IDÉES FIXES

A festschrift dedicated to Christian Bennet
on the occasion of his 60th birthday

Martin Kaså (ed.)
IDÉES FIXES
IDÉES FIXES

A FESTSCHRIFT DEDICATED TO CHRISTIAN BENNET
ON THE OCCASION OF HIS 60TH BIRTHDAY

EDITED BY
MARTIN KASÅ

Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
<table>
<thead>
<tr>
<th>Author/Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDITOR’S PREFACE</td>
<td></td>
</tr>
<tr>
<td>JAN ALMÄNG: THE REAL LESSON OF KANT’S SECOND ANALOGY</td>
<td>9</td>
</tr>
<tr>
<td>TOVA BENNET &amp; SUSANNA RADOVIC: MORDISKA SJÖLEJON OCH MARSIANER</td>
<td>23</td>
</tr>
<tr>
<td>RASMUS BLANCK: TWO CONSEQUENCES OF KRIPKE’S LEMMA</td>
<td>45</td>
</tr>
<tr>
<td>ALI ENAYAT: STANDARD MODELS OF ARITHMETIC</td>
<td>55</td>
</tr>
<tr>
<td>FREDRIK ENGSTRÖM: IMPLICITLY DEFINABLE GENERALIZED QUANTIFIERS</td>
<td>65</td>
</tr>
<tr>
<td>BJÖRN HAGLUND: IN DEFENCE OF THE TURING TEST</td>
<td>71</td>
</tr>
<tr>
<td>THOMAS LINGEFJÄRD: THE MEETING BETWEEN A MATHEMATICS EDUCATOR AND A PHILOSOPHER OF MATHEMATICS</td>
<td>75</td>
</tr>
<tr>
<td>HELGE MALMGREN: PARADOX LOST</td>
<td>81</td>
</tr>
<tr>
<td>JÖRGEN SJÖGREN: PHILOSOPHY OF MATHEMATICS AND MATHEMATICS EDUCATION: SOME REFLECTIONS</td>
<td>85</td>
</tr>
<tr>
<td>CHRISTER SVENNERLIND: THE SWEDISH PENAL LAW ON ACCOUNTABILITY AND SEVERE MENTAL DISORDER</td>
<td>101</td>
</tr>
<tr>
<td>ANDERS TOLLAND: PERSUASIVE DEFINITIONS AND SOCIAL EXTERNALISM</td>
<td>117</td>
</tr>
<tr>
<td>DAG WESTERSTÅHL: PAIR GRAMMARS AND COMPOSITIONALITY</td>
<td>121</td>
</tr>
</tbody>
</table>
EDITOR’S PREFACE

Dear Christian,

A reminiscence: I had recently arrived at Dicksonsgatan 4. Having previously studied mathematics and philosophy in Lund (and having read GEB: EGB), I started at the b-level in logic. The teacher was you (of course!), and the set of students = ⋂ Martin, Glenn, Jonas}. The first written exam was for the Mendelson course and I felt quite good after the ordeal – until I woke up in the middle of the night, realizing that had my faulty reasoning on the last assignment worked, then PA ⊢ Con(PA). The next day I humbly asked you to ignore what I had written, whereupon you just smiled, your very own friendly smile, and said “I’m afraid that is much too late”.

A few years later, I tried being you, i.e., being the logic teacher for a semester. Well, in fact, my first several years of teaching was a prolonged imitation game: examples, choice of words, gestures, mannerisms, all the little things that make up a teaching style was just me copying you. I think I’ve found my own voice now, but I am positive that the reason it worked at all the first years was that the pedagogical force of your performance and your vision to some extent managed to enlighten students even when merely channeled through a disciple.

You and I have shared many pleasant events, talks, thoughts and experiences – work related and not – over the last 15 years. One experience of a not so pleasant kind was that of an imminent threat of being laid off. In the end, it worked out OK, sort of. We were not dismissed, and I find the department in most ways a much better place to work at today than it was back then. But something is amiss. You are missed. We miss you. I miss you.

Though I of course hope and believe that this volume will make you happy, I also know that an apology is in order, so here goes. I am sorry for having spent time working as an editor, rather than as the author of my next journal article. In my defence: look at the result! It can’t count as tooting my own horn when I say that this is a splendid book – I am really just praising the authors. Whom, by the way, I had no problem whatsoever recruiting.

Oh, and I hope you have had a reasonably enjoyable birthday celebration, so far. Many happy returns!

Martin Kaså
THE REAL LESSON OF KANT’S SECOND ANALOGY

Jan Almäng

In his second analogy Kant presents us with a distinction between two kinds of perceptual sequences. In the first case one perceives a change of some kind. In the second case one has a succession of different perceptions. This generates what is in the present context coined “Kant’s problem”, or how we should explain this observation. Framed in terms familiar from the modern context, Kant’s distinction is based on the observation that sometimes one can be aware of a change in perceptual content, without perceiving any change in the surrounding environment. In this paper it is argued that this distinction has received too little attention in the modern discussion of time-consciousness. It is also argued that the moral of Kant’s distinction is one that Kant himself did not see, namely that the intentional content of an act of perception must be distinct from its intentional object, and that there is an awareness of the intentional content, which is not an awareness of the object of perception.

Keywords: Kant, Perception, Temporal consciousness

1 INTRODUCTION

In his second analogy Kant (1998, A189–211) sets out to prove that it is impossible to perceive an event without already presupposing the category of causality. Whereas Kant’s proof of the so called causal principle has been extensively discussed in the literature, the relevance of his argument with respect to the phenomenology of time-consciousness has received relatively little attention. If Kant is correct however, there is a clear phenomenological difference between experiencing a change in one’s perceptions and perceiving an object change.

According to Kant the temporal order of a succession of perceptual experiences need not be given as temporally irreversible. When we look at a house, the gaze might wander from the roof to one of the windows and then to a door. In this case we are certainly aware that our perceptions change. But we see no
event or process. When we see a process however, we see something that is perceptually presented as temporally irreversible. If you see a ship move down the river, the successive order of your perceptions is presented as irreversible in the sense that had the order of the perceptual states been different, you would not have been presented with the process you actually were presented with. This indicates according to Kant that an awareness of a change in one’s perceptions is not equivalent to a perception of a change.

In order to see what Kant is getting at, we shall make a tristinction between states, processes and events. Let us say that states and processes are by definition temporally extended, whereas events are temporally punctual. States and processes have temporal parts. Events have no temporal parts.

A state may be constituted by an object exemplifying a property during a certain period of time. Thus, for example, if an object is blue, it is in the state of being blue for as long as it is blue. But a state may also be constituted by two objects bearing a certain relationship to each other during a certain period of time. States however, involve no change. Whereas a state occurs for a certain period of time, its temporal parts are qualitatively identical.

Unlike states, processes involve a change of some kind. A process consists in a change in an object, be it with respect to its properties, the relations it bears to other objects, or to its spatial position. When an object moves from one position to another, it is consequently involved in the process of changing position. Processes are like states in that they are temporally extended; but they are different from states in that not all their temporal parts are qualitatively identical. Processes can consist of a sequence of states, as when an object is first red and subsequently green. But a process can also consist of a continuous change of some kind – as when an object is continuously moving along a certain trajectory.

Events on the other hand are punctual. Events occur when an object changes in some respects. Events can consequently occur when an object ceases to be in one state and enters another state. If an object is red but becomes blue, the change from red to blue is an event. We can express this by saying that the temporal boundary between the states is an event. If an object is a constituent of a continuous process, then each point in time during the process is a separate event. Thus, each point in time during a continuous movement consists of a separate event.

States, events and processes are what W.E. Johnson once called *occurrents*. In this they are to be distinguished from Aristotelian substances, which according to Johnson are *continuants* (Johnson, 1924, ch. vii).1 A continuant is present in its entirety at every point in time at which it exists. So for example, I am wholly present at each point in time at which I exist. An occurrent on the other hand is something that has temporal parts. An occurrent is consequently present in its

---

1This is obviously a contested point in modern philosophy. Not much in the present discussion will however hinge on whether substances actually are continuants or not.
entirety throughout its life span. Substances cannot exist except as constituents of states, events and processes. For either a substance has a property that persists for a period of time – in which case it is a constituent of a state – or it is changing in some respect. When the latter is the case, the substance is a constituent of a process and an event.

Returning to Kant, what he appears to be saying is that a perceptual process does not entail a perception of a process. In the former case, the perceptual content is changing during the process – this is what makes the sequence a perceptual process as opposed to a perceptual state – but this does not entail that we perceive something that is actually changing, namely a process. It is only in some cases that a perceptual process is accompanied by a perception of a process. We can call this Kant’s phenomenological observation. But this gives rise to a problem: How can a perceptual process be a perception of a process in some cases, but not in others? This is in effect Kant’s Problem which Kant’s theory of perception of causation is supposed to solve.

In this paper I shall discuss and present Kant’s observation and the problem it gives rise to. I shall also discuss Kant’s own preferred solution to the problem, namely the positing of a pre-existing category of causation. But it is important to stress that the present paper is not about the exact interpretation of Kant’s solution. I harbour no exegetical ambitions. Whereas the right interpretation of Kant’s problem and his own solution of it are certainly interesting questions, I shall make no attempts towards answering them.

My claim is that the real lesson of Kant’s problem is that the problem cannot be solved unless we make a distinction between the object of an act of perception and the content of the said act of perception. Any attempt of solving the problem which denies this distinction, or which attempts to reduce the notion of content to the notion of object or vice versa, is doomed to fail. In the second section of the paper, I shall present a version of Kant’s problem. In the third section I shall discuss Kant’s own solution to his problem and argue that it fails. And in the fourth section I shall argue that the real lesson of Kant’s problem is that we must distinguish between perceptual content and perceptual object.

2 KANT’S PROBLEM

Kant’s aim in the second analogy is to show that when we perceive an event, we cannot see it if not a category of causality is presupposed. The precise formulation of this claim and what arguments Kant employs to prove it have been discussed for more than two centuries. Kant’s phenomenological observation is built upon some very astute observations regarding the nature of the perception of time. These observations give rise to some intriguing problems.

Kant makes a comparison in the second analogy between perceiving a house and perceiving a moving ship, or, put more generally, between perceiving a substance and perceiving a process. When we perceive a house, we might first take
a look at one of its visible parts, for example, the roof, then we might move our
gaze to the floor, and then to one of the windows. In this case, the temporal or-
der of our perceptions is contingent in the sense that whereas our perceptions
of the object are changing, we are not seeing a changing object. The latter is
however precisely what characterises perceptions of events. Here is Kant:

And in the translation of Norman Kemp Smith: “But, as I also note, in an appearance which
contains a happening (the preceding state of the perception we may entitle A, and the succeeding
B) B can be apprehended only as following upon A; the perception A cannot follow upon B but
only precede it. For instance, I see a ship move down stream. My perception of its lower position
follows upon the perception of its position higher up in the stream, and it is impossible that in the
apprehension of this appearance the ship should first be perceived lower down in the stream and
afterwards higher up. The order in which the perceptions succeed one another in apprehension
is in this instance determined, and to this order apprehension is bound down. In the previous
example of a house my perceptions could begin with the apprehension of the roof and end with
the basement, or could begin from below and end above; and I could similarly apprehend the
manifold of the empirical intuition either from right to left or from left to right. In the series of
these perceptions there was thus no determinate order specifying at what point I must begin in
order to connect the manifold empirically.” (Kant, 1929, A192f/ B237f)
Kant himself seems to frame the difference between our two perceptions as a difference between perceiving a substance and perceiving an event. This way of explicating the difference is however at best misleading since we are seeing a substance in the case of the perception of the ship as well. The difference seems rather to be that when we are seeing the ship, but not when we are seeing the house, we are seeing a substance that is a part of a process of moving along a trajectory. I shall express the difference as the difference between seeing a state and seeing a process.

We would at this point do well to note that the term “perceptual object” is notoriously ambiguous in perceptual theories. They may refer to an object in the sense of an Aristotelian substance, such as a car, a person or a tree. They may however in other contexts refer to a state, an event or a process. In the present context I shall assume that we can perceive not only substances, but also their properties, relations between them, states and processes of which they are a part, and events that happen to them.

Returning to Kant, we can now formulate the difference between the perception of the ship and the perception of the house. In both cases we are seeing substances. It is only that in the first case we are seeing a process and in the second case a state. Processes have temporal parts that are not qualitatively identical. Consequently, the temporal order of these parts determines the nature of the process. If the order of the temporal parts of the process consisting of the ship’s floating downstream is rearranged, the process that these parts constitute would no longer be the process consisting of the ship floating downstream. The temporal parts of a state however are qualitatively identical. Consequently, the order of these parts is irrelevant with respect to which state the perceiver is presented with.

In a similar way, the order of the perceptions presenting a process determines which process is perceptually presented in the sequence of perceptions. But the order of the perceptions presenting a state is irrelevant with respect to which state is perceived.

Now, Kant would obviously not deny that we are in some sense aware of a succession in the case of the house as well. But in this case we are not aware of anything other than a perceptual process. In modern philosophy of mind, the fact that we are in some sense aware of our own perceptions is normally expressed in terms of the perception having a certain ‘phenomenal character’. Kant’s claim to the effect that we are aware of a succession of perceptions can be translated into the preferred parlour of modern philosophy of mind as the claim that this specific awareness has temporal parts. The phenomenal character of a perceptual experience, whatever this may be, is in other words an entity with temporal parts. Following William James, we shall call the temporal period of the process or state we are aware of the specious present.

If this analysis is correct, we may also distinguish between two different kinds of time-consciousness. Our awareness of our own perceptions as extended in time will in the present context be called experiential time-consciousness, whereas
the perception of a process will be called a *perceptual time-consciousness*.

Now, it is obviously a contested point in modern philosophy of mind whether there is an awareness of any features of a perceptual experience that is not an awareness of the object of perception. And it is consequently also contested whether there is an experiential time-consciousness that is not a perceptual time-consciousness. Indeed, we shall see that Kant himself seems to have believed that experiential time-consciousness just is perceptual time-consciousness and that the perceptual content just is the perceptual object. For the time being, I ask the reader to provisionally accept my distinction in order to explain Kant’s position. In the next section I shall argue that a reduction is not possible.

If Kant’s observation is correct, awareness of a perceptual process does not entail a perception of a process. When your eyes move across the house, you experience a perceptual process but you do not perceive a process.

It might be objected that in Kant’s example we are presented with a process in the case of the house as well. For in this case, the relationship between the eye and the house may be in a constant flux, and the changes in the perceptual content may well track that change even though the house itself undergoes no change. A.D. Smith, for example, has argued that often when we experience a perceptual process without perceiving a corresponding process, the experience nevertheless represents a change in the relation between perceiver and perceived. If we for example are closing in on the object of perception, the perceptual experience will change and the perceptual content will present a change of distance between perceiver and perceived. (Smith, 2002, p. 172)

Kant’s example might however easily be reformulated so as to avoid this objection. Consider the case of looking at a picture and not being able to see what it represents. But suddenly you begin to see persons and trees and rivers. In this case the perceptual content changes from less determinate to more determinate. The content is changing, but the spatial relationship between perceiver and perceived remains unchanged. In this case we are aware of a perceptual process, but we do not perceive a process. The object of perception is quite clearly a state.

It might seem that Kant’s own solution forces us to accept that there really is a difference between content and object and between experiential and perceptual time-consciousness. For, as we have seen, in both of our cases there is an awareness of a perceptual process. It is however only in one case that we perceive a process. In the other case we perceive a state. States can however not be identical to processes. So our awareness of our own perceptions must be distinct from our awareness of the object of perception. But as we shall see, matters are not that simple. Kant’s transcendental idealism commits him on most interpretations to an account according to which the perceptual content just is the object of perception, cf. (Strawson, 1966), and (van Cleve, 1973, p. 81ff) for alternative interpretations. So Kant has to account for his problem in other ways. It is to that solution that we now turn.
3 KANT’S SOLUTION TO KANT’S PROBLEM

Let us now turn to Kant’s solution to Kant’s problem. On Kant’s account a perception of a process will always mean that one is aware of a perceptual process. But the opposite is not true. One can be aware of a perceptual process without perceiving a process. When we perceive a process, the order of the temporal parts of the perceptual process is derived from the order of the temporal parts of the perceived process. Here is Kant:

Ich werde also, in unserem Fall, die subjektive Folge der Apprehension von der objektiven Folge der Erscheinungen ableiten müssen, weil jene sonst gänzlich unbestimmt ist, und keine Erscheinung von der anderen unterscheidet. Jene allein beweist nichts von der Verknüpfung des Mannigfaltigen am Objekt, weil sie ganz beliebig ist. Diese also wird in der Ordnung des Mannigfaltigen der Erscheinung bestehen, nach welcher die Apprehension des einen (was geschieht) auf die des anderen (das vorhergeht) nach einer Regel folgt. (Kant, 1998, A193/B238)\(^3\)

So on Kant’s account the order of the temporal parts constituting the perceptual process is irrelevant with respect to the perceived state. It is however not irrelevant when we perceive a process. In the latter case the temporal parts conforms to a rule, (cf. Strawson, 1966, p. 134). This rule determines a specific temporal order. According to Kant, this rule is a rule of causality.

The point of Kant’s argument is thus that the perception of a state leaves us with an arbitrary order of successive perceptions. When we perceive a process however, the succession of perceptions is non-arbitrary and determined. There is no consensus among Kant-scholars about the precise meaning of this claim. It seems however that most commentators agree on three claims. First, when a process is perceived, the succession of perceptions is ordered and determined in the sense that it is presented as irreversible to the perceiver. Second, this order is determined according to a rule. This rule presents the perceiver with a synthetic necessity a priori. Third, the rule in accordance with which the sequence of perceptions is ordered is a rule that presents the perceived events constitutive of the process as caused. The precise details of these claims vary among Kant-scholars. This is in particular true of the third claim, where there are some significant disagreements as to the right interpretation (cf. Bayne, 1994, pp. 381–388).

\(^3\)And in Kemp-Smith’s translation: “In this case, therefore, we must derive the subjective succession of apprehension from the objective succession of appearances. Otherwise the order of apprehension is entirely undetermined, and does not distinguish one appearance from another. Since the subjective succession by itself is altogether arbitrary, it does not prove anything as to the manner in which the manifold is connected in the object. The objective succession will therefore consist in that order of the manifold of appearance according to which, in conformity with a rule, the apprehension of that which happens follows upon the apprehension of that which precedes.” (Kant, 1929, A193/B238)
Now, due to his transcendental idealism Kant is committed to denying that there is any object of perception which is not a content of the perception. Robert Paul Wolff for example has argued that Kant’s commitment to transcendental idealism entails that the object of a perceptual representation is the representation itself. Kant cannot meaningfully distinguish between the perception and its object. (Wolff, 1963, p. 263) Put into modern parlance, the perceptual object is identical to the perceptual content. In Wolff’s interpretation, Kant solves Kant’s problem not by treating the object as something external to perceptual content, but by treating the perceptual object as a specific way of ordering perceptual content. Here is Wolff:

Subjective time order is the order of representations qua mental contents. When we synthesize these representations, we reproduce them in imagination according to a rule, and this reproduction produces an order of manifold qua representations. [...] The succession in this new order is a necessary succession. It is therefore an objective order. (Wolff, 1963, p. 264)

Kant’s solution to the problem is in other words to claim that whereas the succession of representations qua subjective time order does not need to follow any particular rule at all, the succession of representations qua perception of objects does follow a specific rule when we perceive processes.

Kant does not believe that his phenomenological observation warrants an introduction of a distinction between content and object. We could say that in its place he introduces an Ersatz distinction that is supposed to do the same explanatory work with respect to his problem. This is a distinction between a perceptual sequence which follows a rule and perceptual sequences which does not follow a rule. In both the case of the ship and the case of the house, there is no awareness of any object that is not an awareness of a content. The difference between the cases is that in the former case the order of the temporal parts of the perceptual process follows a certain rule.

Kant’s resulting theory has been coined the “cinematographic time order premise” by William Harper (Harper, 1981, p. 469). This means “that time order has to be determined by the content of the instances of the states to be ordered, much as one might attempt to order individual frames cut from a film” (Harper, 1981, p. 469). It is precisely at this point that the rules become important. For this ordering must be performed in accordance with the causal rules that Kant has introduced as the solution to the problem.

So on Kant’s account, experiential time-consciousness can after all be identified with perceptual time-consciousness. We thought that there was a difference because whereas both the ship- and the house-case involved an experience of a perceptual process, only the former also involved a perception of a process. But now we see that the difference between the perception of a process and the perception of a state is not really a difference with respect to the object of perception. Rather, it is a difference between the way the parts of the perceptual
process are ordered. In the case of the perception of a process, the ordering follows certain rules. But this is not the case in the perception of a state. So the perceptual content can after all be identified with the object of perception, and there is no difference between experiential and perceptual time-consciousness.

Whereas there is little agreement among Kant-scholars as to the exact construal of Kant’s argument in the second analogy, there is widespread (but not unanimous, cf. Beck, 1978, ch. 6) agreement that Kant failed to prove that the perceived temporal order can be explained by recourse to a causal rule. I share that opinion for two reasons.

First of all, I don’t think Kant’s own account does justice to the phenomenology involved in his examples. For if perceptual content was really tantamount to the object of perception, then a change in content would be tantamount to a perception of a process. (For a very similar point, see van Cleve, 1973, pp. 76f.) In such a case, there would be no interesting phenomenological difference between the case of the house and the case of the ship. For in both cases, we would be aware of processes. But processes are by definition ordered, so both cases should exhibit a certain order. But, this, alas, is not the case.

