or behaviour that is correct only by accident.

Also: if rules remain implicit, and do not fulfill the roles just outlined (viz. guide our actions, determine our thoughts), then it is not clear what their role is. That is, in that case the rules are presumably idle (Glüer and Wikforss (2009a, p. 60), they refer here to Quine (1979, p.106).)

Premise (3): For any rule x, S can have a pro-attitude towards what is in accordance with x only if S can think that p is in accordance

with x.

The basic idea of this premise is that pro-attitudes involve thoughts (viz. mental content), and so one cannot have a certain pro-attitude without thinking the corresponding thought. In the following I will suggest that it is plausible that intentional states in general involve thoughts, but that the connection between pro-attitudes and thoughts comes with a complication.

Intentional states in general are states where someone is mentally directed at other states. Familiar intentional states are belief possessions, i.e. states of the form 'S's believing that p' where S is directed at the believed state that p. Furthermore, if S believes that p, then it is plausible to suppose that sometimes S thinks the thought that p as well. So at least some intentional states involve mental content, and the question is whether this holds for pro-attitudes as well.

The complication is that it is not easy to see what thoughts might be involved with pro-attitudes towards 'what is in accordance with R'. There is a possibility to get thoughts, but then we have to suppose that these pro-attitudes involve practical inferences of the following format (varieties of these inferences are described in Glüer and Pagin (1999, §1, esp. p. 217):

PA_1 I want what is in accordance with R.

B That p in accordance with R.

PA_2 Hence, I want that p.

The first premise is the initial pro-attitude (PA_1), the second premise is a belief (B; again: this belief need not be true or whatever), and the conclusion is the final pro-attitude (PA_2). Only the latter pro-attitude is an intentional state of the form 'S's wanting that p' where S is directed at the approved state that p. Furthermore, both B and PA_2, but not PA_1, may involve a thought. Believing that p is in accordance with R may involve the thought that p in accordance with R, and wanting that p may involve the thought that p. Also, 'that p' might be general or rather specific. For example, if the rule is 'for any number n, I ought to write n+2', then 'that p' might be general or rather specific:

- that I ought to write n+2 just after n, for any number n;

- that I ought to should write 1002 just after 1000.

Note that I used the thoughts involved with B for premise (3), but the thoughts involved with PA_2 will do as well. In any case, my point is that pro-attitudes involve thoughts (and (3) holds) only if S makes such practical inferences (or at least holds such beliefs).

Yet, why would the CD Normativist not just grant that pro-attitudes are indeed required for rule-following (premise 2), but deny that pro-attitudes involve mental content (premise 3), so that the regress argument is stopped? Perhaps this route is unavailable, because if pro-attitudes would not involve thoughts (with general or specific content), then there is no use to appeal to them to explain why a thought is determined by a certain rule, rather than another rule. Compare the action case once more: "By virtue of what is it true that I use the '+' sign according to the rule for addition and not some other rule?" (Boghossian, 2008, p. 491)

Furthermore, if the CD Normativist bites the bullet in this, then it reduces to the view that content is determined by rules irrespective of any differentiation among the latter. If this is unacceptable, then the motivation of the premises so far can be summarized as follows. Pro-attitudes are needed to explain why thoughts are determined by rules (rather than not), and further thoughts (related to those pro-attitudes) are needed to explain why thoughts are determined by certain rules (rather than others).

Premise (5): S cannot think an infinity of thoughts.

If this holds, then S cannot do what is required to entertain t_1, and so cannot entertain t_1 (or any other thought). But does it hold? Consider the list of thoughts that S should be able to think:

- the thought that p_1 is in accordance with R_1;
- the thought that p_2 is in accordance with R_2;
- the thought that p_3 is in accordance with R_3;

etc.

CD Normativism is not committed to holding that the content of each thought is determined by a different rule (moreover, that would be rather surprising). So, if the rules R_1, R_2, R_3, etc. could just

be the same, one may wonder whether the thoughts just listed are not just the same as well. Moreover, if they are not distinct, then it is not obvious that S cannot have 'so many' of them (and hence it would not be established that S is unable to entertain any thought in the first place).

Yet, it seems they must be distinct after all. The reason is that the content of each thought t_n is determined only thanks to the content of a further thought t_{n+1}, viz. the one that is involved in one's pro-attitude towards the rule which determines the content of tn (cf. Fig. 1). Simply put, if the thoughts were identical, they had to play a role in the determination of their own content. If this is absurd, then the thoughts must all be distinct.

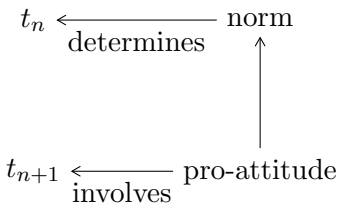

Figure 1:

Summing up, the CD Normativist has in principle the following options:

(i) reject (1) by defending that the content of only one group of thoughts is determined by norms;

(ii) reject (2) by defending that content might be determined by norms even if we do not have pro-attitudes towards the latter;

(iii) reject (3) by defending that the relevant pro-attitudes do not involve mental content;

(iv) reject (5) by defending that it is not impossible to entertain an infinity of thoughts (e.g. if they are identical);

(v) accept the whole argument and the sceptic conclusion (6) that no-one is able to think.