There is however a second and more substantial reason for doubting Kant’s solution. The rules connecting the parts of a perceptual process are supposed to be causal rules. Now, it might well be that we can perceive causal relations between events (cf. Siegel, 2005, 2010) when we perceive a process. But it is certainly not necessary in order to perceive an event that it is perceived as caused by anything. Consider the case when you perceive a computer monitor representing the results of a lottery draw. The screen is blank except for a figure representing a specific number. Initially, the figure “5” is shown, then “3” and then “7”. Clearly, we perceive in this case certain events. Equally clearly, we do not perceive this succession of events as causally related. For the fact that 5 is drawn first is certainly not perceived as the cause of 3 being drawn in the second slot.

The example can easily be generalised. Even though Kant might well be right that we can in principle see causal relations, the fact that we perceive an event does not entail that we see the cause of the event. And even if we did see the cause of the event as causing the event in a sequence of perceptions, we need not experience the sequence of perceptions as necessary in any sense.

Kant can be seen as giving an Ersatz distinction between perceiving states and processes. His is a distinction between two different processes. But then the reasons we had for not identifying perceptual content with perceptual object would evaporate. For that reason was precisely that states are not processes. But now we see that there is according to Kant no awareness of any state in the case of the house. There is an awareness of a different kind of process than in the case of the ship. But then the awareness of the sequence of perceptions may after all be identified with the awareness of the object of perception.

The lesson of the second objection is that even if we accept that there may in principle be an Ersatz distinction that could explain Kant’s phenomenological
observation, Kant’s own distinction cannot do this theoretical work. For Kant’s distinction is drawn in the wrong place. We would like to place the lottery case in the same category as the case of the ship. But since there is no perception of any causation, we will if we follow Kant have to place it in the same category as the case of the house.

Now, if it is not possible to construe an *Ersatz* distinction between perceptual content and perceptual object, it is very doubtful that the distinction between experiencing a succession of perceptions and perceiving a succession of the temporal parts of a process can be drawn without making a real distinction between perceptual content and perceptual object.

### 4 KANT’S PROBLEM AND THE CONTENT-OBJECT DISTINCTION

Kant observed that it seems as though an awareness of a perceptual process does not entail a perception of a process. In some cases we perceive a process. But in other cases we perceive a state. Since states cannot be identical to processes, the object of perception could not be identical with the content of the perception. In order to block that argument, Kant argued that the distinction was not really a distinction between a perception of a state and a perception of a process, but a distinction between two different processes. Kant in other words gave us an *Ersatz* distinction. But we saw that this distinction does not work. And in the absence of any compelling reasons to the contrary, there are good reasons to believe that there is no principled way of drawing such a distinction.

If the argument so far is correct there must be an awareness in perceptions which is not an awareness of the object of perception. This is however something that is denied in much of modern philosophy of perception. Significant parts of modern philosophy of mind are framed in terms of perceptions somehow being constituted by the object of perception. This is often developed in two ways: According to the minority view, there is no object of perception apart from the phenomenal state itself. According to the majority view however, the object of perception is itself a constituent of the perceptual content. But if Kant’s problem is correctly formulated, we must either distinguish between perceptions and their objects, or find an *Ersatz* distinction between states and processes.

One of the most famous and well-developed accounts of the minority view is the account given of time-consciousness by Barry Dainton. According to Dainton there is no mind-independent object of perception. The only object we are aware of in perception, is the phenomenal state itself. Dainton argues that the perceptual object is the perceptual content and that the latter is a mind-dependent entity, even though it seems to us as if the object is mind-independent (cf. Dainton, 2000). But if the perceptual content is held to be identical with the
perceptual object, then a change in content will result in a perception of change. This obviously contradicts the starting point of Kant’s phenomenological observation.

But Kant’s problem is not merely a problem for philosophers who argues that the perceptual object is a mind-dependent entity. It is also a problem for theories of perception according to which we can make a distinction between perceptual content and perceptual object, but which subscribe to the notion that perceptual content is transparent, in the sense that it is only the perceptual object and its properties that is available for introspection. On the latter account, when we perceive we have no awareness (intentional or otherwise) of anything but the object of perception and its properties. Perceptual content is in short transparent. We are aware of the object through the content, but not of any separate qualities of the content.

One influential defender of this thesis is Michael Tye (2002). Here is Tye: “Whatever the nature of the qualities of which we are directly aware when we focus upon how the surfaces before us look, these qualities are not experienced as qualities of our experiences but rather as qualities of the surfaces.” (Tye, 2002, p. 138) So on Tye’s account we have a perceptual content which presents us with an object. But we have no separate nonintentional experience of that content. But this no more does justice to the phenomenology of perception than does Dainton’s position. As we have just seen, we can frequently experience a change in content, without simultaneously perceiving a change in the object of the perception. This cannot be squared with the notion that perceptual content is transparent in the sense that we have no introspective access to it at all except from those properties of the object that it presents us with.

If my argument in this section is correct we must distinguish between being aware of a perceptual process and perceiving a state or process. Experiential time-consciousness is in other words a phenomenon that is quite distinct from perceptual time-consciousness. This is the real lesson of Kant’s problem. But if this is correct, perceptions cannot be transparent in the sense that when having them we are not (intentionally or non-intentionally) aware of anything but the perceived objects and its properties.

If these arguments are correct, it gives some support for an account that makes a clear distinction between intentional content and intentional object, and that claims that our awareness of our perceptions, is an awareness of the content of the perceptions. Such an account has been formulated and defended by several philosophers working in an intentionalistic tradition and perhaps in

---

4Tye is, to be sure, not alone in defending this view. For other defences, cf. (Kriegel, 2009) and (Harman, 1990).

5Cf. A.D. Smith’s criteria for whether a feature of visual experience is an apparent feature of a perceived object or not: “In general, some feature of experience is an apparent feature of an object of awareness only if: (1) if it ceases to be a feature of the experience, but the object remains an object of awareness, the object apparently changes; and (2) if you take your experience at face value, you will believe that the object does have that feature.” (Smith, 2008, p. 211)
particular by Tim Crane (2003) and David Woodruff Smith (1989). The phenomenal character of a perceptual experience depends on the specific way the intentional content is entertained in the act of perception.

The general idea behind this reasoning is that the same intentional content can be entertained in different ways, or in different modes. The origins of the distinction between mode and content can be found in Husserl’s *Logische Untersuchungen*, where Husserl makes a distinction between the quality of an intentional act, and the matter of the act (Husserl, 1984, pp. 425–431). The terms “mode” and “content” however derive from the work of John Searle (1983), whose distinction roughly corresponds to Husserl’s distinction. In the following, we shall use Searle’s terminology.

The idea is that the same content can be entertained in different modes, where visual perception is one mode, auditory perception another mode, visual imaging a third mode, and so on. Acts with different modes can have the same content. Because the phenomenal character of the acts is a function of mode and content, they will however have different phenomenal properties. Put succinctly: Seeing that the bus is arriving has a different phenomenology from imagining that the bus is arriving even though the content may be the same. Phenomenal character is a feature a perceptual state has in virtue of the intentional content being entertained in a specific perceptual mode. (Cf. Smith, 1989, p. 97)

A full blown defence of this theory will however have to be the topic of another paper. In this paper I have only tried to show some of the constraints set upon modern philosophy of perception by Kant in his second analogy. And if that argument is correct, intentionalistic theories are at least vindicated to the extent that we must distinguish between intentional content and intentional object, where the former is at least not exhausted by the latter.6

REFERENCES


6 Thanks are due to Ingvar Johansson, Helge Malmgren, Kristoffer Sundberg and members of the seminar in theoretical philosophy at the University of Gothenburg for valuable comments on an earlier version of this paper. The writing of this paper was funded by The Swedish Research Council (Research Grant 421-2011-1587).


MORDISKA SJÖLEJON OCH MARSIANER
OM FALSKA FÖRESTÄLLNINGAR OCH STRAFFANSVAR

Tova Bennet & Susanna Radovic

1 INLEDNING

En man dödar sin flickvän för att han har sett på en bild att tistlar ursprungligen kommer från Grönland. En annan man dödar sin (panka) man för att han vill ärva hans pengar. En man dödar sin mamma för att han tror att hon är en ond demon som kommer att skada honom. Ytterligare en man begår mord för att han tror att han befinner sig i krig och att offer är en soldat från fiendesidan.


Hur falska förreställningar ska behandlas i straffrättsliga sammanhang är en kompleks fråga. Huruvida en falsk förreställning utgör en vanförreställning i medicinsk mening eller ej verkar ha en avgörande betydelse, men även när det gäller bedömnings av vanförreställningar visar rättspraxis på en påtaglig variation.

Mot bakgrund av detta kommer vi i det följande beskriva hur begreppet förreställning används i en svensk straffrättslig kontext, hur man hittills har hanterat psykotiska vanförreställningar i förhållande till andra felaktiga förreställningar, samt hur vanförreställningars betydelse för frågan om straffansvar skulle komma att förändras vid införandet av ett tillräkelighetsrekvisit. Texten kommer dessvärre inte mynnas ut i ett klart och redigt förslag på en enhetlig och rätts säker hantering av falska förreställningar, men förhoppningsvis i en någorlunda klarhet i varför det inte lätt låter sig göras.

2 UPPSÅT

Inom straffrätten tillmäts trosförreställningar en helt avgörande betydelse och de analyseras också ingående under rättsprocessen. Det är framförallt vid bedöm-
ningen av gärningspersonens uppsät som denna analys äger rum, och vi ska därför inledningsvis ge oss på en liten utläggning om det straffrättsliga uppsätsbegreppet. För att en gärning (en handling eller en underlåtenhet) ska betraktas som ett brott krävs att gärningen utgör en otillåten gärning, det vill säga att den uppfyller de rekvisiter som uppställs i en brottsbeskrivning. Brottet mord är i Brottsbalken (3 kap 1 §) formulerat så här: "Den som berövar annan livet, döms för mord till fängelse på viss tid, lägst tio och högst arton år, eller på livstid." Rekvisiten för otillåten gärning här är att någon (en människa) ska ha berövat annan (en annan människa) livet. Här kan man tänka sig en rad olika situationer, en maffiaboss som skjuter ett vittne till döds, en missbrukare som knivhuggar en annan missbrukare, eller en jägare som skjuter ett skott mot en älg varpå skottet går rakt igenom älgen och träffar en skidåkare.

I alla uppräknade fall har någon berövat annan livet. Alla dessa fall skulle dock inte med nödvändighet betraktas som mord. Brottet mord blir inte komplett med mindre än att även rekvisiten för personligt ansvar är uppfyllda. Detta följer av 1 kap 2 § i Brottsbalken: "En gärning skall, om inte annat är särskilt föreskrivet, anses som brott endast då den begås uppsätligt." För att ett berövande av annans liv ska betraktas som mord krävs alltså att gärningen har begått med uppsät.


Idag beskrivs tre former av uppsät i svensk straffrätt: avsikt uppsät, insiktsuppsät och likgiltighetsuppsät. Vilken variant som är aktuell för ett specifikt brott framgår av brottsbeskrivningen, och om inget annat anges så gäller att åtminstone ett likgiltighetsuppsät måste kunna visas. De tre formerna av uppsät kan gälla en effekt (en följd) av en gärning, t.ex. någons död, men det kan också vara handlandet i sig eller vissa faktiska omständigheter, som ska "täckas" av uppsät.

Avsikt uppsät föreligger om den brottsliga effekten är åsyftad av gärningspersonen. Så kan vara fallet om effekten är själva målet med gärningen, eller

---

1Även underlåtenhet att göra något visst kan vara straffbart och resonemanget gäller också dessa fall.

2(Bergström et al., 2002)

3I kombination med begreppet oaksamhet, se not 7.

4Se exempelvis (Strahl, 1976, s. 89).
om effekten utgör ett medel för att uppnå målet. Om ett avsiktsuppsåt måste vi-sas uttrycks det ofta i lagtexten med formuleringar som avsiktlig eller i syfte att.  

Om den brottsliga effekten inte är åsyftad i sig, utan istället utgör en nödvändig biverkan av gärningen, föreligger ett insiktsuppsåt. Handlandet kan syfta till en annan brottslig effekt, till något som inte är brottsligt eller inte till något särskilt alls. Ett exempel som ofta anges är att någon avser att döda en viss person med en sprängladdning, men inser att detta kommer att skada eller döda andra människor runtomkring. Målet med gärningen, en persons död, omfattas då av ett avsiktsuppsåt medan skadorna på andra kan omfattas av insiktsuppsåt.  


Sammanfattningsvis ställer avsiktsuppsåtet upp ett krav på kännedom om eller insikt i sakförhållandena samt en viljeinriktning hos gärningspersonen. För insiktsuppsåt krävs att gärningspersonen insåg eller var övertygad om att effekten skulle inträda som en biverkan, en säker följd av, eller var nödvändigt förbunden med gärningen. Vid insiktsuppsåt är det alltså gärningspersonens föreställning om följen och omständigheterna kring gärningen som utgör hela grunden för bedömningen. När det gäller likgiltighetsuppsåtet får gärningspersonens föreställning av risken för en viss följd stor betydelse. Har gärningspersonen insett att risken var mycket hög och trots det genomfört gärningen är det i praktiken ofta tillräckligt för att det ska anses visat att personen varit likgiltig på ett sådant sätt att uppsåt föreligger.  

Uppsåtet ska "täcka" gärningsrekvisiten enligt den så kallade "täckningsprincipen". I ett fall av mord ska gärningspersonen ha uppsåt till, och i och med det vissa föreställningar om, både handlingen t.ex. ett knivhugg, omständigheten att det är en annan människa som tar emot hugget samt effekten att personen dör.  

---

5 Se t.ex. (Asp et al., 2013, s. 287).  
6 Se t.ex. BrB kommentar 1:2 s. 5.  
7 Detta är vad bedömningen gäller i exemplet ovan med jägaren som skjuter en älg, varpå skottet går rakt igenom älgén och träffar en skidåkare, som oturligt nog åker förbi i precis fel ögonblick. Det är inte, som man kan tro, ett läsöket skolexempel utan ett verkligt fall, där jägaren frikändes av en oenig hovrätt. Resonemanget om fôreställningar gäller i delar även oaktamhetsbegreppet, men vi behandlar inte det närmare i denna framställning, främst av utrymmesskäl.  
8 Se t.ex. (Lejonhuvud & Wennberg, 2008, s. 60), och NJA 2004 s. 176.  
9 (Asp et al., 2013, s. 290f)  
10 Se mer i t.ex. (Asp & Ulväng, 2009).
3 DEN STRAFFRÄTTSLIGA FÖRESTÄLLNINGEN OM OCH
BEDÖMNINGEN AV FÖRESTÄLLNINGAR

Gemensamt för de olika formerna av uppsåt är att gärningspersonen måste ha en någorlunda korrekt föreställning om gärningen. Nils Jareborg kallar denna komponent för en kognitiv inställning, och i praxis formuleras ofta ett krav på tillräcklig medvetenhet.

Uppsåtet har tidigare uppfattats som ett uttryck för gärningspersonens samhällsfarliga vilja. Ivar Agge skriver 1961 att det för uppsåt krävs att ”gärningsmannen med sitt psyke omfattat det objektiva skeendet på ett sådant sätt, att detta kan sägas utgöra ett uttryck för en samhällsskadlig vilja eller en tydlig likgiltighet för de krav som lagen ställer på medborgarnas handlande i det berörda hänseendet.”11 I senare juridisk doktrin har man tagit avstånd från ett viljebaserat uppsåtsbegrepp och poängterar istället handling och handlingsskäl i enlighet med en mindre libertarianisk syn på viljefrihet.12 En uppsåtlig gärning beskrivs idag ofta som ”en medveten, kontrollerad handling”.13

I linje med att fokus flyttas från viljan till handlingen har uppsåtsbedömningen mycket kommit att handla om den tilltalades medvetenhet. I lagkommentarer till Brotsbalken beskrivs en första förutsättning för att ansvar ska kunna utdömas, att gärningspersonen befunnit sig i ett ”medvetet tillstånd”.14 Exempel som anges på tillstånd som inte är medvetna är medvetenslöst, koma, sömn och omtöckningstillstånd t.ex. vid uppvaknande ur narkos. I rättspraxis finns exempel på fall där gärningspersonen befunnit sig i ett tillstånd av sänkt medvetenhet till följd av blodsockersänkning vid diabetes (RH 1993:80). Madeleine Lejonhufvud och Suzanne Wennberg skriver i Straffansvar att reflexrörelser och rörelser under sömn, medvetenslöst eller hypnos inte inbegrips under begreppet uppsåtlig handling15 då rörelserna utförs omedvetet eller ”utanför jagets kontroll”.16

I kommentaren till Brotsbalken påpekas att detta krav på medvetenhet inte ska tolkas så att gärningspersonen aktivt måste tänka på de omständigheter som ska vara täcka av uppsåt. Det räcker att personen kan ”erinra sig dem om han tänker efter”. Ivar Strahl beskriver denna grundläggande föreställning om gärningen som ett krav på viss kännedom. Han ger ett exempel med en person

---

11(Akke, 1961, s. 251)
12(Asp et al., 2010, s. 304)
13(Asp et al., 2013, s. 285)
14Kommentar till BrB 1:2 s. 4
15Man kan här fråga sig om en sådan kroppsrörelse som företas utanför jagets kontroll överhuvudtaget är att betrakta som en handling eller en gärning i juridisk mening. I teorin är gärningen en del av gärningsrekvisiten, och det är först då det föreligger en gärning som vi kan tala om huruvida den är täckt av uppsåt eller inte. En rörelse som är en ren reflex betraktas inte som en gärning. Denna åtskillnad är tydlig i den juridiska doktrinen, där det också har poängterats att det är viktigt att hålla isär begreppen, men skillnaden förefaller inte upprätthållas i praktiken i någon större utsträckning. (Agge, 1961, s. 267f, Jareborg, 1994, s. 256, Jareborg 1969).
16(Lejonhuvud & Wennberg, 2008, s. 55)
som går. Att själva gåendet är uppsåtligt följer av att den gående har omedelbar kännedom om att hon går. Personen kanske inte direkt tänker på att hon går, men hon är medveten om det. Strahl preciserar begreppet medvetenhet så att en människa får anses vara medveten om "allt det hon en gång lärt känna eller tänkt på och inte sedermera glömt".\textsuperscript{17} Strahl markerar tydligt mot ett viljebaserat uppsåtsbegrepp: "Någon föresats eller något slags beslut kräves emellertid inte. Någon viljeaktig behöver inte påvisas. Den som slår annan är att döma för misshandel utan att det behöver påvisas mer än att han är medveten om detta."\textsuperscript{18}


\textsuperscript{17}(Strahl, 1976, s. 112)  
\textsuperscript{18}(Strahl, 1976, s. 110)  
\textsuperscript{19}(Jareborg, 2008, s. 2f)  
\textsuperscript{20}(Strahl, 1976, s. 113)  
\textsuperscript{21}(Jareborg, 2008, s. 11, Asp et al., 2013, s. 289, NJA 1977 s. 630)
det finns en risk för en viss följd, den kognitiva inställningen till följd en beskrivs i Kriminalrättens grunder som lägst en misstanke om att följen kan komma att inträffa.22

I de flesta fall är bedömningen av en persons föreställningar om gärningens förhållandevis oproblematisk. En maffiaboss som skjuter ihjäl ett vittne har högst sannolikt både en viljemässig inställning till och någorlunda korrekta föreställningar om gärningen, omständigheterna och effekterna. En rejält överförfriskad person som knivhugger sin dryckeskamrat har kanske inte lika tydliga eller korrekta föreställningar om vad som äger rum. Domstolen har då att analysera omständigheterna i fallet för att hitta ledträd till om gärningspersonen var tillräckligt medveten om händelseförloppet för att kunna hållas ansvarig. Resonemanget i fall som detta torde ofta koka ner till ett konstaterande att den som hugger en person med kniv har uppsåt till åtminstone skadan, de flesta av oss delar föreställningen att ett knivhugg kan orsaka allvarlig kroppsskada. Även föreställningen att det vid knivhugg mot vissa delar av kroppen finns en betydande risk för att personen avlider förutsätts vara rätt allmänt spridd. Förinställningen om risken, i kombination med genomförandet av knivhugget kan i sig vara tillräckligt för att visa att gärningspersonen förhöll sig likgiltig till effekten (att personen dör) på ett sådant sätt som krävs för likgiltighetsuppsåt. Omständigheter som domstolen kan tillmatta betydelse vid bedömningen är t.ex. antalet knivhugg, placeringen av huggen och hur kraftiga de var.23


Enligt lag har vi rätt att försvara oss i en situation där vi är utsatta för ett överhängande brottsligt angrepp, vi kan då tillgripa våld i nödvärn för att rädda oss själva (under förutsättning att våldet inte är uppbergt oförsvaret). Samma rätt gäller i nödsituationer när fara hotar liv, hälsa, egendom eller något annat viktigt av rättsordningen skyddat intresse. Nöd- och nödvärntsättnings är två så kallade ansvarsfrihetsgrunder, som kan rättfärda ett beteende som i normala fall skulle bedömas som en uppsåtlig brottslig gärning. I vårt exempel föreligger ingen äkta nödvärnssituation, utan en inbildad sådan, så kallat putativt nödvärn.

22(Asp et al., 2013, s. 291)
För att ett uppsåtligt brott ska vara begången krävs att rekvisiten för otillåten gärning som ställs upp i brottsbeskrivningen är uppfyllda och att uppsåtet täcker dessa rekvisiter. Utöver det krävs också att det inte föreligger några omständigheter som befriar från ansvaret (ansvarsfrihetsgrunder). Här skiljer man mellan **rättfärdigande omständigheter**, som innebär att en annars otillåten gärning faktiskt är tillåten, t.ex. att dödandet skedde i en krigssituation, att slaget utdelades som en del i en boxningsmatch (samtycke) eller i en nödvärrsitu- tion, och **ursäktande omständigheter**, som innebär att gärningspersonen är ur- säktad trots att gärningen var otillåten t.ex. frivilligt tillbakaträdande (i vissa fall då gärningspersonen avbryter gärningen). De rättfärdigande omständigheterna tillsammans med gärningsrekvisiten utgör det som ibland kallas brottets ”objektiva” sida, medan kravet på uppsåt och de ursäktande omständigheterna tillhör den ”subjektiva” sidan. Som vi vet krävs att uppsåtet täcker rekvisiten för otillåten gärning. För att gärningen verkligen ska vara otillåten krävs i själva verket att hela den objektiva sidan täcks av uppsåtet, alltså gärningsrekvisiten samt **frånvaro av rättfärdigande omständigheter**. Med andra ord, tror gärnings- personen att en rättfärdigande omständighet, t.ex. en nödsituation, föreligger, så föreligger inte en fullständig täckning, inget uppsåt och därmed inget brott.24

4 STRAFFRÄTTSLIG BEDÖMNING AV VANFÖRESTÄLLNINGAR

Begreppet **föreställning** i straffrätten beskrivs således som en attityd hos subjek- tet när hon antar att något är fallet. En föreställning kan vara både medveten och dispositionell, den ska vara underbyggd av skäl och den kan föreligga hos subjektet med olika grad av övertygelse eller ”styrka”. En särskild sorts före- ställning som ibland hamnar i fokus vid en rättegång är s.k. vanföreställningar, vilka på avgörande punkter förefaller skilja sig från mer vardagliga felaktiga föreställningar.

En vanföreställning är per gångse psykiatrisk definition en fix idé om den yttre verkligheten som vidmakthållits trots att det finns obestridliga eller uppenbara bevis för motsatsen och trots att den strider mot vad flertalet finner rimligt att tro givet en specifik kulturell, religiös eller sub-kulturell kontext. Om vanför- reställningen innebär ett ”extremt värdeomdöme” kan den bara betraktas som vanföreställning om ”omdömet är så extremt att det trotsar all trovärdighet”.25 En vanföreställning skiljer sig alltså från andra föreställningar genom att per- sonen som hyser en sådan är ovillig eller snarare oförmögen att ompröva sin uppfattning oavsett vad andra tror och oavsett vad vi i allmänhet anser vara evidens mot något. Personer med vanföreställningar drar inga slutsatser från dessa föreställningar och förefaller inte heller bekymra sig över att de inte är

24(Asp & Ulväng, 2009, s. 265–273)
25(DSM-IV 2000, s. 765)
konsistenta med övriga trouppfattningar som personen själv kan ha. Vidare är de inte sällan även ”tröga” från ett beteendeperspektiv, subjektet tenderar att låta bli att agera i enlighet med sina vanföreställningar. Sammanfattningsvis kan sägas att en vanföreställning tenderar att hållas med en mycket hög grad av visshet, men utan goda skäl i en vanlig mening. De är ”fixa idéer” i den meningen att de hålls för sanna med en överdriven grad av övertygelse.

Är en psykotisk vanföreställning jämställd med en troföreställning vilken som helst ur ett straffrättsligt perspektiv? Det enkla svaret på frågan är ja. I teorin. Vad som är upphovet till en viss föreställning, eller vilken sorts föreställning det är, är egentligen juridiskt irrelevant. Är kravet på en kognitiv inställning till gärningen inte uppfyllt kan gärningspersonen inte hållas ansvarig för ett uppsåtligt brott, det gäller oavsett om anledningen är att gärningspersonen hade en felaktig uppfattning om gärningen och omständigheterna däromkring på grund av en psykisk störning eller om hon helt enkelt inte kunde föreställa sig t.ex. att ett knivhugg i läret skulle kunna leda till någons död. Om den räddhågse slår en termos i huvudet på en oskyldig medmänniska ska det, i princip, inte heller spela någon roll om han misstog en stålkam eller om han trodde att han höll på att bli uppsäten av ett mordiskt sjölejon.


26(Sims, 1995)
27(I NJA 2012 s. 45 konstaterar HD dock att det vid putativa nödvärssituationer bör krävas att de yttre förhållanden ger någon form av stöd för att den tilltalade har missuppfattat situationen.
Om det saknas en rimlig förklaring till varför den tilltalade skulle ha misstagit sig bör invändningen, enligt HD, normalt kunna lämnas utan avseende eller anses motbevisad.
28(Asp et al., 2013, s. 294 ff)
då gärningspersonen lider av en psykisk störning. Man kan emellertid ifrågasätta om t.ex. en insikt om en följd verkligen kan utläsas ur gärningspersonens beteende och när det gäller föreställningar som ligger långt från vardagliga erfarenheter blir det klart svårere att dra slutsatser om föreställningars innehåll utifrån beteendet. Det är för rätten lättare att förstå att en kam kan misstas för en kniv än att man tycker sig se ett mordiskt sjöljeon.

Man skulle kunna föreställa sig att föreställningar kring brottsituationens verkliga egenskaper som är så fjärran från verkligheten, som den att en människa är ett mordiskt sjöljeon, skulle kunna innebära att uppsåt saknas. Strahl skriver att man inte generellt kan hävda att ”sinnessjuka eller sinnsslöa” inte kan uppfylla kravet på uppsåt, men att sinnessjukdom kan innebära en ”så ofullständig eller oriktig föreställning om sakförhållandena” att uppsåt inte föreligger. Han drar dock, utifrån studier av praxis, slutsatsen att ”uppsåtsrevisitit kan anses uppfyllt ehuru den tilltalades uppfattnings- och omdömesförmåga var i ganska hög grad nedsatt.”

En princip som brukar anges när det gäller bedömning av uppsåt hos psykiskt störda lagöverträdare är att uppsåt inte föreligger om en person utan den psykiska störningen som hade samma föreställning om sakförhållandena inte skulle anses handla uppsåtligt. Det sägs att prövningen av uppsåtsfrågan i de flesta fallen torde ge till resultat att uppsåt föreligger. HD har dock poängterat att detta inte ska förstås så, att uppsåtsprövningen i de fall gärningspersonen hade en psykisk störning ska ske med mindre noggrannhet än i andra fall.

Vad menar man här egentligen? Vi tänker oss en ung man som lider av psykotiska vanföreställningar. Han befinner sig i vardagsrummet tillsammans med sin mamma och tittar på TV. Övertygelsen om att mamman i själva verket är en fruktansvärd demon, redo att angripa och döda honom, blir plötsligt helt överväldigande. Han går ut i köket, hämtar en kniv och oskadliggör sedan den onda demonen i förebyggande syfte. Vi ska nu tänka oss att en psykiskt frisk människa hade samma föreställning, alltså att det faktiskt verkligen var en demon som angreps. Skulle det vara att betrakta som ett uppsåtligt angrepp?

Vad man verkar vara ute efter här är att bedömningen av uppsåt gäller den brottsliga gärningen i ett snävare perspektiv. Även om sonen har en förvrängd verklighetsuppfattning inser han och är medveten om att han hugger. Man kan till och med säga att han har ett avsiktsuppsåt, då han avser att oskadliggöra sitt offer. Har det då ingen betydelse att föremålet för attacken i sonens föreställningsvärld är ett monster och inte hans mamma?

29 Per Ole Träskman diskuterar problematiken i ”Kan gärningspersonens uppsåt bevisas med hållpunkter i sinnevärlden?” (Träskman, 1985).
30 (Strahl, 1976, s. 78)
31 (Strahl, 1976, s. 94, vår kursivering)
32 T.ex. (Berg et al., 2011). Principen formulerades av Högsta domstolen i NJA 2004 s. 702.
33 NJA 2012 s. 45
34 Detta är ett exempel på ett verkligt fall, avgjort av Södertälje tingsrätt 2009, där gärningspersonen begick ett mord under påverkan av psykotiska vanföreställningar.
Vi nämnde tidigare täckningsprincipen, som uppställer ett krav på att gärningsrekvisiten ska vara täckta av uppsåt. Kan man verkligen säga att uppsåtet att döda ett monster täcker gärningen att döda mamman? Om vi delar upp gärningen i handling, följd och omständigheter kan vi se att handlingen, huggen, täcks av uppsåt. Detsamma gäller följden, död. När det gäller omständigheterna kan vi hävda att de inte är täckta, då sonen i sin psykotiska vanöreställningsvärld inte förstår att offret är en människa.35

Vid tillämpningen av täckningsprincipen anses det dock inte vara rimligt att ställa upp ett krav på fullständig täckning. Det räcker med en huvudsaklig överensstämmelse. En rekommendation som ges av Asp m.fl. är att vi ska anta att den verklighetsbild som gärningspersonen har är riktig och ställa oss frågan: skulle personen i så fall ha förövat samma rättsstridiga gärning som han/hon verkligen gjorde?36 Frågan vi slutligen ställs inför blir alltså huruvida man begär ett mord om man dödar ett monster. Lagen svarar inte på den frågan, kanske kan man tänka sig att ett monster är ett slags djur och om man dödar ett monster i sitt hem har man snarare haft uppsåt till djurplågeri, eller kanske jaktbrott?


35I det faktiska fallet dömdes sonen för att uppsåtligt ha dödat sin mamma. Domstolen menade att mamman i sonens ögon "såg ut som och talade till honom som en mor", varför han måste ha haft en tillräckligt korrekt föreställning om att det var hans mamma.

36(Asp et al., 2013, s. 64)
När han anklagar grannarna för detta kommer han typiskt inte bli det minsta påverkad av dessa nya fakta, utan fortsätta att tro att så är fallet med en oförminskad (känsla av) övertygelse.


Om vi då istället antar att vanöreställningar inte är förreställningar i den straffrättsliga meningen utan fantasier, torde det mentala tillståndets innehåll inte spela någon roll vid bedömning av uppsåt. Om mannen i exemplet ovan fantiserar om att mamma är ett monster samtidigt som han dödar henne (avsiktligt) verkar det rimligt att hålla honom ansvarig för sin gärning. Å andra sidan, om han inte själv förstår att det är en fantasi utan fel-identifierar sitt mentala tillstånd (fantasi) som en trosförreställning så tror han, från en subjektiv synvinkel, att mamma är ett monster, även om han ”egentligen” inte tror det, dvs ur en funktionell synvinkel. Hur ska rätten förhålla sig till detta?

5 FÖRESTÄLLNINGAR OCH TILLRÄKELIGHET

Svårigheten att identifiera vad en vanöreställning är förefaller ställa till med problem när man ska avgöra om det finns ett uppsåt som täcker gärningsrekvisiten i fall där gärningspersonen har vanöreställningar kopplade till gärningen. I de flesta länder i världen löser man detta problem genom att hantera vanöreställningar på ett helt annat sätt än andra förreställningar. Om en persons förreställning om gärningen och omständigheterna därormkring är så fjärran från vad som faktiskt är fallet så kan personen undantas från ansvar på grund av otillräknelighet.

37 (Currie & Ravenscroft, 2002)
38 (Stephens & Graham, 2006)
39 (Bortolotti, 2009)
Även svensk strafflagstiftning har omfattat ett tillräknelighetsrekvisit, men det avskaffades då Brotsbalken infördes 1965 efter omfattande kriminalpolitisk debatt. Istället för att fokusera på den brottsliga gärningen och ansvarsfrågan ville man flytta fokus till brottslingen och vad som skulle hända med denne efter domen. Den distinktion man tidigare gjort mellan sjuka och friska flyttades från ansvarsledet till påföljdsledet, där personer som led av en sinnessjukdom skulle beredas rättspsykiatrisk vård som påföljd för ett brott istället för fängelsestraff.\footnote{Upplysande redogörelser för denna reform ges av t.ex. (Lernestedt, 2013) och (Svennerlind, 2009).}

Denna ideologiska reform innebar att intresset att bereda vård för psykiskt störda lagöverträdare överordnades principen att ingen ska straffas som inte haft skuld till gärningen (skuldprincipen) eller haft möjlighet att hantera lagen (konformitetsprincipen). Överväganden gällande vilken brottspåföljd som skulle passa varje kategori av lagöverträdare betonades, och som Strahl uttrycker det: "Vid övervägandena har humanitära och praktiska synpunkter anlagts, inte filosofiska."\footnote{(Strahl, 1976, s. 77)}

Redan tidigt insåg man dock att reformen medförde en del problem. Kravet på att gärningspersonen skulle ha begått brottet uppsätligt kvarstod och skulle nu tillämpas lika för alla lagöverträdare, även psykiskt störda och tillfällig sinnesförvirrade. Psykiskt störda personer förväntades i hög utsträckning anses ha haft uppsät vilket skulle leda till att de dömdes till rättspsykiatrisk vårds. När det gällde tillfällig sinnesförvirrade, som inte vid domstillfället hade något vårdbehov, framgår av förarbetena att man förväntade sig att de skulle frikännas på grund av brist på uppsät.\footnote{Se t.ex. (Jareborg, 1986, s. 38) och (Lernestedt, 2013).}

Rättsutvecklingen har, kort uttryckt, lett till att vi idag har ett system för ansvarsbedömning som bygger på skuld- och konformitetsprincipen och ett påföljdssystem som bygger på en behandlingsideologi. Denna inbyggda intressekonflikt har inneburit att gärningspersonens föreställningar har bedömts på olika sätt beroende av vilken typ av sinnestillstånd personen var i vid gärningstillfället. I fall av tillfällig sinnesförvirring måste, enligt den juridiska doktrinen, en oskruven undantagsregel gälla. En grund för detta är uttalandet i förarbetena till brotsbalken och även uttalanden i senare förarbeten, en annan är att det helt enkelt skulle leda till orimliga konsekvenser om tillfällig sinnesförvirring inte friade från ansvar.\footnote{(Asp et al., 2013, s. 370f, Strahl, 1976, s. 95, SOU 2012:17, s. 545f)} Ett exempel kan vara en normalt mycket fredlig person som utan vetskap får i sig en drog, fullkomligt tappar besinnen och misshandlar sin granne med ett vedtrå. Här resonerar man så att personen kan ha haft uppsätt att misshandla grannen, som kanske hade sågat ner träd på hans mark, men att överreaktionen och uppsätet till gärningen helt berodde på en tillfällig sinnesförvirring som inte vallades av personen själv. Att döma personen för misshandel skulle vara orimligt och strider mot skuld- och konformitetsprincipen. Det är dock oklart om en sådan oskriven undantagsregel godtas och tillämpas i praktiken. Det finns oss veterligen inga exempel på fall där undantagsregeln

\footnote{40}{Upplysande redogörelser för denna reform ges av t.ex. (Lernestedt, 2013) och (Svennerlind, 2009).}

\footnote{41}{(Strahl, 1976, s. 77)}

\footnote{42}{Se t.ex. (Jareborg, 1986, s. 38) och (Lernestedt, 2013).}

\footnote{43}{(Asp et al., 2013, s. 370f, Strahl, 1976, s. 95, SOU 2012:17, s. 545f)}
Det verkar snarare vara så att det sker en anpassning av uppsåtsbedömningen, där den tillfälliga sinnesförvirrade inte anses vara i tillräcklig grad medveten för att hållas ansvarig.\footnote{I NJA 2001 s. 899 konstaterar HD att den tilltalade vid gärningen led av en icke självvallad tillfällig sinnesförvirring, någon ansvarsfrihetsregel tillämpas inte. Se också t.ex. Hovrättens diskussion i (RH 1995:150) där man konstaterar att tillfällig sinnesförvirring enligt nuvarande lagstiftning inte kan anses undanta från ansvar, om gärningen begåtts med uppsåt.}

När det gäller långvariga allvarliga psykiska störningar gäller dock ingen ansvarsfrihetsregel, här var ju tanken att sinnestillståndet, i enlighet med behandlingsideologin, istället skulle påverka påföljdbedömningen. För att kunna utdöma en påföljd krävs då att den som lider av en allvarlig psykisk störning anses ha begått ett uppsåttligt brott och inte befrias från ansvar.\footnote{(Asp et al., 2013, s. 371) samt (SOU 2012:17, s. 546f).}

Flera statliga utredningar, senast Psykiatrilagsutredningen, har föreslagit en återgång till ett system liknande det som föregick Brottsbalken och som finns i de flesta andra rättsystem, där en person som är 

\begin{itemize}
  \item en allvarlig psykisk störning,
  \item en tillfällig sinnesförvirring,
  \item en svår utvecklingsstörning eller
  \item ett allvarligt demenstillstånd
\end{itemize}

har saknat förmåga att förstå gärningens innebörd i den situation i vilken han eller hon befann sig eller att anpassa sitt handlande efter en sådan förståelse.

En återgång till en tillräknelighetslära skulle innebära att gärningspersonens föreställningar får en annan betydelse för ansvarsbedömningen. Här är det inte bara en föreställning om de specifika gärningsrekvisiten som är relevant, utan även en förståelse för gärningens innebörd i en vidare mening. En förståelse för gärningens innebörd i den situation som gärningspersonen befann sig, formulerar man i Psykiatrilagsutredningens betänkande *Psykiatrin och lagen – tvångsvård, straffansvar och samhällsskydd*.\footnote{(SOU 2012:17, s. 561f)}

I betänkandet anges en rad problem med dagens lagstiftning som skulle kunna åtgärdas med en tillräknelighetsreglering. Bland annat skriver man att ett
problem med den nuvarande regleringen är att "en person som vid gärningstillfället var akut psykotisk och vars uppfattning om sakförhållanden helt saknade förankring i verkligheten ändå kan vara straffrättsligt ansvarig."\textsuperscript{48} Ett exempel som anges är en människa som skadar någon, och förstår att hon skadar någon, men felaktigt har förstått att hon befinner sig i krig och att gärningen därför är ursäktad eller kanske till och med påbjuden. Man påtalar också som ett straffrättsligt problem att det finns en risk att domstolarna vid uppsättsbedömningen i mål med psykiskt störda lagöverträdare ställer upp ett för lågt krav för att uppsätta ansvar. I syfte att möjliggöra åtgärderna för vård och samhällsskydd. Det finns visst stöd för att det också är fallet.\textsuperscript{49}

Kravet på tillräknelighet ska enligt utredningen gälla gärningspersonens "förmåga att på ett rationellt sätt förstå vad gärningen innebär i den situation som han eller hon befinner sig i."\textsuperscript{50} Man talar om situationer där gärningspersonen haft en bristande realitetsvärdering (vanföreställningar eller förvirring) som varit avgörande för dennes förståelse av situationen, då gärningspersonen "inte kunnat relatera gärningen till dess faktiska sammanhang". Gärningspersonen har inte alls haft förmåga att skilja mellan vad som är rätt eller fel, eftersom personen saknat förmåga att relatera gärningen till dess faktiska sammanhang. Exempelvis kan omgivning och tidpunkt ha uppfattats på ett "annat sätt än det verkliga". Gärningspersonen har "befunnit sig i en annan verklighet". Detta vill utredningen tydligt skilja från de föreställningar som anses vara relevanta för bedömningen av uppsåt, nämligen föreställningar om "betydelsen och konsekvenserna av den straffbelagda gärningen som sådan."\textsuperscript{51}

6 BEHÖVS EN TILLRÄKENLIGHETSREGLERING FÖR EN TILLFREDSSTÄLLANDE BEHANDLING AV VANFÖRESTÄLLNINGAR?

Bedömningen av om en person ska dömas och straffas för en gärning är i högsta grad beroende av personens föreställningar om gärningen. När det gäller personer som är mentalt friska gäller bedömningen endast innehållet i föreställningen. Fanns en insikt om att offret skulle avlida? Trodde personen att kammen var en kniv? Om personen däremot lider av en allvarlig psykisk störning, blir det istället med dagens lagstiftning det mentala tillståndets natur och varaktighet som hamnar i fokus.

I förslagen om att övergå till en tillräkenlighetsreglering skiljer man mellan föreställningar relevanta för uppsåt och de som är relevanta för tillräknelighet.

\textsuperscript{48}(SOU 2012:17, s. 512)
\textsuperscript{49}(SOU 2012:17, s. 527f, Westin, 1985, Lernestedt, 2013)
\textsuperscript{50}(SOU 2012:17, s. 542)
\textsuperscript{51}Utredningens beskrivning av vad som avses men sådan bristande förmåga finns på s. 542–544 och s. 561–562.
I straffrättsteorin förklaras skillnaden i hanteringen av olika sorters föreställningar med att man skiljer mellan föreställningar om de relevanta gärningsrekvisitken för ett brott och föreställningar om gärningen i en "vidare bemärkelse". Nu frågar vi oss, är denna distinktion rimlig? Och är den egentligen nödvändig?

Utan att gå in i en omfattande diskussion om kriminaliseringens syften och uppsåtsbegreppets avgränsande funktion i förhållande till dessa syften, kan vi för enkelhetens skull utgå från att vi bör hålla en person straffrälsligt ansvarig om den relevanta gärningsrekvisiten för eftersom han eller hon begått och föreställningar om gärningen i en "vidare bemärkelse".

Utan att gå in i en omfattande diskussion om kriminaliseringens syften och uppsåtsbegreppets avgränsande funktion i förhållande till dessa syften, kan vi för enkelhetens skull utgå från att vi bör hålla en person straffrälsligt ansvarig om den relevanta gärningsrekvisiten för eftersom han eller hon begått och föreställningar om gärningen i en "vidare bemärkelse".

Nu frågar vi oss, är denna distinktion rimlig? Och är den egentligen nödvändig?

Utan att gå in i en omfattande diskussion om kriminaliseringens syften och uppsåtsbegreppets avgränsande funktion i förhållande till dessa syften, kan vi för enkelhetens skull utgå från att vi bör hålla en person straffrälsligt ansvarig om den relevanta gärningsrekvisiten för eftersom han eller hon begått.