Not all of these options are equally worth exploring, but my main point is that the dialectical situation is somewhat more complicated

than a 'dilemma of regress and idleness' (Glüer and Wikforss, 2009a, p. 54). In particular, idleness is only related to one of the five options listed above, viz. (ii). Even if all options are carefully dismissed, the situation for CD Normativism is one of five implausible horns, rather than two.

IV Conclusion

The aim of this paper was to clarify G&W regress argument against CD Normativism. I employed an argument schema and showed how the argument can be spelled out in terms of it (§II). Also, I evaluated its premises in some detail, and argued why the situation is slightly more complicated than a dilemma between two implausible horns (§III). Let me conclude with three general remarks.

First, as was already clear from Glüer and Wikforss (2009a), 'the' regress of rules does not exist. In this paper I spelled out two rule regresses, and the Appendix hosts two more. (Another rule regress which is worth mentioning is the well-known Kripke (1982, ch. 2).)

Second, regress arguments are strong arguments, not because they cannot be resisted, but because they can be used against substantive positions like CD Normativism.

Third, argument schemes such as the one I presented in this paper for a group of regress arguments are useful for at least the following four, related reasons. On the basis of the schema it can be seen (i) what specific arguments have in common; (ii) what their conclusions are, and what not; (iii) which premises are responsible for those conclusions, and which not; and (iv) which options are available to resist the arguments, i.e. which premises may be attacked.

References

Boghossian, P. A., 2008. Epistemic rules. Journal of Philosophy 55, 472–500.

Carroll, L., 1895. What the tortoise said to achilles. Mind 4, 278–280.

Glüer, K., Pagin, P., 1999. Rules of meaning and practical reasoning. Synthese 117, 207–227.

Glüer, K., Wikforss, Å., 2009a. Against content normativity. Mind 118, 31–70.

Glüer, K., Wikforss, Å., 2010. Es braucht die regel nicht. wittgenstein on rules and meaning. In: Whiting, D. (Ed.), The Later Wittgenstein on Language. Palgrave Macmillan, Basingstoke, pp. 148–166.

Glüer, K. . Å. W., 2009b. The normativity of meaning and content. In: Zalta, E. N. (Ed.), The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/.

Kripke, S. A., 1982. Wittgenstein on Rules and Private Language. HUP, Cambridge, MA, Ch. The Wittgensteinean Paradox, pp. 7–54.

Quine, W. V. O., 1979. The Ways of Paradox and Other Essays. HUP, Cambridge, MA, Ch. Truth by Convention, pp. 77–106.

Wieland, J. W., In preparation. And So On. Two Theories of Regress Arguments in Philosophy. PhD dissertation, Ghent University.

Wittgenstein, L., 1953. Philosophical Investigations. Blackwell, Oxford.

Wright, C., 2007. Rule-following without reasons: Wittgenstein's quietism and the constitutive question. Ratio 20, 481–502.

Appendix

In the following I provide two more instances of the argument schema presented in §II in order to illustrate its applicability. The first is from Glüer and Wikforss (2009a, pp. 55-6).

Instance 3: Motivations

1. For any belief x, if S can form x only if S can be motivated by a rule.
2. For any belief x, S can be motivated by a rule only if S can form a belief y that to believe that p is in accordance with x.
3. For any belief x, if S can form x only if S can form an infinity of beliefs. (1, 2)
4. S cannot form an infinity of beliefs.
5. S cannot form any belief. (3, 4)

Glüer and Wikforss (2009a, pp.55, fn. 55) note that this motivations' regress is similar to the one by Carroll (1895). I am not sure. I take the moral of Carroll's regress to be that Achilles never demonstrates that the Tortoise is forced to accept a conclusion if he adds extra premises of the form 'if the foregoing premises are true, the conclusion must be true' to the argument. No such problem seems at play in G&W's case. The version of Boghossian, different from any of the arguments discussed above, is closer to the Carroll case (as he himself acknowledges):

> If on the Intention View, rule-following always requires inference; and if inference is itself always a form of rule-following, then the Intention View would look to be hopeless: under its terms, following any rule requires embarking upon a vicious infinite regress in which we succeed in following no rule. (Boghossian, 2008, pp. 492-3)

My reconstruction:

Instance 4: Inferences

1. For any rule x, if S can follow x only if S can infer what x calls for in the circumstances in which S finds herself.

2. For any rule x, S can infer what x calls for only if S can follow another rule y (i.e. 'from x and the circumstances, one ought to infer such and such').

3. For any rule x, if S can follow x only if S can follow an infinity of rules. (1, 2)

4. S cannot follow an infinity of rules.

5. S cannot follow any rule. (3, 4)

Note: G&W do not buy this one, and reject (2). Specifically, they deny that inference involves following a rule (2009a, p.57, fn. 58); (2010, pp.162-4).