Dagens system för tillskrivande av straffansvar genom uppsåtsbedömningen fängar upp de fall där den tilltalades föreställning om gärningen i en striktare mening avviker från verkligheten. Vi håller inte personen som "bara" trodde att hugget skulle orsaka skada ansvarig för mord. Täckningsprincipen ger oss också möjlighet att frikänna den som trodde sig vara under angrepp, även om den föreställningen skulle råka vara en psykotisk vanföreställning. Om gärningspersonens föreställningar någorlunda överensstämmen med verkligheten när det gäller gärningsrekvisiten i en striktare mening, har vi däremot inte möjlighet att undanta personen från ansvar. Den oskrivna undantagsregel som gäller tillfällig sinnesförvirring kan möjligen ursäkta den som i strikt mening haft ett uppsåt, men där uppsåtet betraktas som sprunget ur sinnesförvirringen. I fall då sinnesförvirringen inte betraktas som tillfällig finns dock inte den möjligheten.

En övergång till ett tillräcknelighetssystem skulle kunna möjliggöra en mer konsekvent bedömning av gärningspersonens föreställningar. Låt oss undersöka i vilka situationer detta begrepp behövs. Om vi betraktar gärningspersonens föreställningar som trosföreställningar i en straffrättslig mening, när uppstår egentligen problemet att gärningspersonens föreställningar om gärningen uppfyller kravet på uppsåt men inte kravet på tillräcknelighet? Låt oss titta närmare på de exempel som Psykiatrilagsutredningen hänvisar till, utredningen skriver:

[E]n del fall som inte bör leda till ansvar enligt skuldprincipen är sådana där det normalt anses finnas uppsåt (eller oaksamhet). Som exempel kan nämnas ett fall där en person felaktigt trodde sig delta i krig. Gärningsmannen var medveten om att han eller hon sköt på människor och hade därför uppsåt till just detta, men trodde som

52 En fråga för sig är då gärningspersonen inte uppfyller dessa förutsättningar på grund av eget vållande. HD har nyligen kommit med ett par klargörande domar där man fastslår att kravet på uppsåt gäller i samma utsträckning för alla. Den som saknar uppsåt, men själv vållat detta genom t.ex. droger eller alkohol, ska inte dömas. En mer ingående diskussion om detta är både relevant och intressant men utelämnas här av utrymmeskäl.

53 (SOU 2012:17, del 2, s. 525)
sagt att det skedde i en krigssituation; en föreställning som saknade all föranckring i verkligheten.

Det kan vara så att en person som tror att hon befinner sig i krig normalt anses ha haft uppsåt, men om vi tillämpar uppsåtsbegreppet och täckningsprincipen så som den beskrivs i den juridiska doktrinen kan vi nå en annan slutsats. Förutom nöd och nödvärn anges i Brotsbalkens 24 kapitel samtycke, rättlig befogenhet och förmanes befällning som allmänna ansvarsfrihetsgrunder. Den sistnämnda innebär att en gärning som någon begår på order av den under vars lydnad han står inte ska medföra ansvar. Den möjlighet att frikänna någon som befann sig i en inbillad, så kallad, putativ nödsituation borde rimligtvis även gälla övriga ansvarsfrihetsgrunder. "Putativ krigssituation" låter kanske långsökt, men en person som tror sig agera på order under de förutsättningar som ställs upp i lag, har samma rättsliga ställning som den som tror sig agera i nödvärn.  

Så långt förefaller vi klara oss med den nuvarande lagstiftningen.

En förutsättning för att befria någon från ansvar i en putativsituatation, oavsett om det handlar om nödvärn, order eller samtycke (t.ex. putativ boxningsmatch), är dock att personen agerar inom ramen för den rättfärjandares omständigheten. Det är inte tillåtet att använda hur mycket våld som helst i en nödvärnssituation, och att argumentera för ett frikännande baserat på föreställningen att den var tillåtet att starta ett eget krig mot t.ex. en granne, kan alltså betraktas som förhållandevis utsiktslös.

Låt oss testa ett annat av Psykiatrilagsutredningens exempel:

Ett annat fall kan vara, att någon attackerar en annan person i tron att denne är besatt av djävulen. Uppsåt att attackera en annan människa finns ju i en sådan situation, men gärningsmannens egen föreställning om gärningen och den situation i vilken den begås skiljer sig så markant från vad som rent faktiskt är fallet att det knappast ter sig rimligt att utkräva något straffrättsligt ansvar för denna.

På vilka grunder anser man att denna person ska frias från ansvar? År det för att angerpeppet på något sätt är berättigat? Kan det vara så att personen uppfattar situationen så att den som är besatt av djävulen befinner sig i en nödsituation? Eller att hon själv är hotad av djävulen och kan tillgripa nödvärn? I så fall handlar det återigen om en rättfärjandig omständighet. Om så inte är fallet och personen t.ex. avskyr den besatta och passar på att attackera denne av någon annan anledning är det då rimligt att personen ska frikänna? Låt oss se på ytterligare ett exempel som lyfts fram av Psykansvarskommittén, i en tidigare utredning, för att illustrera skillnaden mellan föreställningar som är relevanta för bedömningen av uppsåt respektive bedömningen av tillräknelighet:

54Putativ krigssituation har oss veterligen aldrig prövats i praktiken, men det finns exempel på fall där putativt samtycke och putativ laga befogenhet har prövats i domstol.
55(SOU 2012:17, del 2, s. 525)
56Även Psykansvarskommittén föreslog införandet av ett tillräknelighetsrekvisit (SOU 2002:3), ett
Ansvar för hälleri kräver att gärningsrekrivisitet "frånhänt annan genom brott" är täckt av uppsåt. Det innebär att gärningspersonen måste inse, eller vara medveten om, att det hon befattar sig med, t.ex. en cykel som hon förvarar på uppdrag av någon, är "frånhänt annan genom brott". I det sammanhanget spelar det för uppsåts-prövningen ingen roll om hon tar hand om cykeln i tron att hon agerar på uppdrag av marsianer som stulit denna från människorna. Denna verklighetsfrämmande föreställning kan däremot ha betydelse för prövningen av personens tillräknelighet.

Personen i exemplet förvarar en stulen cykel och förstår att hon gör det och har följaktligen uppsåt. Vidare har hon vanföreställningar om marsianer och exemplet ovan föreslår att hon skulle kunna bedömas som otillräknelig. Det är emellertid inte uppenbart på vilket vis det faktum att hon tror att hon agerar på uppdrag av marsianer ska vara ursäktande. Det finns inget i exemplet som tyder på t.ex. en hotsituation.


De flesta formuleringar av tillräknelighetsrekrivisit inbegriper att den tilltalade ska förstå att gärningen är fel för att kunna hållas ansvarig för sina gärningar. Det är oklart om en förmåga att skilja mellan rätt och fel innefattas i formuleringen "förmåga att förstå gärningens innebörden" som framförs i lagförslaget.

57 Se t.ex. M’Naghten-kriterierna som används ibland annat i England. De säger (i översättning): ”för att etablera ett försvar på grund av sinnessjukdom måste det klart bevisas att den åtalade vid tiden för brottslasten agerade under en så pass allvarlig förnuftsnedläggning på grund av sinnessjukdom att han inte insåg gärningens natur eller kvalitet och om han insåg det, att han inte insåg att det han gjorde var fel.”
och det finns kanske anledning att specificera om så är fallet. Definitionen av vanöreställningar från DSM (se sidan 29) anger att vanöreställningar inte endast handlar om vad som finns och inte finns, även värdeomöden kan utgöra vanöreställningar om de är så extrema "att de trotsar all trovärdighet". Det blir dock problematiskt för domstolen att avgöra huruvida ett visst omdöme dels är en vanöreställning, dels trotsar all trovärdighet. Rätten att döda monster är ju ex en norm som ofta förmedlas i spelfilmer och kan därmed kanske sägas ha en viss trovärdighet?

7 AVSLUTNING

Bedömningar av uppsåt i samband med vanöreställningar verkar kunna göras på ett par olika sätt. I det nuvarande svenska systemet förefaller följande varianter vara möjliga: Domstolen kan behandla vanöreställningen som ett globalt tillstånd som påverkar gärningspersonens medvetenhet och konstatera att personen hade en så verklighetsfrämmande verklighetsuppfattning att kravet på medvetenhet helt enkelt inte är uppfyllt, vilket leder till att personen frikänns på grund av bristande uppsåt. Att vanöreställningen betraktas främst som ett symptom på en allvarlig psykisk störning kan emellertid också leda till att domstolen tillmärker storningens betydelse först vid påföljdsvallet och istället dömer till ansvar efter att ha konstaterat att vanöreställningarna inte var så omfattande att personen inte var i tillräcklig grad medveten om sitt handlande. Betraktas vanöreställningen istället som en förställning i straffrättslig bemärkelse kan domstolen komma till slutsatsen att personen hade tillräcklig kännedom eller insikt om t.ex. huggandet av mamman, trots att hon uppfovjades som ett monstor, och hålla personen ansvarig. Man kan dock också tänka sig att domstolen tillämpar täckningsprincipen fullt ut och konstaterar att vanöreställningarna inte är tillräckligt i överensstämmele med gärningsrekvisiten för att en uppsåtlig gärning ska ha begåtts, varpå personen frikänns.

De olika sättet att resonera kan leda till att fall som förefaller vara ganska lika, när det gäller både personens sinnestillstånd och personens förståelningar om gärningen, bedöms på olika sätt. Här följer ett par exempel från den rättsliga praktiken. I ett fall från 1994 led den tilltalade av en manodepressiv psykos och var bland annat övertygad om att han var messias och odödlig. Han åtalades för bland annat vårdslöshet i trafik men frikändes.58 Enligt egen uppgift var han medveten om att han körde bil, han trodde dock att de mötande bilarna var flygplan som landade. Ett liknande fall är vansinneskörningen i Gamla Stan i Stockholm 200359 då en man som led av paranoid schizofreni körde in på en gågata med resultatet att två personer dog och flera skadades. Även denna man uppgav att han var medveten om att han satt i bilen men också att han var övertygad om att bilen styrdes med fjärrkontroll av en fiende till honom som

58(RH 1994:78)
59(Svea hovräts dom i mål B 6504-03)
kallades Fugi. Både tingsrätt och hovrätt dömde i det här fallet för mord och försök till mord. I ett fall från 1985 åtalades en man för försök till drap efter att ha huggit sin fru med kniv. Mannen led av en djup manodepressiv psykos och båda instanser friade på grund av bristande uppsåt till följd av att den tilltalades psykotiska tankevärd var så förvirrad.60 I ett till synes liknande fall från 2010 åtalades en man som inte tidigare varit våldsamt och plötsligt och utan särskild anledning knivhögg sin partner.61 Mannen led av psykotiskt syndrom med paranoida vanöställningar, han frikändes av tingsrätten på grund av bristande uppsåt men dömdes av hovrätten för försök till mord.62

En anledning till att utfallet varierar kan naturligtvis vara att vanöställningar anses kunna påverka gärningspersonens medvetenhet i olika utsträckning, men det tycks dock finnas skillnader i bedömningen som bäst förklaras med att rätten förhåller sig på olika sätt till frågan om huruvida vanöställningar är att bedöma som öreställningar i straffrättsligt mening.


60(RH 1985:62)
61(Göta hovrätts dom i mål B 2256-10)
Det finns helt säkert ytterligare tänkbara alternativ för hur vanföreställningar kan hanteras vid bedömning av straffansvar och ett flertal aspekter av problematiken som inte har behandlats här. Här måste vi trots detta dessvärre avsluta denna text. Vi kan trösta den otillfredsställda läsaren med att vi definitivt kommer att återkomma i frågan.

REFERENSER


SOU 2012:17. Psykiatrin och lagen – tvångsvård, straffansvar och samhällsskydd.


RÄTTSFALL

Hovrätten över Skåne och Blekinges dom i mål B 2100-02
Svea hovrätts dom i mål B 6504-03
Svea hovrätts dom i mål B 6660-07
Göta hovrätts dom i mål B 2256-10
Hovrätten för Övre Norrlands dom i mål B 504-12

NJA 1976 s. 183
NJA 1977 s. 630
NJA 1995 s. 48
NJA 2001 s. 899
NJA 2004 s. 176
NJA 2004 s. 702

NJA 2012 s. 45
RH 1985:62
RH 1993:80
RH 1995:150
RH 1999:139
RH 2009:11
TWO CONSEQUENCES OF KRIPKE’S LEMMA

Rasmus Blanck

1 INTRODUCTION

There are several results on independent formulae in the literature. These formulae are “flexible” in the sense that their extensions as sets are left undetermined by the formal system. Prominent references include, in order of appearance, e.g. (Mostowski, 1961), (Scott, 1962), (Visser, 1980), (Montagna, 1982), and most recently, (Woodin, 2011). (Lindström, 1984) gives a short exposé of, and a general proof for, some of these results. Different terminology has been used by different authors, e.g. flexible predicates, flexible formulae, free formulae, and independent formulae. For the purposes of this note, we will stick to the term flexible formulae.

The arguably first occurrence of a flexible formula is in Kripke’s “Flexible” predicates of formal number theory (Kripke, 1962).1 Kripke’s paper is very short,
only four pages, and its main technical content is an innocent-looking lemma. We will not delve into the technicalities and history of flexible formulae here, but rather use Kripke’s lemma as the main tool for proving two theorems from quite separate areas of mathematical logic. The first is a new result, which is a minor variation on the flexible formula-theme, whereas the second is a well-known theorem of Chaitin’s. First, we state some preliminaries.

Throughout the paper, let T be some p.r., consistent extension of \( I \), and fix some Gödel numbering of syntactical objects. \( \{ \phi \} \) is the (numeral representing the) Gödel number of \( \phi \), and \( n \) is the numeral representing the natural number \( n \). \( \text{Pr}_T(x) \) is a \( \Sigma_1 \)-provability predicate for T, satisfying Löb’s derivability conditions. Proofs of the following well-known facts can be found in e.g. (Smoryński, 1985).

**Fact 1** (Formalised \( \Sigma_1 \)-completeness). For each \( \sigma(x_0, \ldots, x_n) \in \Sigma_1 \),

\[
 T \vdash \forall x_0, \ldots, x_n (\sigma(x_0, \ldots, x_n) \rightarrow \text{Pr}_T(\{ \gamma \rightarrow \sigma(x_0, \ldots, x_n) \})).
\]

**Fact 2** (Existence of a \( \Sigma_1 \)-satisfaction predicate). Given \( n > 0 \), there is a \( \Sigma_1 \)-formula \( \text{Sat}_{\Sigma_1}^n \) such that for all \( \Sigma_1 \)-formulae \( \phi(x_0, \ldots, x_{n-1}) \) with exactly the variables \( x_0, \ldots, x_{n-1} \) free,

\[
 T \vdash \forall x_0, \ldots, x_{n-1} (\phi(x_0, \ldots, x_{n-1}) \leftrightarrow \text{Sat}_{\Sigma_1}^n(\gamma \phi, x_0, \ldots, x_{n-1})).
\]

We will omit the superscript \( n \), as the arity will always be clear from the context.

**Fact 3** (The selection theorem). For each \( \Sigma_1 \)-formula \( \phi(x_0, \ldots, x_n) \), there is a \( \Sigma_1 \)-formula \( \text{Sel}\{\phi\} \) with exactly the same free variables, such that

1. \( T \vdash \forall x_0, \ldots, x_n (\text{Sel}\{\phi\}(x_0, \ldots, x_n) \rightarrow \phi(x_0, \ldots, x_n)) \),
2. \( T \vdash \forall x_0, \ldots, x_n (\text{Sel}\{\phi\}(x_0, \ldots, x_n) \land \text{Sel}\{\phi\}(x_0, \ldots, x_{n-1}, y) \rightarrow x_n = y) \),
3. \( T \vdash \forall x_0, \ldots, x_{n-1} (\exists x_n \phi(x_0, \ldots, x_n) \rightarrow \exists x_n \text{Sel}\{\phi\}(x_0, \ldots, x_n)) \).

To be able to talk about partial recursive functions in a language that does not contain functions symbols other than \( +, \times \) and \( S \), we will refer to functions by formulae defining their graphs. Hence, let the \( n \)-ary partial recursive function \( \phi_i \) with index \( i \) be the function with graph defined by

\[
 \text{Sel}\{\text{Sat}_{\Sigma_1}^i\}(i, x_0, \ldots, x_n).
\]

Let \( R(x, y, z) \) be the formula \( \text{Sel}\{\text{Sat}_{\Sigma_1}^i\}(x, y, z) \). As it happens, this formula strongly represents (a notational version of) Kleene’s universal function (for unary partial recursive functions) in T. It follows that, for each \( m, n, k \in \omega \),

\[
 T \vdash R(m, n, k) \land \exists z R(m, n, z) \text{ iff } \Phi(m, n) = k \text{ iff } \phi(n) = k \text{ and } m = \gamma \phi^\gamma.
\]

Now for the main lemma, which is Lemma 1 of (Kripke, 1962). We give what is essentially the original proof.
Kripke’s lemma. There is an index $e$ such that for each $k$, $T + R(e, e, k) \land \exists! z R(e, e, z)$ is consistent.

Proof. Suppose that $T$ is consistent. Define a partial recursive function $f(n)$ by the stipulation that $f(n) = k$ iff $T \vdash (R(n, n, k) \land \exists! z R(n, n, z))$.

If more than one formula of the form $(*)$ is provable, choose the one with the least Gödel number. Let $e = \lceil f \rceil$, and suppose that $(*)$ is provable for $n = e$. Then $f(e) = k$ is defined, so by universality, $f(e) = \Phi(e, e) = k$. Hence $T \vdash R(e, e, k) \land \exists! z R(e, e, z)$ by strong definability, which contradicts the consistency of $T$.

2 FLEXIBLE FORMULAE

We will state and prove a rather straightforward proto-result on the existence of flexible formulae, which is essentially Kripke’s Theorem 1.

Theorem 2.1. There is a $\Sigma_1$-formula $\gamma(x)$ such that for each $\Sigma_1$-formula $\delta(x)$, the theory $T + \forall x (\gamma(x) \leftrightarrow \delta(x))$ is consistent.

Proof. Let, for the remainder of this note, $\gamma(x)$ be the formula

$$\exists z (R(e, e, z) \land \text{Sat}_{\Sigma_1}(z, x))$$

and let $\delta(x)$ be any $\Sigma_1$-formula. By Kripke’s lemma, the theory $T + R(e, e, \lceil \delta \rceil) \land \exists! z R(e, e, z)$ is consistent, and in this theory $\gamma(x)$ is equivalent to $\text{Sat}_{\Sigma_1}(\lceil \delta \rceil, x)$. It follows that the theory $T + \forall x (\gamma(x) \leftrightarrow \delta(x))$ is also consistent.

Quite a few generalisations of Theorem 2.1 are available. For example, the same argument can be used to prove that there is a $\Sigma_n$- ($\Pi_n$-) formula that can mimic every other $\Sigma_n$- ($\Pi_n$-) formula, for all $n \geq 1$. Mostowski gives another kind of generalisation, constructing formulae that are flexible (in a slightly different sense) over certain kinds of r.e. sequences of theories. Details of these generalisations, and others, can be found in Chapter 2 of Lindström (2003). Yet another kind of generalisation is due to Hugh Woodin (2011), who has recently proved that a weaker result holds uniformly for countable models of PA. We present it here in a different wording.

Theorem 2.2. There is a $\Sigma_1$-formula $\psi(x)$ such that

1. $\mathbb{N} \models \forall x \lnot \psi(x)$,
2. For all countable models $\mathcal{M}$ of PA, and for all $\delta(x)$ such that $\mathcal{M} \models \psi(x) \rightarrow \delta(x)$ and the $\mathcal{M}$-extension of $\delta(x)$ is $\mathcal{M}$-finite, there is an end-extension $\mathcal{N} \models \text{PA}$ of $\mathcal{M}$ such that $\mathcal{N} \models \forall x (\psi(x) \leftrightarrow \delta(x))$. 

47
An optimal extension of this kind of result would of course be one that holds uniformly for all models of $T$, and for any r.e. set. Recently, Ali Enayat and Volodya Shavrukov have made a step in this direction, by proving that the countability restriction can be removed, if the flexibility is still restricted to finite sets. We will prove an intermediate version which holds for extensions of $\Sigma_1$-formulae, but only for models of $T + \text{Con}(T)$.

**Theorem 2.3.** The formula $\gamma(x)$ of Theorem 2.1 is such that for all models $M$ of $T + \text{Con}(T)$,

1. $M \models \forall x \neg \gamma(x)$, and

2. for each $\Sigma_1$-formula $\delta(x)$, there is an end-extension $N \models T$ of $M$ such that $N \models \forall x (\gamma(x) \leftrightarrow \delta(x))$.

The main effort of this section is to show that a version of Kripke’s Lemma is formalisable in $T$. Let $\phi(x,y) = \text{Pr}_T(\neg \text{Sat}_{\Sigma_1}(x, y))$, and let $e = \neg \phi$. The following Lemmas 2.4 and 2.5 will establish, respectively, the properties of $e$ corresponding to the claims made in parts 1 and 2 of Theorem 2.3.

**Lemma 2.4.** $T \vdash \forall y(\text{Con}(T) \rightarrow \neg R(e, e, y))$.

**Proof.** We argue by contraposition. By definition of $R$,

$$T \vdash \forall y(R(e, e, y) \leftrightarrow \text{Sel}(\text{Sat}_{\Sigma_1})(e, e, y)) \quad (1)$$

which by Fact 3 gives

$$T \vdash \forall y(R(e, e, y) \rightarrow \text{Sat}_{\Sigma_1}(e, e, y)) \quad (2)$$

and by Fact 2,

$$T \vdash \forall y(R(e, e, y) \rightarrow \phi(e, y)) \quad (3)$$

so by construction of $\phi$,

$$T \vdash \forall y(R(e, e, y) \rightarrow \text{Pr}_T(\neg \text{Sat}_{\Sigma_1}(e, e, y))). \quad (4)$$

But by (2), and Fact 1,

$$T \vdash \forall y(R(e, e, y) \rightarrow \text{Pr}_T(\neg \text{Sat}_{\Sigma_1}(e, e, y))). \quad (5)$$

so (4) and (5) together with the derivability conditions give

$$T \vdash \forall y(R(e, e, y) \rightarrow \neg \text{Con}(T)). \quad (6)$$

\[ \square \]
Lemma 2.5. \( T \vdash \forall y (\text{Con}(T) \rightarrow \text{Con}(T + R(e, e, y))) \).

Proof. Again, we argue by contraposition. Since \( T \vdash \text{I}_1 \),
\[
T \vdash \forall y (-\text{Con}(T + R(e, e, y)) \leftrightarrow \text{Pr}_T (\neg R(e, e, y))) \tag{7}
\]
and by Facts 2 and 3,
\[
T \vdash \forall y (-\text{Con}(T + R(e, e, y)) \rightarrow \text{Pr}_T (\neg \text{Sat}_{\Sigma_1}(e, e, y))) \tag{8}
\]
The definition of \( \phi \), together with (8) give
\[
T \vdash \forall y (-\text{Con}(T + R(e, e, y)) \rightarrow \phi(e, y)) \tag{9}
\]
which by Fact 2 gives
\[
T \vdash \forall y (-\text{Con}(T + R(e, e, y)) \rightarrow \text{Sat}_{\Sigma_1}(e, e, y)) \tag{10}
\]
so by Fact 1 we get
\[
T \vdash \forall y (-\text{Con}(T + R(e, e, y)) \rightarrow \text{Pr}_T (\neg \text{Sat}_{\Sigma_1}(e, e, y))) \tag{11}
\]
whence (8) and (11) together with the derivability conditions give
\[
T \vdash \forall y (-\text{Con}(T + R(e, e, y)) \rightarrow \neg \text{Con}(T)). \tag{12}
\]

Proof of Theorem 2.3. Suppose that \( M \models T + \text{Con}(T) \), and recall that \( \gamma(x) \) is the formula
\[
\exists z (R(e, e, z) \land \text{Sat}_{\Sigma_1}(z, x)).
\]
(1) By Lemma 2.4, \( M \models \forall y (\neg R(e, e, y)) \), so \( M \models \forall y \neg \gamma(x) \).

(2) Let \( \delta(x) \) be any \( \Sigma_1 \)-formula. By Lemma 2.5, we have \( M \models \text{Con}(T + R(e, e, \gamma \delta)) \).
By the Low Arithmetised Completeness Theorem, there is some subset \( A \) of \( M \), such that \( M \) satisfies “\( A \) is a consistent completion of \( T + R(e, e, \gamma \delta) \)”.
Moreover, the expanded model \((M, A)\) satisfies \( \Sigma_1(A) \), i.e., induction holds for all \( \Sigma_1 \)-formulae of the language obtained by augmenting the language of arithmetic with a new predicate \( A \) (interpreted by \( A \)). The desired model \( N \) can then be read off \( A \), and the fact that \((M, A)\) satisfies \( \Sigma_1(A) \) ensures the existence of an embedding of \( M \) onto an initial segment of \( N \).

It follows that \( N \models R(e, e, \gamma \delta) \), and as in the proof of Theorem 2.1, we obtain \( N \models \forall x (\gamma(x) \leftrightarrow \delta(x)) \).

Note that this result can easily be extended to \( \Sigma_n \)- and \( \Pi_n \)- formulae as well, by using the corresponding satisfaction predicates.

\(^2\) See p. 104 of (Hájek & Pudlák, 1993) for a proof of the Low Arithmetised Completeness Theorem.
Chaitin’s incompleteness theorem states that for each consistent theory T, there is a constant c, such that T does not prove any true statements of the form \( C(k) > c \), where \( C \) is some suitably chosen measure of the algorithmic complexity of \( k \). Following Raatikainen (1998), we define the algorithmic complexity of \( k \), \( C(k) \), to be the least \( \lceil \varphi \rceil \) such that \( \varphi(0) = k \). Now, Chaitin’s theorem is a easy consequence of Kripke’s Lemma. First, note that Kripke’s construction can be easily modified to yield an index \( e_0 \), such that, for each \( k \), the theory \( T + R(e_0, 0, k) \land \exists z R(e_0, 0, z) \) is consistent. Then, observe that the expression \( C(k) = c \) can now be formalised as

\[
R(c, 0, k) \land \forall j < c \lnot R(j, 0, k).
\]

Hence \( C(k) > c \) can be formalised as

\[
\exists p(R(p, 0, k) \land \forall j < p \lnot R(j, 0, k) \land p > c).
\]

**Theorem 3.1.** No sentence of the form \( C(k) > e_0 \) is provable in \( T \).

**Proof.** Let \( \lceil \varphi \rceil > e_0 \), let \( k \) be arbitrary, and suppose for a contradiction that

\[
T \vdash R(\lceil \varphi \rceil, 0, k) \land \forall j < \lceil \varphi \rceil \lnot R(j, 0, k).
\]

By Kripke’s lemma, the theory \( T' = T + R(e_0, 0, k) \land \exists z R(e_0, 0, z) \) is consistent. But \( T \vdash \lceil \varphi \rceil > e_0 \), so

\[
T' \vdash \exists j < \lceil \varphi \rceil R(j, 0, k),
\]

which makes \( T' \) inconsistent. Hence if \( \lceil \varphi \rceil > e_0 \), then \( T \) does not prove \( C(k) = \lceil \varphi \rceil \).

Now, reason in \( T \), and pick any \( \lceil \varphi \rceil \) and \( k \). If \( \lceil \varphi \rceil > e_0 \), then

\[
R(\lceil \varphi \rceil, 0, k) \land \forall j < \lceil \varphi \rceil \lnot R(j, 0, k)
\]

is false. Hence

\[
T \vdash \forall p (p > e_0 \rightarrow \lnot (R(\lceil \varphi \rceil, 0, k) \land \forall j < \lceil \varphi \rceil \lnot R(j, 0, k)))).
\]

So if \( T \vdash C(k) > e_0 \), \( T \) is inconsistent.

Chaitin’s theorem is often stated in terms of randomness, as in “only finitely many strings can be proven to be random”. To prove this version from Kripke’s lemma we need a definition of randomness, and borrow an informal one from (Li & Vitányi, 1993):

We write ‘\( x \) is random’ if the shortest [binary] description of \( x \) with respect to the optimal specification method \( [D_0] \) has length at least \( x \).
In order to make this formal, we have to define and construct such an optimal
specification method. Let \( l(n) \) denote the length of \( n \). Given a recursive function
\( \psi \), we define the \( \psi \)-complexity of \( k \), \( C_\psi(k) \), to be the least \( l(n) \) such that \( \psi(n) = k \), i.e. such that \( T \vdash R(\Gamma \psi^n, n, k) \). We also have to specify a pairing function
sending pairs of numbers \((x, y)\) to single numbers \( \langle x, y \rangle \). Raatikainen assumes
that “one can simply concatenate the two […]” giving a code with length \( x + y \).
For binary codes one could instead define \( \langle x, y \rangle \) to be the number consisting of
\( x \) ones, followed by a zero, followed by \( y \). Hence it is possible to read off the
values of both \( x \) and \( y \) from the number \( \langle x, y \rangle \). We will for these reasons assume
that one could take \( l(\langle x, y \rangle) \) to be at most \( l(x) + l(y) + 1 \).

Now we can define \( \theta \) to be \textit{additively optimal} iff for each \( \phi \), there is a constant
\( c_{\theta, \phi} \) such that for all \( k \),
\[
C_\theta(k) \leq C_\phi(k) + c_{\theta, \phi}.
\]

The constant \( c_{\theta, \phi} \) may depend on \( \theta \) and \( \phi \), but not on \( k \).

**Lemma 3.2.** \textit{The function \( \psi \) defined by the stipulation that}
\[
\psi(\langle \Gamma \phi^n, n \rangle) = k \text{ iff } T \vdash R(\Gamma \phi^n, n, k)
\]
is \textit{additively optimal}.

**Proof.** Pick a \( \phi \) and a \( k \), and let \( l(n) \) be the least length such that \( \phi(n) = k \), i.e. so
that \( C_\phi(k) = l(n) \). But since \( \phi(n) = k \), \( T \vdash R(\Gamma \phi^n, n, k) \) and \( \psi(\langle \Gamma \phi^n, n \rangle) = k \),
by definition of \( \psi \). Hence the \( \psi \)-complexity of \( k \) is \( \leq l(\langle \Gamma \phi^n, n \rangle) \). We want to
show that there is a constant \( c_{\psi, \phi} \) such that
\[
C_\psi(k) \leq C_\phi(k) + c_{\psi, \phi},
\]
but since
\[
C_\psi(k) \leq l(\langle \phi, n \rangle) = l(\Gamma \phi^n) + l(n) + 1
\]
and \( C_\phi(k) = l(n) \), we can take \( c_{\psi, \phi} \) to be \( l(\Gamma \phi^n) + 1 \). Hence the constant does
not depend on \( k \), so \( \psi \) is additively optimal.

We let the Kolmogorov complexity of \( k \), \( K(k) \), be the least \( l(\Gamma \phi^n) \) such that
\( \psi(\langle \Gamma \phi^n, 0 \rangle) = k \). Note the relationship between the Kolmogorov complexity
and the algorithmic complexity defined above: The latter is based on any given
effective enumeration of recursive functions, while the former works with the
subclass consisting of optimal enumerations. The reason for the Kolmogorov
complexity to be based on the length of \( \Gamma \phi^n \) rather than on just the Gödel num-
ber is that this allows for easy construction of the constant \( c_{\psi, \phi} \), without any
additional constraints on the Gödel numbering. Since \( \psi \) is an optimal specification
method, we can define \( k \) to be random iff \( K(k) \geq k \), and prove the desired
theorem.

**Theorem 3.3.** \textit{No sentence of the form} \( K(k) > e_0 \) \textit{is provable in} \( T \). \textit{In particular,
if} \( k \) \textit{is random and} \( > e_0 \), \textit{T cannot prove that} \( k \) \textit{is random.}
Proof. Suppose that $K(k) = \lceil \phi \rceil > e_0$. Then $\lceil \phi \rceil$ is the least numeral such that $\psi(\lceil \phi \rceil, 0) = k$, and by definition of $\psi$, $\lceil \phi \rceil$ is also the least numeral such that $T \vdash R(\lceil \phi \rceil, 0, k)$. Now use the argument from Theorem 3.1 to conclude that $T \nvdash K(k) > e_0$.

Let $k$ be any random number $> e_0$, so that we have $K(k) \geq k > e_0$. $T \nvdash K(k) > e_0$, and since $T \vdash k > e_0$ it follows that $T$ cannot prove $K(k) \geq k$.

4 POST SCRIPTUM

Michiel van Lambalgen (1989) credits the proof of the following theorem to Albert Visser and Dick de Jongh:

**Theorem.** There is a constant $c$ such that no sentence of the form $K(k) > c$ is provable in PA.

Their proof also makes use of the function $f(x)$ (with Gödel number $e$) defined in Kripke’s Lemma. By optimality of the function $\psi$ used to define $K(x)$, it follows that

$$K(k) = K_\psi(k) \leq K_f(k) + e + 1.$$  

Since for each $k$, PA+$“f(e) = k”$ is consistent, PA+$K(k) \leq 2e+1$ is consistent, so PA cannot prove the statement $K(k) > 2e + 1$.

REFERENCES


V. Yu. Shavrukov. *On risks of accruing assets against increasingly better advice*. Personal communication.


STANDARD MODELS OF ARITHMETIC

Ali Enayat

A ZF-standard model of PA (Peano arithmetic) is a model of arithmetic that is of the form $\mathbb{N}^M$ for some model $M$ of ZF (Zermelo-Fraenkel set theory), where $\mathbb{N}^M$ is the standard model of arithmetic in the sense of $M$. We provide a recursive axiomatization $\Phi_T$ of the set of arithmetical consequences of any prescribed recursively axiomatizable extension $T$ of ZF, and then we establish the following characterization of ZF-standard models of countable cofinality.

**Theorem.** Let $T$ be a recursively enumerable extension of ZF. The following conditions are equivalent for a nonstandard model $A$ of PA that has countable cofinality:

(a) There is a model $M$ of $T$ such that $A = \mathbb{N}^M$.

(b) $A$ is a recursively saturated model of $\Phi_T$.

Gödel’s incompleteness theorems demonstrated once and for all that our knowledge of a sufficiently rich structure is highly dependent on the choice of higher order formal frameworks through which the structure is viewed. This Gödelian insight has come to shape a great deal of research in the foundations of mathematics, particularly in set theory, where the higher order frameworks are provided by large cardinal axioms. For example, consider Gödel’s constructible universe $L$ of set theory. It is consistent with ZF that the class $V$ of all sets coincides with $L$, and yet it is well-known (Jech, 2003) that in the presence of

It is a pleasure to present this paper in a Festschrift volume for Christian Bennet. The results presented here arose from a question posed by Joel Hamkins and Victoria Gitman that asked whether there is a characterization of models of arithmetic that arise as the standard model of arithmetic in some model of ZF; this question was communicated to the author by Roman Kossak at the Mittag-Leffler Institute, during September of 2009. Thanks to Roman, Menachem Magidor, and Jouko Väänänen for inspiring conversations, to Jim Schmerl for reminding me of the relevance of his remarkable paper Schmerl (1995) to the questions dealt with here, to Fredrik Engström for a careful reading of the penultimate draft which resulted in weeding out many infelicities, and to Martin Kaså for his fine efforts in putting this volume together.
sufficiently large cardinals (e.g., a Ramsey cardinal) 0 exists, and in particular the satisfaction-predicate of L is definable in V, which in light of Tarski’s undefinability of truth theorem is a strong form of asserting $V \neq L$.

The discovery of Paris-Harrington phenomena in the late 1970s, and the striking work of Harvey Friedman (as recently summarized in (Friedman, 201x)) has further accentuated the pervasiveness of the Gödelian insight regarding the “standard model of arithmetic”, whose existence and uniqueness up to isomorphism, is only possible in the context of a background theory of sets. The proof of existence and uniqueness (up to isomorphism) of the standard model of arithmetic was originally carried out in the pioneering work of Dedekind (Dedekind, 1963). In the ZF-setting, Dedekind’s proof is amalgamated with von Neumann’s implementation of ordinal numbers (von Neumann, 1923), where one defines the set $\omega$ of natural numbers as the intersection of all inductive sets\(^1\), and then one uses appropriate recursion theorems to define addition and multiplication on $\omega$\(^2\).

With the above background in place, we are in a position to probe the class of models of arithmetic that are “the standard model of arithmetic” of some model $\mathcal{M}$ of ZF, i.e., models of arithmetic that arise as $\mathbb{N}^\mathcal{M}$ (read: $\mathbb{N}$ in the sense of $\mathcal{M}$), where $\mathbb{N} = (\omega, +, \cdot)$, $\omega$ is the set of finite von Neumann ordinals, and $+$ and $\cdot$ are respectively ordinal addition and multiplication on $\omega$. We begin with some definitions and observations.

1. **Definition.** Suppose $T$ is a recursively axiomatizable extension of ZF (in the same language).
   
   (a) A $T$-standard model of arithmetic is a model of the form $\mathbb{N}^\mathcal{M}$, where $\mathcal{M} \models T$.
   
   (b) $\text{PA}^T$ is the theory of the class of $T$-standard models of arithmetic, i.e.,
   
   $$\text{PA}^T := \cap \{ \text{Th}(\mathbb{N}^\mathcal{M}) : \mathcal{M} \models T \} .$$

2. **Remark.** It is easy to see, using the completeness theorem for first order logic, that
   
   $$\text{PA}^T = \{ \varphi : \varphi \text{ is an arithmetical sentence and } T \vdash \varphi^\mathbb{N} \} .$$

The above characterization makes it clear that $\text{PA}^T$ has a recursively enumerable axiomatization, and therefore, by Craig’s trick, $\text{PA}^T$ has a primitive recursive axiomatization. This motivates the search for a “natural axiomatization” of $\text{PA}^T$. Note that $\text{PA}^T$ far extends PA, since, e.g., $\text{PA}^T$ includes the

\[^1\text{In this setting, a set } X \text{ is inductive if it satisfies (i) } \emptyset \subset X, \text{ and (ii) } \forall x \in X \quad x \cup \{x\} \subset X.\]

\[^2\text{Dedekind’s uniqueness proof has been recently investigated in the context of reverse mathematics in (Simpson & Yokoyama, 2013) and (Kołodziejczyk & Yokoyama, 201x).}\]
statements $\text{Con}(\text{PA})$, $\text{Con}(\text{PA} + \text{Con}(\text{PA}))$, etc. Moreover, since Gödel’s con-structible universe $L$ satisfies $\text{GCH}$ (generalized continuum hypothesis) and $\text{AC}$ (axiom of choice), routine absoluteness consideration show that

$$\text{PA}^\text{ZF} = \text{PA}^\text{ZFC+GCH}.$$ 

Similarly, since the truth of arithmetical statements is not altered by forcing extensions, we may conclude that

$$\text{PA}^\text{ZF} = \text{PA}^\text{ZFC+¬CH}.$$ 

On the other hand, let $I$ denote the set-theoretical sentence that expresses “there is an inaccessible cardinal”. By coupling the well-known fact that $\text{Con}(\text{ZF})$ is provable in $\text{ZFI}$ with Gödel’s second incompleteness theorem we can conclude:

$$\text{PA}^\text{ZF} \subsetneq \text{PA}^\text{ZFI}.$$ 

Our first result provides a “natural” recursive axiomatization of $\text{PA}^T$. In order to describe it, we need the following definition.

3. Definition. Suppose $\{\theta_n : n < \omega\}$ is a recursive enumeration of the sentences of some extension $T$ of ZF, and let $\Phi_T$ be the collection of sentences of the form

$$\varphi \rightarrow \text{Con}(T_n + \varphi^\mathbb{N}),$$

where $T_n = \{\theta_i : i < n\}$, $\varphi$ ranges over sentences of arithmetic, $\varphi^\mathbb{N}$ is the sentence in the language of set theory that expresses “$\mathbb{N} \models \varphi$”, and $\text{Con}(X)$ is the arithmetical sentence that expresses the formal consistency of $X$. Note that $\Phi_T$ is recursive (recall that $T$ is assumed to be recursively axiomatizable).\(^3\)

4. Theorem. For every recursively axiomatizable extension $T$ of ZF, $\text{PA} + \Phi_T$ axiomatizes $\text{PA}^T$.

Proof. Each axiom of $\Phi_T$ is a theorem of $\text{PA}^T$, essentially because of the reflection theorem of ZF (Jech, 2003), and the fact that ZF can verify that a theory is consistent if it has a model; hence the set of logical consequences of $\text{PA} + \Phi_T$ is a subset of the set of logical consequences of $\text{PA}^T$. To show the other direction, suppose $\theta$ is an arithmetical sentence such that $\text{PA}^T \vdash \theta$. Then there is some natural number $k$ such that $T_k \vdash \theta^\mathbb{N}$. Therefore, since $\text{PA}$ is $\Sigma_1$-complete,

$$\text{PA} \vdash \neg\text{Con}(T_k + \neg\theta^\mathbb{N}).$$

\(^3\)A similar idea was discovered earlier by Azriel Lévy (1961). I realized this about two years after the preparation of the first draft of this note, as indicated in my FOM posting: http://www.cs.nyu.edu/pipermail/fom/2011-June/015532.html

Later, Solovay presented another axiomatization of $\text{PA}^T$ that is mentioned at the end of Remark 5; for more detail see his FOM posting: http://www.cs.nyu.edu/pipermail/fom/2010-March/014489.html

57
On the other hand, $\neg \theta \rightarrow \text{Con}(T_k + \neg \theta^N)$ is one of the axioms of $\Phi_T$, which makes it clear that $\text{PA} + \Phi_T \vdash \theta$, as desired. □

5. Remark. By putting Gödel’s second incompleteness theorem together with the fact that the $\Sigma_n$-reflection scheme is provable in the fragment of $\text{ZF}$ in which the replacement scheme is limited to $\Sigma_{n+1}$-formulas we can show that $\Phi_{ZF}$ is not axiomatizable over $\text{PA}$ by any collection of arithmetical sentences of bounded quantifier complexity; and in particular $\Phi_{ZF}$ is not finitely axiomatizable over $\text{PA}$. Also, since for each natural number $n$ there is an arithmetically definable truth predicate for $\Sigma_n$-arithmetical sentences one can use the proof strategy of Theorem 4 to show that $\text{PA}^T$ is also axiomatized by sentences of the form $\{\psi_n : n \in \omega\}$, where $\psi_n$ is the arithmetical sentence that expresses:

“$\text{Con}(T_n + \varphi^N)$ holds for all true $\Sigma_n$ sentences $\varphi$”.

6. Proposition. Every $\text{ZF}$-standard model of $\text{PA}$ that is nonstandard is recursively saturated. □

Proof: ZF can define the Tarskian satisfaction predicate for every set structure; and in particular it can do so for $\mathbb{N}$. In light of this observation, recursive saturation follows from a routine overspill argument, as in (Kaye, 1991, Proposition 15.4).

7. Theorem. Every countable recursively saturated model of $\text{PA} + \Phi_T$ is a $T$-standard model of $\text{PA}$.

Proof: The quickest way to prove this theorem is by a “resplendence argument”, as in the proof of Corollary 9 (since every countable recursively saturated model is resplendent (Kaye, 1991, Theorem 15.7)), but we take this opportunity to present a direct proof (indeed, by a variation of this proof, one can show the resplendence property of countable recursively saturated models). Given a countable recursively saturated model $A \models \text{PA} + \Phi_T$, first use recursive saturation to find $c$ in $A$ that codes $\text{Th}(A)$ by realizing the recursive type $\Sigma(v)$, where

$$\Sigma(v) := \{\varphi \leftrightarrow \Gamma\varphi \in \text{Ack} v : \varphi \in \text{Sent}\},$$

where $\Gamma\varphi$ is the Gödel number for $\varphi$, $\in \text{Ack}$ is “Ackermann’s $\in$”, i.e., $a \in \text{Ack} b$ holds iff the $a$-th bit of the binary expansion of $b$ is 1; and Sent is the (recursive) set of sentences in the usual language of arithmetic. Next, let $\langle \varphi_n : n \in \omega \rangle$ be a recursive enumeration of arithmetical sentences, and consider the arithmetical formula $\gamma(i)$ defined below:

$$\gamma(i) := \text{Con}(T_i + \{\varphi_j : j < i \land \varphi_j \in \text{Ack} c\})$$

This proposition is a descendent of a result of Ehrenfeucht and Kreisel (1966) that states that no model of $\text{PA}$ that is expandable to a model of second order arithmetic is finitely generated (recall that an infinite recursively saturated model cannot be finitely generated).
It is easy to see, using the assumption that $\Phi_T$ holds in $A$ and our choice of $c$ that $A \models \gamma(n)$ for each $n < \omega$, and therefore by overspill $A \models \gamma(d)$ for some nonstandard element $d$ of $A$. By invoking the Hilbert-Bernays arithmetized completeness theorem (Kossak & Schmerl, 2006, Thm 1.12) within $A$, we can conclude that there is a model $\mathcal{M}$ of $T$ with the following three properties:

1. $\text{Th}(\mathcal{A}) = \text{Th}(\mathbb{N}^\mathcal{M})$.
2. $\mathbb{N}^\mathcal{M}$ is recursively saturated.
3. $\text{SSy}(\mathcal{A}) = \text{SSy}(\mathbb{N}^\mathcal{M})$.

Note that (1) holds by design and (2) holds by Proposition 6. The definability of $\mathcal{M}$ within $A$ combined with an internal recursion within $A$, implies that there is an $A$-definable embedding from $A$ into an initial segment of $\mathbb{N}^\mathcal{M}$, which in turn shows that (3) is true since standard systems of models of arithmetic are invariant under end extensions.

Since the isomorphism type of a countable recursively saturated model of PA is uniquely determined by (a) its first order theory, and (b) its standard systems (Kossak & Schmerl, 2006, Prop. 1.8.1), we may conclude that $A \cong \mathbb{N}^\mathcal{M}$, as desired. □

8. Corollary. The following statements are equivalent for a countable nonstandard model $A$ of arithmetic:

(a) $A$ is a $T$-standard model of PA.
(b) $A$ is a recursively saturated model of PA + $\Phi_T$.

Proof: This follows from Proposition 6 and Theorem 7. □

9. Corollary. Every resplendent model of PA + $\Phi_T$ is a $T$-standard model.

Proof: This follows from Corollary 8 from a routine “reduction to the countable” argument. More specifically, given a resplendent model $A$ of PA + $\Phi_T$, augment the language of PA with a new function symbol $f$, and binary relation $E$, and let $\Sigma$ be the recursive set of sentences in the expanded language of arithmetic that says that $T$ holds in the structure $(A, E)$ (where $A$ is the universe of $A$), and $f$ is an isomorphism between $\mathbb{N}^{(A, E)}$ and $A$. By Corollary 8, $\Sigma$ is consistent, and therefore by resplendency, $A$ is a $T$-standard model. □

10. Remark.

(a) The countability condition is crucial for the $(b) \Rightarrow (a)$ direction of Corollary 8. To see this, recall that, by a theorem of Kaufmann, every completion of PA has a recursively saturated model $A$ all of whose classes are parametrically definable in $A$ (Kossak & Schmerl, 2006, Thm 10.1.5). Since ZF-standard models of PA always carry a full inductive satisfaction class, and by Tarski’s theorem no such satisfaction class is first order definable, this shows that $A$ is not a ZF-standard model (however, Theorem 11 presents an uncountable analogue of Corollary 8).

---

$^5$A subset $X$ of a model $A$ of arithmetic is said to be a class of $A$ if $X$ is piecewise coded in $A$, i.e., $\forall a \in A \exists b \in A$ such that $\{x \in X : x < a\} = \{x \in X : x \in \text{Ack} \ b\}$. 

59
(b) As shown in Corollary 16, the converse of Corollary 9 does not hold.

11. **Theorem.** Corollary 8 remains true if “countable” is weakened to “of countable cofinality” and \( T \) includes DC\(_\omega\) (dependent choice of length \( \omega \)).\(^6\)

**Proof:** It is sufficient to prove that every cofinal elementary extension of a \( T \)-standard model is also a \( T \)-standard model, since using a Löwenheim-Skolem argument one can easily show that a recursively saturated model of PA of cofinality \( \omega \) has a cofinal countable recursively saturated elementary submodel. More specifically, by a theorem due independently to Kotlarski and Schmerl (2006, Theorem 1.3.7), we have:

If \( A \models PA, X \subseteq A \) with \( (A, X) \models PA(X) \) [i.e., the induction scheme continues to hold for formulae mentioning a predicate \( X \), interpreted by \( X \)], and \( A \prec_{\text{cofinal}} B \), then there is a (unique) \( Y \subseteq B \) such that \( (A, X) \prec (B, Y) \).

Kotlarski (1983) used this theorem to show that a cofinal extension of a ZFC-standard model is also a ZFC-standard model. His argument can be refined to establish the following result.

11.1. **Lemma.** Suppose \( A = N^M \) for some model \( M \) of ZF + DC\(_\omega\), and \( A \prec_{\text{cofinal}} B \). Then \( B = N^{\overline{M}} \) for some \( \overline{M} \equiv M \).

**Proof:** Use the ZF-reflection theorem within \( M \) to find an increasing sequence of “ordinals” \( \alpha_n \) of \( M \) such that

\[ M \models \text{“}(V_{\alpha_n}, \in) \prec_{\Sigma_n} (V, \in)\text{”}. \]

Then, using an external induction, and an internal repeated application of the Löwenheim-Skolem theorem\(^7\), build an increasing chain \( \langle X_n : n < \omega \rangle \), with all \( X_n \subseteq V_{\alpha_n} \), satisfying the following properties:

(a) \( M \models \text{“}X_n \text{ is a countable subset of } V_{\alpha_n}\text{”}; \)
(b) \( M \models \text{“}(X_n, \in) \prec (V_{\alpha_n}, \in)\text{”}. \)

Note that by the choice of \( \alpha_n \)'s, (b) implies

(c) \( M \models \text{“}(X_n, \in) \prec_{\Sigma_n} (X_{n+1}, \in)\text{”}. \)

The \( M \)-countability of each \( X_n \), and the fact that within \( M \) each \( (V_{\alpha_n}, \in) \) is an \( \omega \)-model (since ZF proves that each \( V_\alpha \) is transitive) coupled with (c), shows that

\(^6\)Indeed this theorem holds even without assuming that \( T \) includes DC\(_\omega\), but the proof becomes more involved (one would either need to take advantage of the omitting types theorem, or bring Shoenfield absoluteness theorem in to the picture).

\(^7\)The Löwenheim-Skolem theorem is provable in ZF + DC\(_\omega\). Indeed, it is well known that over ZF, the Löwenheim-Skolem theorem is equivalent to DC\(_\omega\).
within \( \mathcal{M} \) we can copy each \((X_n, \in)\) over \( \omega \), thereby getting hold of relations \( E_n \subseteq \omega \times \omega \), and functions \( f_n : \omega \to \omega \) such that for each \( n < \omega \) we have:

(d) \( (A, E_n, E_{n+1}, f_n) \models (f_n : (\omega, E_n) \to \Sigma_n (\omega, E_{n+1}) \land N(\omega, E_n) = N(\omega, E_{n+1})) \),

where \( f : X \to \Sigma_n Y \) is the abbreviation for the set-theoretical statement expressing "\( f \) embeds \( X \) as a \( \Sigma_n \)-elementary submodel of \( Y \)." By the aforementioned Kotlarski-Schmerl theorem, (d) implies that there exist \( E_n, E_{n+1} \), such that

(e) \( (B, \overline{E}_n, \overline{E}_{n+1}, \overline{f}_n) \models \ldots \)

\[ \ldots \left( \overline{f}_n : (\omega, \overline{E}_n) \to \Sigma_n (\omega, \overline{E}_{n+1}) \land N(\omega, \overline{E}_n) = N(\omega, \overline{E}_{n+1}) \right). \]

The desired model \( \mathcal{M} \) is the direct limit of the models \( (\omega, \overline{E}_n) \).  \( \square \)

In order to present the next result (Theorem 14) we need a definition.

12. **Definition.** Suppose \( \mathcal{M} \) is a model of ZF.

(a) \( \mathbb{N}^\mathcal{M} \) is \( \mathcal{M} \)-rather classless if all the classes of \( \mathbb{N}^\mathcal{M} \) are parametrically definable in \( \mathcal{M} \) (and therefore each class is coded as a real number of \( \mathcal{M} \)).

(b) \( \mathbb{R}^\mathcal{M} \) is the field \( \mathbb{R} \) of real numbers in the sense of \( \mathcal{M} \).

(c) \( \mathbb{R}(\mathcal{M}) \) the external Scott completion (also known as the Cauchy completion) of the field of rational \( \mathbb{Q}^\mathcal{M} \) of \( \mathcal{M} \). The Scott completion of an ordered field \( F \) is the largest ordered field that contains \( F \) as a dense subfield (by a theorem of Dana Scott (1967), every ordered field has a Scott completion).

13. **Remark.** Generally speaking, \( \mathbb{R}^\mathcal{M} \) is a proper subfield of \( \mathbb{R}(\mathcal{M}) \); and when \( \mathcal{M} \) is countable, \( \mathbb{R}^\mathcal{M} \) is also countable but \( \mathbb{R}(\mathcal{M}) \) is of cardinality \( 2^{\aleph_0} \). Moreover, one can show, using (Schmerl, 1985, Proposition 1.4) that \( \mathbb{R}^\mathcal{M} \cong \mathbb{R}(\mathcal{M}) \) iff \( \mathbb{N}^\mathcal{M} \) is \( \mathcal{M} \)-rather classless.

14. **Theorem.** Let \( \kappa \) be an infinite cardinal, \( T \) be an extension of ZFC, and suppose \( \mathcal{A}_0 \) is a \( T \)-standard model of cardinality less than \( \kappa \).

(a) There is an elementary extension \( \mathcal{A} \) of \( \mathcal{A}_0 \) such that \( \mathcal{A} \) is \( \kappa \)-like and end extends \( \mathcal{A}_0 \).

(b) Moreover, if \( \text{cf}(\kappa) > \aleph_0 \), then \( \mathcal{A} \) can be required further to be \( \mathcal{M} \)-rather classless, where \( \mathcal{A} = \mathbb{N}^\mathcal{M} \), for some model \( \mathcal{M} \) of \( T \).

**Proof sketch:** Suppose \( \mathcal{A}_0 = \mathbb{N}^{\mathcal{M}_0} \), for a model \( \mathcal{M}_0 \) of \( T \). Fix a nonprincipal ultrafilter \( \mathcal{U} \) in \( \mathcal{M}_0 \), and let \( \mathcal{M} \) be the \( \kappa \)-th iterated ultrapower of \( \mathcal{M}_0 \) modulo \( \mathcal{U} \). The desired model \( \mathcal{A} \) is \( \mathbb{N}^\mathcal{M} \). The moreover clause uses the same argument as in the well-known PA-case (Kossak & Schmerl, 2006, Thm 2.2.14). \( \square \)

15. **Remark.** The full force of AC is not needed for the above construction. What is needed is enough choice to have a workable theory of ultrapowers, i.e., (i) AC\(_\omega\), and (ii) the existence of a nonprincipal ultrafilter on \( \mathcal{P}(\omega) \).
16. **Corollary.** ZF-standard models need not be resplendent.

**Proof:** By Friedman’s self-embedding theorem (Kaye, 1991, Theorem 12.4) every countable nonstandard model of $\mathcal{PA}$ is isomorphic to a proper initial segment of itself. This shows, by a “reduction to the countable” argument (as in the proof of Corollary 9), that every resplendent model of $\mathcal{PA}$ is isomorphic to a proper initial segment of itself, a property that cannot be true of any $\kappa$-like model, so thanks to Theorem 14(a) the proof is complete. □

17. **Remark.** Lemma 11.1 and Theorem 14 can also be derived as corollaries of general results of Schmerl’s (1995). Schmerl’s paper deals with sequential theories (i.e., theories in which a “$\beta$-function” that codes finite sequences is available) and which include the full scheme of induction in the extended language. It is known that such theories are reflexive (Hájek & Pudlák, 1998, Thm 3.30, Ch. III).

**REFERENCES**


We give a new elementary proof of the main theorem of (Feferman, 2012): Quantifiers implicitly definable in pure second-order logic equipped with Henkin semantics are (explicitly) definable in first-order logic.

Dear Chris,

You have recently, together with Jörgen Sjögren in (2011; 2013), proposed that concepts should be considered mathematical iff they have a unique explication (in the sense of Carnap). In this short paper I will discuss a somewhat similar idea: Symbols (or operations) should be considered logical iff they can be defined uniquely by inference rules. This idea can be traced back at least to Belnap’s response (1962) to Prior’s argument in his famous Tonk paper (1960); but let us start from the beginning.

Logic considers the form, or to be more precise, the logical form, of sentences and arguments. To determine this form we need to distinguish the logical constants from the other symbols; we need to characterize the logical constants, answering the fundamental question:

Which symbols should be considered logical?

For example, what is the reason for us think of $\forall$ and $\land$ as logical but not $P$ and $c$ in the formula $\forall x (P(x) \lor x = c)$?

Several answers to this questions has been proposed; we shall concentrate on a proof theoretic idea (or rather a model theoretic take on the proof theoretic idea): Symbols that can be defined by inference rules are logical. Belnap, in (1962), proposed that “defined” should be understood as being governed by conservative inference rules that define the symbol uniquely. Regarding uniqueness he writes:
It seems rather odd to say we have defined plonk unless we can show that A-plonk-B is a function of A and B, i.e., given A and B, there is only one proposition A-plonk-B.

A possible, and natural, interpretation of what it means for a symbol $Q$ to be uniquely definable is the following: If $Q'$ is another symbol governed by the same inference rules as $Q$, then $\varphi \vdash \varphi'$; where $Q'$ is not mentioned in $\varphi$ and $\varphi'$ is like $\varphi$ but with all occurrences of $Q$ replaced by $Q'$.

In (Feferman, 2012) Feferman argues for a model theoretic take on this proof theoretic explanation of logicality. In model theoretic terms, conservativity can be understood as existence of a sound interpretation of the symbol\(^1\) and the uniqueness can be understood as uniqueness of that interpretation. Thus, the proof theoretic notion of unique definability corresponds, in some sense, to the model theoretic notion of *implicit definability*.

In this short paper I will give a new elementary proof of the main result of Feferman in (2012) that implicitly definable quantifiers (in a second-order language) are exactly the definable ones (in a first-order language). We end with some open questions and remarks.

1 TECHNICAL PRELIMINARIES

Throughout the paper $M$ will denote the domain of a model $M$.

We will focus on quantifier symbols, i.e., symbols that bind variables. Their model theoretic interpretations are generalized quantifiers:

**Definition** (Lindström 1966). A generalized quantifier $Q$ of type $\langle n_1, n_2, \ldots, n_k \rangle$ is a (class) function mapping sets to sets such that

$$Q_M = Q(M) \subseteq \mathcal{P}(M^{n_1}) \times \mathcal{P}(M^{n_2}) \times \ldots \times \mathcal{P}(M^{n_k}).$$

The syntax and semantics of the extension of first-order logic with a quantifier symbol $Q$ of type $\langle n_1, n_2, \ldots, n_k \rangle$, $L(Q)$, can be exemplified by the formula

$$Q\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k(\varphi_1, \varphi_2, \ldots, \varphi_k)$$

together with the truth condition corresponding to an interpretation $Q$ of the symbol $Q$:

$$M, s \models Q\bar{x}_1, \ldots, \bar{x}_k(\varphi_1, \ldots, \varphi_k) \text{ iff } ([\bar{x}_1|\varphi_1]^M, s, \ldots, [\bar{x}_k|\varphi_k]^M, s) \in Q_M,$$

where $[[\bar{y}|\varphi]^M, s = \{ \bar{a} \in M^l \mid M, s[\bar{a}/\bar{y}] \models \varphi \}$, and $l$ is the length of the tuple $\bar{y}$.

The quantifiers of first-order logic, $\forall$ and $\exists$ are examples of generalized quantifiers:

---

\(^1\)This is not a perfect translation; even when given a sound and complete proof system, model theoretic conservativity (i.e., the existence of a sound interpretation) is strictly stronger than proof theoretic conservativity.
\[ M = \{ M \} \]
\[ \exists M = \{ A \subseteq M \mid A \neq \emptyset \} \]

Other examples of generalized quantifiers include the following cardinality quantifiers:

\[ (Q_0)_M = \{ A \subseteq M \mid |A| \geq \aleph_0 \} \]
\[ (Q_\alpha)_M = \{ A \subseteq M \mid |A| \geq \aleph_\alpha \} \]
\[ I_M = \{ \langle A, B \rangle \mid A, B \subseteq M, |A| = |B| \} \]

2 SECOND-ORDER IMPLICIT DEFINITIONS

If we agree that the model theoretic notion of implicit definability in some way is the right model theoretic translation of Belnap’s proof theoretic notion of unique definability, then we need to know what it means for a generalized quantifier to be implicitly definable. To answer that we are going to introduce a very basic second-order language:

The second-order language \( L_2 \) is pure second-order logic with two sorts of variables:

- Individual variables: \( x, y, z, \ldots \), and
- Predicate variables (including 0-ary) \( P, P_1, P_2, \ldots \)

The atomic formulas are predicate variables applied to individual variables, as in \( P(x, y) \). Formulas are built from atomic formulas using \( \neg, \lor, \land, \to, \forall, \exists \). Observe that we may quantify using both individual variables, as in \( \forall y \), and predicate variables, as in \( \exists P \).

The semantics for \( L_2 \) used in (Feferman, 2012) is so-called Henkin semantics: An \( L_2 \) model \( M \) consists of a set \( M \) together with subsets \( \text{Pred}_{k}(M) \) of \( P(M^k) \) for the predicate variables to range over. We follow Feferman in that the subsets \( \text{Pred}_{k}(M) \) may be chosen in a completely arbitrary way.\(^2\)

Example. Let the domain of \( M \) be \( \{ 1, 2 \} \) and \( \text{Pred}_1(M) = \{ \emptyset \} \), i.e., the unary predicates are only allowed to be interpreted as the empty set. Then

\[ M \models \forall P \forall x \neg P(x). \]

The language \( L_2(Q) \) is \( L_2 \) extended with a second-order predicate symbol \( Q \) of some specified type \( \langle n_1, \ldots, n_k \rangle \). For example, \( \forall P Q(P) \) is a sentence of \( L_2(Q) \), if \( Q \) is of type \( \langle k \rangle \) and \( P \) is of arity \( k \).

A model of \( L_2(Q) \) consists of a model \( M \) of \( L_2 \) together with an interpretation of \( Q \) as a second-order predicate, i.e., a subset of \( \text{Pred}_{n_1}(M) \times \cdots \times \text{Pred}_{n_k}(M) \). A generalized quantifier \( Q \) of type \( \langle n_1, \ldots, n_k \rangle \) defines a second-order predicate over any \( L_2 \) model \( M \) as follows:

\[ Q_M = Q_M \cap (\text{Pred}_{n_1}(M) \times \cdots \times \text{Pred}_{n_k}(M)). \]

\(^2\)Observe that in the rather restricted language of \( L_2 \), there are no first-order formulas (a formula without second-order quantifiers) without free predicate variables. Thus, a very restricted form of comprehension holds in all models for trivial reasons.
Definition. A sentence $\sigma$ of $L_2(Q)$ implicitly defines a generalized quantifier $Q$ if, for every $L_2$ model $M$, the only second-order predicate over $M$ satisfying $\sigma$ is $Q_M$.

Compare this with the notion of (explicit) definability:

Definition. A formula $\sigma(P_1, \ldots, P_k)$ of $L_2$ defines a generalized quantifier $Q$ if for every $L_2$ model $M$, and every $R_1 \in \text{Pred}_{n_1}(M), \ldots, R_k \in \text{Pred}_{n_k}(M)$:

$$(M, R_1, \ldots, R_k) \models \sigma(P_1, \ldots, P_k) \text{ iff } \langle R_1, \ldots, R_k \rangle \in Q_M.$$ 

We can now state and prove a slightly weaker form of the main result of (Feferman, 2012):

Lemma. All generalized quantifiers of type $\langle n_1, \ldots, n_k \rangle$, where $n_i \neq n_j$, for $i \neq j$, implicitly definable in $L_2(Q)$ are definable in first-order logic.

Proof. Let us give the full proof for the case where $k = 1$ and $n_1 = 1$, i.e., where $Q$ is of type $\langle 1 \rangle$. The general result follows by the same argument.

Assume $Q$ is implicitly definable in $L_2$ by the formula $\theta$, where $\theta$ is written in prenex normal form. For any $A \subseteq M$ let $M_A$ be the model of $L_2$ in which the first-order variables range over the set $M$, the second-order unary variables range over the singleton set $\{ A \}$, and the polyadic second-order variables range over $\emptyset$.

Let $A \subseteq M$ be such that $A \in Q_M$ (if there are no such $M$ and $A$ then $Q$ is first-order definable). Since $\theta$ implicitly defines $Q$ we know that $Q_M \cap \text{Pred}_1(M_A) = \{ A \}$ satisfies $\theta$, but $\emptyset$ does not. We may conclude that there cannot be any polyadic variables occurring in $\theta$ since then $\theta$ would be true or false in $M_A$ regardless of how $Q$ is interpreted.

Let $\phi$ be the first-order formula we get by removing the second-order prefix from $\theta$ and replacing all second-order variables by the single unary predicate symbol $P$. We also replace all occurrences of $Q(P)$ by some true sentence, for example $\exists xPx \lor \neg \exists xPx$.

It should be clear that for any $A \subseteq M$,

$$(M, A) \models \phi \text{ iff } (M_A, \{ A \}) \models \theta.$$ 

Note that $(M_A, X)$ is to be interpreted as the model $M_A$ in which $X$ is the interpretation of the quantifier symbol (second-order predicate symbol) $Q$.

We know that $\theta$ implicitly defines $Q$, meaning that $(M_A, \{ A \}) \models \theta$ iff $\{ A \} = Q_M \cap M_A$. Clearly $\{ A \} = Q_M \cap M_A$ iff $A \in Q_M$.

Thus, for any $A \subseteq M$,

$$(M, A) \models \phi \text{ iff } A \in Q_M.$$ 

The first-order sentence $\phi$ defines $Q$. 

\[\square\]
An easy observation now gives us the full result:

**Theorem** (Feferman). *All generalized quantifiers implicitly definable in $L_2(Q)$ are definable in first-order logic.*

**Proof.** Given a quantifier of type $\langle n_1, \ldots, n_k \rangle$ we define $Q^{+l}$ of type $\langle n_1, \ldots, n_{l-1}, n_l + 1, n_{l+1}, \ldots, n_k \rangle$ as follows:

$$(Q^{+l})_M = \{ \langle R_1, \ldots, R_{l-1}, R_l \times M, R_{l+1}, \ldots, R_k \rangle | \langle R_1, \ldots, R_k \rangle \in Q_M \} .$$

It should be clear that $Q^{+l}$ is implicitly definable in $L_2$ if $Q$ is, and that $Q$ is definable in first-order logic if $Q^{+l}$ is. The theorem then follows directly from the lemma. \qed

### 3 CONCLUSION AND DISCUSSION

The proof of the theorem rests on the simple idea that this general version of Henkin semantics does not allow second-order quantification to be used in a substantial way, not because of deep results but of completely elementary reasons. I believe the above argument suggests that the specific choice of Henkin semantics for $L_2$ is plainly wrong. Maybe a more natural choice would be to include some comprehension axioms for the Henkin semantics: Only consider models $M$ satisfying

$$\forall \bar{R} \exists P \forall \bar{x} (\varphi(\bar{x}, \bar{R}) \leftrightarrow P(\bar{x})) ,$$

where $\varphi$ is any formula of $L_2$. This leads us to a stronger notion of implicit definability and the following open question:

**Question.** Which quantifiers are implicitly definable in $L_2$ equipped with this stronger form of Henkin semantics?

Regardless of the choice of semantics to use there is clearly something odd with thinking of arbitrarily complicated formulas of $L_2$ as somehow corresponding to inference rules. Instead, to find the correct model theoretic take on Belnap’s idea to characterize logicality in terms of unique definability, we may need to analyze the concept of an *inference rule* to see what the right logic to express such rules is. A first attempt in this direction might be the observation that many inference rules can be formalized by a $\Pi_1^1$ formula. This leads us directly to the following question:

**Question.** Which quantifiers are implicitly definable in full second-order logic (i.e., second-order logic with standard semantics) with a $\Pi_1^1$ sentence?

Clearly, any first-order definable quantifier is implicitly definable with a $\Pi_1^1$ sentence. In fact, any $\Delta_1^1$ definable quantifier is.

**Question.** Are there $\Delta_1^1$ definable quantifiers that are not first-order definable?
REFERENCES


S. Feferman. Which quantifiers are logical? A combined semantical and inferential criterion. 2012.


IN DEFENCE OF THE TURING TEST

Björn Haglund

1 THE TEST

In 1949 Gilbert Ryle’s The Concept of Mind was published, with the professed intention of ‘exorcizing the ghost in the machine’. So when Alan Turing’s article Computing Machinery and Intelligence appeared the following year, many regarded it as yet another expression of the behavioristic vogue prevailing in mid-century academic Britain. The upshot of Ryle’s ideas seemed to be that human intellectual or psychological properties do not require or presuppose an immaterial soul or spirit, but can be explained in terms of (complex, multi-track) patterns of behavioral dispositions.

If you accept that, it is but a small further step to assume that anything that can exhibit a complex enough behavior also might possess intellectual capacities. Like being intelligent, for instance. But how can one go about to decide whether something – like a piece of computing machinery, say – possesses intelligence? To answer this kind of question, Turing invented what has since been called the Turing test.

The main hypothesis behind the test is that anything behaviorally indistinguishable from an intelligent being must itself be intelligent. A further assumption is that normal human persons possess intelligence. If you therefore, after having engaged in conversation with two parties, one of which is human and the other a (suitably programmed) machine, cannot tell which of the two is human, then – according to the hypothesis – both must be considered to be intelligent. In such a case we say that the machine has passed the test.

To make the test more reliable, we may repeat the experiment several times, with different people involved. After each session the interrogator guesses which responder is the computer. If no more than 50 percent of the guesses are correct, we say that the computer has passed the test.

To ensure that the test is reasonable the conversation between the interrogator and the two responders must be made by means of typed messages, and the responders may of course not be seen by the interrogator.
2 AN OBJECTION

Behaviorism is certainly not fashionable any more, and the Turing test is today often regarded with suspicion. The probably most common criticism launched against the Turing test is based on the following reasoning. There is a decisive difference between an ever so accurate computer model of, e.g., photo synthesis, and the biological phenomenon modeled. The computer model does not produce oxygen and carbon hydrates from water and carbon dioxide with the help of sunlight when running. The model may be excellent for determining, e.g., how the rate of oxygen production depends on the temperature and intensity of light etc., but it can not produce real oxygen. Computer simulated photo synthesis is simply not a kind of photo synthesis! Similarly an accurate computer model of a cognitive system can be used to simulate mental activity. But such a simulation exhibits no more real intelligence than the simulation of a rainstorm exhibits real moisture. Computer simulated intelligence is not a kind of intelligence at all! The objection amounts to the claim that a cleverly programmed computer that passes the Turing test may well appear to be intelligent, but does not exhibit real intelligence at all!

3 APPEARANCE PROPERTIES

The above objection appeals to an instance of the distinction between appearance and reality. Everyone understands the difference between ‘looking young’ and ‘being young’, or between ‘looking expensive’ and ‘being expensive’. And of course a perfectly healthy person can simulate illness (as any military physician can tell you). Appearance (often) deceives! But for certain properties the distinction between appearance and reality does not apply. There seems to be something fundamentally wrong with the claim that a picture certainly looks beautiful but really isn’t. Likewise it would be odd to say that a dish certainly tastes good but isn’t really savory. And the claim that a person looks good but really isn’t good-looking comes very close to an outright contradiction. Let us call properties where the distinction between appearance and reality isn’t applicable appearance properties. The idea is of course that such properties are used to characterize the appearance of an object, even though they are formally predicated of the appearing object itself.

4 A PRELIMINARY DEFENCE

The defence of the Turing test offered here rests on the hypothesis that cognitive properties often are appearance properties. Competence, skill, and intelligence are examples of this. To exhibit competence and to be competent is one and the same thing! It would be odd indeed to claim that a physician who constantly offers correct diagnoses, and successfully treats his or her patients, only appears
competent but really isn’t. Or that he or she merely simulates competence, and thus may be justly accused of practicing quackery. And it would certainly be strange to claim that a carpenter or cabinetmaker, who over and over again produces beautiful pieces of furniture, isn’t really skilled, but only appears to be so.

Similarly, if someone systematically answers questions in a reasonable way, quickly and correctly solves intellectual problems, reacts adequately to new proposals, i.e., behaves in intelligent ways, than he/she/it really is intelligent. One simply cannot appear intelligent without being intelligent, according to the hypothesis. Or, in other words, the only way to successfully pass the Turing test is by being intelligent.

5 A MODIFIED DEFENCE

Our first line of defence rested on a hypothesis against which two natural objections may be raised. To begin with, the idea to equate intelligence with intelligent behavior may seem far too behavioristic in spirit to be useful in a credible, independent, non-circular defense of the (behavioristically flavored) Turing test. And furthermore, the hypothesis is probably too strong to be plausible. It seems intuitively more reasonable to assume that intelligent behavior is a manifestation of an ability, that may however also (at least sometimes) remain latent.

So instead of assuming an equivalence between intelligence and intelligent behavior, it seems more plausible to equate intelligence with an ability to behave intelligently. (And correspondingly for competence and skill.)

Without entering into a discussion of the true nature of abilities, it seems safe to assume that behaving in a certain way demonstrates the ability to behave in that way. And so we are done! Passing the Turing test is a manifestation of intelligence!

It should perhaps be added that the Turing test is a strictly positive test, i.e., a failure in the test doesn’t tell us much. In particular, a failure in the test is not a reliable indication of lacking intelligence! Many believe that, e.g., elephants and dolphins are intelligent creatures, but since their ability to handle typed messages is (in all likelihood) rather limited, they could hardly pass the Turing test. But this is of course no valid objection to the test.

REFERENCES

THE MEETING BETWEEN A MATHEMATICS EDUCATOR AND A PHILOSOPHER OF MATHEMATICS

Thomas Lingefjärd

What does it mean to understand a mathematical concept? What does it mean to understand? What is really a mathematical concept? Does it exist outside man? Or is a concept internal? Why do we say a mathematics educator but a philosopher of mathematics? Together with the intriguing question “Was Immanuel Kant wrong or was he wrong?” these inquiries could in a way summarize my collaboration with Christian Bennet.

I believe that we are both interested in the similarities we discover in our thinking with regards to different topics in our daily work. We can discuss the two meanings of geometry as well as the GeoGebra construction of kissing circles or Soddy’s theorem. The theoretic side of Immanuel Kant’s life is a somewhat interesting starting point.

Ethics, or human morality, is always interesting. In fact we humans are all both moral and immoral. Immanuel Kant is often regarded as one of several central figures in modern philosophy. He fused early modern rationalism and empiricism, set the terms for the majority of nineteenth and twentieth century philosophy, and still has some influence today in metaphysics, epistemology, ethics, political philosophy, aesthetics, and other fields.

The fundamental idea of Kant’s “critical philosophy” – especially in his three Critiques: the Critique of Pure Reason (1781, 1787), the Critique of Practical Reason (1788), and the Critique of the Power of Judgment (1790) – is the human autonomy.

According to Kant the human understanding is the source of the general laws of nature that structure all our experience; and that human reason gives itself the moral law, which is our basis for belief in God, freedom, and immortality. Therefore, scientific knowledge, morality, and religious belief are mutually consistent and secure because they all rest on the same foundation of human autonomy, which is also the final end of nature according to the teleological worldview of reflecting judgment that Kant introduces to unify the theoretical and practical parts of his philosophical system. The ideal human was consistent in Immanuel Kant’s ideas.
He further argued that morality is derived from rationality and rationality in turn leads us to the categorical imperative which is absolutely not possible to negotiate about. What is right is right and what is wrong is wrong and there are NO grey zones what so ever. So Immanuel Kant derived three maxims:

- Universality check – whatever you do is correct if you are sure that you could live with the fact that everybody in universe would do the same as you did if they were in your position...

- You are never ever supposed to manipulate anybody even once – and even the greater good of humanity is not relevant. You do not manipulate. Period. It means that you could never lie. Not once! Never!

- We are all our own moral agents and need not take responsibility for any other moral agent.

The capacity that underlies deciding what is moral (not as easy as it sounds) is called pure practical reason, which is contrasted with pure reason (Immanuel Kant called that the capacity to know without having been shown) and mere practical reason (which allows us to interact with the world in experience). Hypothetical imperatives tell us which means best achieve our ends. They do not, however, tell us which ends we should choose. The typical dichotomy in choosing ends is between ends that are “right” (as when helping someone) and those that are “good” (if we are enriching ourselves).

Kant considered the “right” superior to the “good”; to him, the “good” was morally irrelevant. In Kant’s view, a person cannot decide whether the conduct is “right,” or moral, through empirical means. Such judgments must be reached a priori, using pure practical reason from outside.

Reason, separate from all empirical experience, can determine the principle according to which all ends can be determined as moral. This fundamental principle of moral reason is known as the categorical imperative. Pure practical reason in the process of determining it dictates what ought to be done without reference to empirical contingent factors. Moral questions are determined independent of reference to the particular subject posing them. It is because morality is determined by pure practical reason rather than particular empirical or sensuous factors that morality is universally valid.

Besides, Kant remarks that free will is inherently unknowable. Since even a free person could not possibly have knowledge of their own freedom, we cannot use our failure to find a proof for freedom as evidence for a lack of it. The observable world could never contain an example of freedom because it would never show us a will as it appears to itself, but only a will that is subject to natural laws imposed on it. But we do appear to ourselves as free. Luckily! Otherwise we would have nothing to fight for and discuss about. Kant argued for the idea of transcendental freedom – that is, freedom as a presupposition of the question “what ought I to do?”
Personally I do have severe problems with Kant’s suggestions that “We are all our own moral agents and need not take responsibility for any other moral agent.” I argue that we all would try to act against that if we saw someone trying to harm or kill someone we care about.

Even that Christian Bennet venerates Immanuel Kant does not mean that he has the same view of happiness. To venerate derives from the Latin verb, venerare, meaning to regard with reverence and respect. So, we respect Kant. But sometimes Kant feels a little bit stone aged. As in his definition of happiness. Makes one wonder if poor Immanuel ever was happy?

Kant defines happiness as “the state of a rational being in the world in the whole of whose existence everything goes according to his wish and will” (Kant, 1996, p. 240). Happiness is not pleasure. It is not the virtuous, joyful feeling associated with living a moral life. Happiness is simply getting what you want. It is important to keep Kant’s definition of happiness in mind when thinking about his ethics and his morality.

Christian and I are, as already said, working closely together and even if we are closely related when we look at mathematics and at science or perhaps even when we look at human knowledge and maybe especially at what constitutes a good red wine, we are totally different in terms of other virtues. Take for instance the issue of the alpacas.

Whenever Christian is showing something on his MacBook Air all of us have to look at a vast number of photos from the alpaca farm where Christian lives. I mean, seriously? What on earth would Immanuel Kant have said about the alpacas?

Alpacas have been domesticated for thousands of years. So Immanuel Kant must have met one or two of them through his travelling. But how did Kant encounter animals? After all, can we judge animals according to the same moral standards as we judge humans? Because human behavior and cognition share deep roots with the behavior and cognition of other animals, any try to find sharp behavioral or cognitive boundaries between humans and other animals would be controversial. For this reason, attempts to establish human uniqueness by identifying certain capacities are not the most promising when it comes to thinking hard about the moral status of animals.

Nonetheless, there is something important that is thought to distinguish humans from non-humans although not reducible to the observation of behavior best explained by possessing a certain capacity, namely our “personhood.” The notion of personhood identifies a category of morally considerable beings that is thought to be coextensive with humanity. In some perspectives, Immanuel Kant is one of the most noted defenders of personhood as the single quality that makes a being valuable and thus morally considerable. In the Groundwork, Kant writes:

\[
\text{[...]} \text{every rational being, exists as an end in himself and not merely as a means to be arbitrarily used by this or that will [...] Beings}
\]
whose existence depends not on our will but on nature have, nevertheless, if they are not rational beings, only a relative value as means and are therefore called things. On the other hand, rational beings are called persons inasmuch as their nature already marks them out as ends in themselves. (Kant, 1988, p. 428)

And in the Lectures on Anthropology:

The fact that the human being can have the representation “I” raises him infinitely above all the other beings on earth. By this he is a person [...] that is, a being altogether different in rank and dignity from things, such as irrational animals, with which one may deal and dispose at one’s discretion. (Kant, 2014, p. 7, 127)

So possibly Christian is running his heard of alpacas as a way to crystalize his superiority. That in turn could be related to his nobleman status. Since Christian is the ancestor to a long line of noblemen, he is easy to understand in terms of a deep and natural born need for control. And what is easier to control than alpacas?

There is nothing noble in being superior to your fellow man; true nobility is being superior to your former self. (Ernest Hemingway)

I do not at all mean that Christian deliberately want to take control over his alpacas, it is just in his blood – and in the alpacas. Compared to other stock species, alpacas are very low maintenance. Basic care includes annual shearing, toenail trimming 3-4 times annually, annual vaccination, worming, and bimonthly drenching with vitamins A, D & E, and general protection for predators. They should have a daily visual check, and need supplementary hay feeding in winter.

Probably Immanuel Kant would twist his hand and shiver in the cold wind if he saw what have become of philosophers in our time. Shepherd in his spare time and lecturer of logic when at work. What happened with the moral imperative? Well, it is actually quite logical (which it should be, right?) and rather Immanuel Kant streamlined although his strong morality is somewhat hidden. The moral imperative nevertheless leads us to humans’ rights and equity and law and order. And humans rights and equity movement often like to go back to basics and sort of share deep dreams of sewing the own clothing and such. And who will be there and sell them the very fine yarn? Well, Baron Christian Bennet of course. Instead of buying clothes from China because it is cheap, we will buy alpaca yarn from HARSÅS GÅRD - Alpacka Sjuhärad.

Is this right or wrong? You probably have learned by know that we cannot decide that. Immanuel Kant would probably say that it holds for the:

- Universality check – Christian probably could live with the fact that everybody in universe would do the same thing...Right?
• Christian would not lie about the quality of his herd of alpacas. Not once! Never!

• Christian is his own moral agent and need not take responsibility for any other moral agent.

Happy 60 years Christian and keep staying with Kant and with the alpacas.

REFERENCES


http://www.youtube.com/watch?v=xw0CmJevigw
PARADOX LOST

Helge Malmgren

Every proposition says about itself that it is true. Hence if the Liar statement, “This proposition is false” expresses a proposition, it says about itself both that it is true and that it is false. Since it is a contradiction, it does not follow from its being false that it is true. Hence the second step in the standard argument is blocked, and there is no paradoxicality left.

1 INTRODUCTION

The Liar statement\(^1\) has a strikingly evasive content. This evasiveness is most obvious in the propositional (as opposed to sentential) formulations of it, e.g.,

\begin{flalign*}
P1: \text{This proposition is false}.\(^2\)
\end{flalign*}

What is the proposition that P1 says is false? Well, it is the proposition that P1 is false, but that does not give us any clue to the identity of P1 except that it belongs to a certain half of all possible propositions. And then the standard argument shows that it belongs to the other half. No other help in identifying the proposition is available. Indeed, one might be tempted to conclude that “this proposition” in P1 lacks reference, just like “This Swedish sentence” lacks reference in

\begin{flalign*}
\text{SW: This Swedish sentence is true}
\end{flalign*}

---

\(^{1}\)I am grateful to my wife Lena for stimulating discussions around the theme of this paper.

\(^{2}\)Nothing of importance in what I say below hinges on self-reference, but the choice of one self-referential statement rather than several statements referring to each other makes my argument simpler and more direct.

\(^{2}\)I will occasionally use “P1” to stand for the sentence (symbol string) “This proposition is false”, for example in its fourth occurrence in the next paragraph, but most often it is a name of the proposition expressed by that sentence. Analogous reservations hold for the S symbols.
and that, therefore, P1 itself is not a proposition. That would be another way of reaching the (fairly) common conclusion that P1 is neither true nor false.

However, in this brief paper I will take another main route, supposing that P1 is a proposition. I will show that on this supposition, P1 is a contradiction.

My main argument hinges on also accepting the following assumption:

D: For any p, it is true that p is the same proposition as p

D follows from the so-called deflationary theory of truth, at least on one natural reading of that theory. However, it can also be derived from a number of theories of propositions, for example the simplistic one that a proposition is identical with the tripartition that it induces on the set of all possible worlds. Be that as it may, if the reader does not accept D, she will not regard my argument as valid.

2 THE MAIN ARGUMENT

It follows from D that “it is true that” can be eliminated from all sentences of the form “it is true that p” without there being any change of the identity of the proposition that the sentence expresses. But conversely, “it is true that” can be prefixed to any propositional descriptor without change of propositional identity. Let us try this on P1. The result becomes:

P2: It is true that this proposition is false

P2 says about itself that it is true, and P1 of course says about itself that it is false. But according to D, P1 and P2 are the same proposition. Therefore P1 (if you prefer to call it so) says about itself not only that it is false but also that it is true. Hence it is a contradiction, and is as such well expressed through

P3: This proposition is true and false

But stop and behold, the reader may think and say. Isn’t it obvious that the proposition that P2 says is true is different from the one that P1 says is false? So how can P1 and P2 themselves be the same?

Well: we are not talking about sentences, but about propositions. The sentence that expresses P1 in four words is different from the sentence that expresses P2 in eight words. But the propositions themselves, P1 and P2, are identical. If the reader is not convinced, assign the name \( p_0 \) to the proposition (supposedly) expressed by the sentence P1. Then use substitution to rewrite P1 as P4: \( p_0 \) is false, and use D to expand P1 to P5: It is true that \( p_0 \). Finally, from the fact that \( p_0 \) is also the proposition that P1 expresses, it follows that P1 is a contradiction.
3 THE POINT

The second step in the standard Liar argument is that in virtue of P1’s saying about itself that it is false, it follows from its being false that it is true. However, since P1 does not only say about itself that it is false but also that it is true, it does not follow from its being false that it is true. Therefore, the supposition that P1 is false is consistent and there is no paradox left.

4 THE SENTENTIAL INTERPRETATION

Let us look at the sentential Liar sentence

S1: This sentence is false

Truth and falsehood concur to sentences only in so far as they express propositions.\(^3\) My above argument can therefore be reconstructed for this case too. This is obvious if one allows S1 to be rewritten as

S2: The proposition expressed by this sentence is false

Following the same procedure as above, assign the name \(p_0\) to the proposition referred to in S2. Again, according to D, S2 says about \(p_0\) not only that it is false but also that it is true.

However, the translation S2 of S1 can be challenged and has been challenged. The following statement might seem immune to the above argument since it does not use a demonstrative phrase to pick out the purportedly false proposition:

S3: This sentence expresses a false proposition

Now “expresses a false proposition” must here mean “expresses a total proposition that is false”. One might use “express a proposition” in such a way that the sentence “The moon is made of green cheese and the sun is made of orange juice” could be said to express two propositions. But this would not work for S3 since then it would not follow from S3’s being false that it is true. It could be the case that S3 expressed two false propositions. So to get the standard Liar argument running we must be using “express” in a sense in which a sentence (in its context) expresses one and only one proposition. And then we are back in S2.

\(^3\)This thesis is sometimes called “contextualism” because it is usually coupled to the idea that what a sentence expresses depends on the context in which it is used. I think it is too obviously true to deserve the name of a “presupposition” for my argument.
5 AND WHAT IF P₁ IS NOT A PROPOSITION?

As I have hinted above, I do not believe that P₁ is a proposition. Although it surely has the grammatical form of a proposition, it does not have any logical subject (to use a traditional terminology). Indeed, this point seems pretty obvious to me, but I will not try to prove it. There is a simple route from it to the conclusion that P₁ does not generate a paradox: even if it follows from P₁’s being true that it is false, and from its being false that it is true, it is just neither.

However, one would perhaps also like to know how my above argument to the same conclusion fares if one drops the assumption that P₁ expresses a proposition. I would argue that assumption D holds for pseudo-propositions as well and that logic can be used not only on real propositions but also on pseudo-propositions. Consider as an example

E₁: Pirotes carulate elatically

The pseudo-proposition expressed by E is, it seems to me, the same as that expressed by

E₂: It is true that pirotes carulate elatically

and E₁ entails E₂.

So the main argument should still go through, with the slight difference that p₀ is now a pseudo-proposition, and with the new conclusion that P₁ has the form of a contradiction. In no way, therefore, can a paradox be derived from it.⁴

⁴The sentential formulation

S₄: This sentence does not express a true proposition arguably manages to include the essence of P₁ even when P₁ is allowed to be neither true nor false. However, I must leave the treatment of this case and its ramifications to the reader.
PHILOSOPHY OF MATHEMATICS
AND MATHEMATICS EDUCATION:
SOME REFLECTIONS

Jörgen Sjögren

Mathematical concepts are abstract, but have an origin in empirical reality. They are formed via sequences of explications in Carnap’s sense. Regarding concept formation in this way makes it possible to point at differences between mathematics and science. In this paper I present some examples of explications in both mathematics and science, and then give an account of Carnap’s work on explanation. This is followed by a discussion of mathematics education. It highlights the importance of historical and philosophical awareness of mathematics teachers. It also points at some factors that distinguish mathematics from science as, e.g., the role of proof in mathematics.

*I had made an empirical discovery and it had all the force of a mathematical proof.* Paul Auster

*In Strum’s brain it wasn’t mathematics that reflected the world, but the world that was a projection of differential equations; the world was a reflection of mathematics.* Vasily Grossman

1 INTRODUCTION

There are at least three ways in which mathematics differs from empirical science at the conceptual level. In this paper I discuss these differences, and also suggest how these differences can be treated in mathematics education. The three distinguishing features I will focus on are the following.

Mathematical concepts have, just like empirical ones, an origin in empirical reality. This origin is often distant in modern mathematics, but can sometimes be traced in a ‘concept archeology’. Mathematical propositions and theories are
not, however, accepted on empirical grounds; they are not tested against empirical reality. Mathematical theories are proved from axioms or other propositions that are taken for granted. Scientific hypotheses and theories must, on the other hand, have empirical support, if they are to be accepted.

A second way of describing the difference between mathematics and empirical science is to see it as a difference between cumulative and non-cumulative disciplines. It is obvious that we can teach Euclidean geometry in mathematics, but not Aristotelian mechanics in physics but for more than historical purposes. Concept changes in science are more radical and of another character than in mathematics (see below).

Furthermore, mathematics deals with uniquely explicable concepts, while concepts in empirical sciences don’t have this robustness, and change with theory (r)evolution. Mature mathematical concepts have a bedrock position in a mathematical structure or theory. The road to this position may have been long as the evolution of, e.g., the function concept from the seventeenth century to modern times makes evident (see below).

None of these differences are unquestioned, even if the first one seems to be rather standard. Concerning the first issue, there is an ongoing discussion of the role of proof, what is to be counted as a proof, and of the possible empirical character of mathematics. Concerning the second issue, Kuhn’s characterization of the development of science as a series of revolutions is not universally accepted, and there is, furthermore, a discussion whether there are revolutions in mathematics or not. The third point seems, however, not to have been noticed earlier.

In any case, these three differences will be taken for granted in this paper, and in the sequel I suggest a way to clarify these differences between mathematics and science. The explanation is based on a theory of concept formation, having its origin with Rudolf Carnap, and I use his ideas of concept formation as explication to describe and explain the differences mentioned.

It is important in mathematics education to point at the special character and methods of mathematics, and not to treat it as an empirical science. Treating concept formation as in this paper may help in facilitating an understanding of mathematics. An awareness of philosophical and historical aspects of concept formation on behalf of mathematics teachers would, furthermore, be fruitful in mathematics education, and I also try to illustrate this below.

The structure of the paper is as follows. Section two presents some examples to highlight the process of concept formation in mathematics and science, respectively. This is followed up with an account of Carnap’s idea of explication. Finally, in section three, some consequences for mathematics education

---

1See e.g. the contributions of Michael Detlefsen, Jonathan Borwein, and Joseph Auslander in (Gold & Simons, 2008) on proof. See (Jenkins, 2008) for an empiricist view of mathematics, and (Sjögren & Bennet, 2014) for some comments.

2See (Gillies, 1992).

3This theme is discussed in (Bennet & Sjögren, 2013), and is developed mainly by Christian Bennet.
are discussed on the assumption that these differences between mathematics and science are important, and that they are essential for developing a reasonable understanding of mathematics.

2 ON EXPLICATIONS

Mathematical concepts, as well as concepts of empirical science, are abstractions. Aristotle states that these ‘objects’ cannot be separated from the substances (individual objects) in which they inhere, but they can be isolated or separated in thought. He also indicates the difference between objects of mathematics and objects of physics in that physics treats accidental properties while mathematics treats essential ones. Also, according to Aristotle, results in mathematics and empirical sciences are grounded in different ways. It is

for the empirical scientists to know the fact and for the mathematical to know the reason why; for the latter have the demonstrations of the explanations, and often they do not know the fact ...

for the doctor to know the fact that circular wounds heal more slowly, and for the geometrician to know the reason why.

The process of formulating these abstractions can be described in a metaphysically neutral way as sequences of explications. Carnap introduced this method of making concepts more precise in the middle of the twentieth century. The idea of an explication is the frequent need in mathematics and science to replace vague, or otherwise non-clear, concepts with more exact ones. Before turning to Carnap’s idea some examples of explications in mathematics and empirical science will be considered.

2.1 CONCEPT FORMATION IN MATHEMATICS

Consider first the concept function; a concept with a long history of evolution. Preconcepts date back to antiquity in, e.g., the tables of chords (trigonometric tables) of Ptolemy and his predecessors. An empirical origin may be the dependency of some entity $B$ on some entity $A$, or the affection of $A$ on $B$. This in turn is natural to connect with a relation of causal dependency, i.e., the concept has an empirical origin.

---

4 See e.g. Aristotle, Phys., Book II, 193b – 194a.
5 An. Post. B. 13. 79a3 – 6. (I use (Barnes, 1984) when I quote Aristotle.)
7 See (Kline, 1972) for a more comprehensive treatment.
In the period dating from the seventeenth to the early twentieth centuries, the function concept evolved from a geometric one, via an algebraic, to the modern, set-theoretic conception. The term “function” was first used by Leibniz in phrases such as “a tangent is a function of a curve”, and the spirit in the infinitesimal calculus as developed by Newton, Leibniz, and their contemporaries was geometric in the sense that functions were identified with curves (Kleiner, 1989). The first algebraic definition is due to Johann Bernoulli (1718) and reads:

One calls here a Function of a variable a quantity composed in any manner whatever of this variable and constants.8

A couple of decades later Leonard Euler (1748) defined a function as an “analytical expression”:

A function of a variable quantity is an analytical expression composed in any manner from that variable quantity and the numbers or constant quantities.

Euler does not define “analytical expression”, so it is not completely clear what expressions he regards as functions, but presumably he meant an algebraic formula. The approach is, however, clearly algebraic.

Two important events in the evolution of the function concept are the vibrating string controversy, and the development of Fourier series as representations of functions over an interval. At the time of these discussions (1822), Joseph Fourier had a very general conception of functionality.

In general, the function \( f(x) \) represents a succession of values or ordinates each of which is arbitrary. […] We do not suppose these ordinates to be subject to a common law; they succeed each other in any manner whatever, and each of them is given as it were a single quantity.

Some years later, Dirichlet is very explicit concerning the arbitrary correspondence between the dependent and the independent variables, a terminology that is hardly relevant any longer with its connection to causality. This general concept of a function from a set \( A \) to a set \( B \) is nowadays usually defined \( \forall x \in A \exists! y \in B \ y = f(x) \), or that \( f \) is a certain type of subset of \( A \times B \). While Dirichlet regarded functions from reals to reals there are no restrictions made on the sets \( A \) and \( B \) in this definition. The function concept has thus evolved from the conception of a function as a geometric curve to a widest possible (?) set-theoretic version.

Not everyone was pleased with this wide concept. New types of ‘pathological’ functions, like Dirichlet’s function and plane filling curves, were defined to highlight different issues, and Poincaré, e.g., protested against these “bizarre” functions which served no purpose.9 In the beginning of the twentieth century

---

8This quotation, and the following ones, are from (Rüting, 1984).

9Dirichlet’s function is defined \( D(x) = \begin{cases} c, & \text{if } x \text{ is rational} \\ d, & \text{if } x \text{ is irrational} \end{cases} \). See also (Kline, 1972, p. 973).
Baire, Borel, and Lebesgue also protested against the “lawless” connection between the variables, but Hadamard retorted in their controversies that denying the set-theoretic function concept would lead back to the eighteenth century conception. The demand for definiteness, or possibility to determine the value of a function is at present emphasized by various schools of constructivism.

In the twentieth century new concepts like, e.g., $L^2$-functions and generalized functions have been introduced to meet new needs of mathematics and physics. These entities are however not functions in the meaning of one entity uniquely determining another.\(^{10}\)

It thus seems that with the set-theoretic function concept the most fruitful conception of the original, vague idea has been found. From an empirical origin as some kind of dependency, the function concept has evolved via geometric and algebraic ones to the modern set-theoretic. The concept has in the process become more and more general to meet up with mathematical needs as well as to be able to handle pathological cases. Also, the original idea of dependency may have been broken with the set-theoretic concept with the ‘arbitrariness’ (Fourier) or ‘lawlessness’ (Baire et al.) of the connection between the variables. Some, however, cling to the original intuition in its algebraic version demanding the possibility of determining the values of functions. Others, the great majority of mathematicians, want a function concept as wide as possible, and accept the set-theoretic one.

I have treated this example in some detail because of the importance of the concept in mathematics education. However, for several mathematical concepts akin processes of evolution have occurred, and, consequently, similar stories can be related. Consider, e.g., the concept number. In antiquity a number was regarded as what we today describe as a positive, rational number. Later on real numbers, negative numbers, and complex numbers, \textit{etc.} have been introduced, but, e.g., the complex numbers were not fully accepted by mathematicians until the early nineteenth century. This development of the number concept is described by Kitcher (1984) as a reinterpretation of the concept, but I prefer to regard it as a sequence of explications.

When a mathematician, or group of mathematicians, is engaged in explicating a concept, their freedom is utterly limited. The ‘new’ concept must fit into an already existing, deductive structure. This was one reason behind the intense discussion of pathological functions in the nineteenth century. It is also possible to read Lakatos’s \textit{Proofs and Refutations} as an exposition of the problem how to find the most fruitful explication of the concept \textit{polyhedron}. The concept is to be as general as possible while it also fits in a known structure (Lakatos, 1963–64).

\(^{10}\)For $f, g \in L^2$, $f$ and $g$ are identical iff $f = g$ \textit{a.e.}, \textit{i.e.} they differ at most in a set of zero measure. As an example of a generalized function, consider Dirac’s $\delta$ function. This is an entity satisfying $\delta(x) = 0$, if $x \neq 0$, and $\int_{-\infty}^{\infty} \delta(x)dx = 1$. No \textit{function} satisfy these two conditions.
2.2 CONCEPT FORMATION IN SCIENCE

Turning then to concept formation in empirical sciences, Carnap exemplifies with the concept fish (animal living in water) as explicated by pisces (in biological taxonomy). This explication starts with an unclear concept, and replaces explicandum with a concept in a scientific system of concepts. To consider another concept, regard species. Aristotle defines different species by genus and differentia specifica, an idea that Linnaeus worked out in great detail, but this, of course, doesn’t give a definition of the concept species. The classical idea of the stability of species dates from approximately the end of the sixteenth century, since before that time no clear criteria existed that could be used to distinguish different species. Discussions whether there are species or just individuals occurred, just like questions on the stability of species.\(^{11}\) In the 1690’s John Ray suggested a genealogical definition of species, meaning that two individual plants belong to the same species if they have a common ancestor, i.e. genealogically originate from the same seed.

On the possibility of defining species, Darwin states that

> no one definition [of species] has as yet satisfied all naturalists; yet every naturalist knows vaguely what he means when he speaks of species. (Darwin, 1859, ch. 2)

And a little bit further down in the text he suggests that

> [...] in determining whether a form should be ranked as a species or a variety, the opinion of naturalists having sound judgement and wide experience seems the only guide to follow.

Darwin himself does not enter into discussions in On the Origin of Species of various definitions of species, but this issue is of course essential in discussing evolution of species. Turning to modern textbooks the concept species is defined, e.g., as

> a group of actively or potentially interbreeding populations that is isolated reproductively from other such groups.\(^ {12}\)

> A species, in the modern view, is a genetically distinctive population, a group of natural demes that share a common gene pool and are reproductively isolated from all other such groups. A species is the largest unit of population within which effective gene flow occurs or can occur. (Gould & Keeton, 1996, p. 488)

There is still, however, no universally accepted definition of this classificatory concept.

\(^{11}\)See (Toulmin & Goodfield, 1965), especially chapter VIII.

\(^{12}\)(Ricklefs & Miller, 2000, p. 384) This definition originates from Ernst W. Mayr in 1942.
This may be compared with the situation in physics. Carnap exemplifies with the concept *warmth*, where the classificatory concept *warm*, via the comparative concept *warmer*, has the quantitative concept *temperature* as explicatum. In Newtonian physics an everyday concept such as *work* has received a quantitative explication as the physical concept work \( W = \int F \cdot dr \). Still, even concepts of physics change; they are theory dependent. Gravitation in Newton’s physics is one thing, but it is something else in Einstein’s general theory of relativity, and it will probably be something different in a quantized theory of gravitation.

In concept changes as these described above, there are radical breaks; phenomena are explained with new kinds of entities. With new knowledge in genetics it is possible to regard *species* in a way that is radically different from Mayr’s, and his understanding of the concept is totally different from Ray’s (see also below). Concerning gravitation Newton makes essential use of the concept *force*, while Einstein’s idea of gravitation is geometric. Both these theories are classical, *i.e.*, non-quantized, and what a quantized concept of gravitation would amount to we don’t know. Note that one must not accept Kuhn’s idea of paradigms to regard the concept changes as these above as radical.

Before entering into a discussion of differences between concepts of physics and mathematics, an account of Carnap’s idea of concept formation via explications will be given.

### 2.3 CONCEPT FORMATION AS EXPLICATION

Rudolf Carnap has throughout his career been engaged in promoting precision and exactness in scientific and mathematical disciplines. Here focus will be on his later work on *explication* as a means to generate more exact concepts.

Carnap introduces the concept *explication* in a paper on *probability* in 1945 (Carnap, 1945). In explicating a concept the question is not, as is often the case in science and mathematics,

one of defining a new concept but rather of redefining an old one. Thus we have here an instance of that kind of problem […] where a concept already in use is to be made more exact or, rather, is to be replaced by a more exact new concept. (*Ibid.*)

In an explication the *explicandum* is the more or less vague concept, and the new, more exact one, is the *explicatum*. As an example Carnap mentions Frege’s and Russell’s explication of the cardinal number *three* as the class of all triplets.

In *Meaning and Necessity* Carnap describes the concept *explication* in the following manner.

The task of making more exact a vague or not quite exact concept used in everyday life or in an earlier stage of scientific or logical development, or rather of replacing it by a newly constructed, more
exact concept, belongs among the most important tasks of logical analysis and logical construction. We call this the task of explicating, or of giving an explication for, the earlier concept; this earlier concept, or sometimes the term used for it, is called the explicandum; and the new concept, or its term, is called an explicatum of the old one. (Carnap, 1947, pp. 7f)

As before Carnap exemplifies with cardinal number, but he now adds truth, his own efforts to handle concepts like L-truth (logical truth, analytic), and phrases of the form the so-and-so, etc. He also briefly mentions how the relation between explicandum and explicatum concerning meaning ought to be understood.

Generally speaking, it is not required that explicatum has, as nearly as possible, the same meaning as explicandum; it should, however, correspond to explicandum in such a way that it can be used instead of the latter. (Ibid., p. 8.)

A bit further down in the text Carnap discusses the possible correctness of an explication, and states that

there is no theoretical issue of right or wrong between the various conceptions, but only the practical question of the comparative convenience of different methods. (Carnap, 1950, p. 33)

Finally, Carnap devotes chapter one of Logical Foundations of Probability to the concept explication Carnap (1950). The main problem in this book is to construct explications of concepts like degree of confirmation, induction, and probability. The process of making explications is described as above, and now Carnap also emphasizes the need to clarify explicanda in order to make clear which sense of a vague and unclear explicandum it is that needs to be explicated. It was a clarification of this type the 1945 paper discussed. Carnap does not present, but in some vague phrases, any criteria explicatum must fulfil in the 1945 paper, or in Meaning and Necessity. This is taken care of in Logical Foundations of Probability.

1. The explicatum is to be similar to the explicandum in such a way that, in most cases in which the explicandum has so far been used, the explicatum can be used; however close similarity is not required, and considerable differences are permitted.
2. The characterization of the explicatum, that is the rules of its use [...], is to be given in an exact form, so as to introduce the explication into a well-connected system of scientific concepts.
3. The explicatum is to be a fruitful concept, that is, useful for the formulation of many universal statements (empirical laws in the case of a nonlogical concept, logical theorems in the case of a logical concept).
4. The explicatum should be as simple as possible; this means as simple as the more important requirements (1), (2), and (3) permit. \((Ibid.,\ p.\ 7.)\)

Concerning the possible correctness, or truth, of an explication, Carnap reinforces the statement from *Meaning and Necessity* that there is no question of right or wrong. Since explicandum is not an exact concept, the problem of explication is not stated in exact terms, so

the question whether the solution is right or wrong makes no good sense because there is no clear-cut answer. The question should rather be whether the proposed solution is satisfactory, whether it is more satisfactory than another one, and the like. \((Ibid.,\ p.\ 4.)\)

Carnap’s position concerning the possible correctness of an explication as it is expressed here may be too defensive. In some cases it is even possible to prove the correctness of an explication.\(^{13}\)

Later on Carnap points out that there is no sharp dividing line between scientific and non-scientific languages (Carnap, 1963, pp. 933–940). Scientific languages arise from non-scientific ones, and scientific vocabulary works its way into non-scientific languages. An explication does not have to be formulated in a precise, formal language. The point is that an explication replaces the imprecise explicandum by a more precise explicatum. Therefore, whenever greater precision in communication is desired, it will be advisable to use the explicatum instead of the explicandum. The explicatum may belong to the ordinary language, although perhaps to a more exact part of it. (Carnap, 1963, p. 935)

2.4 DIFFERENCES BETWEEN MATHEMATICS AND SCIENCE

By regarding the examples above as analysed in a Carnapian style as explications, the three initially described differences between mathematics and science can be clarified. Concerning the first one, several mathematical concepts have, just like *function* and *number*, an origin in empirical reality. As concepts receive new, and often more abstract, explications they are distanced from this origin. When a new explication of a concept is suggested, the question is normally whether it is more suitable than the old one in explanations of the facts or problems that are at issue. There were no empirical questions involved in the discussions concerning, e.g., the reasonableness of including ‘pathological functions’ among functions, and Dirichlet’s explication was the most fruitful one in this case. The acceptance of complex numbers came with the introduction of the Argand plane, and of course there were no empirical questioned involved

\(^{13}\)See (Sjögren, 2011, pp. 19–24) for a discussion.
The introduction of scientific concepts follows another route. In this case the question is if they are better suited to explain empirical phenomena, and this is certainly an empirical question. Failure here may imply that concepts are rejected.

Concerning the cumulative character of mathematics it is obvious that we, as mentioned in the introduction, can use and teach Euclidean mathematics. New concepts of geometry have been introduced as well as new geometries, but this doesn’t mean that the old concepts are misplaced. Turning to science, Ray’s concept of species is useless, and the genetic conception is radically different from Mayr’s, but can possibly ‘explain’ it. Furthermore, if we are to describe the planetary orbits of our solar system in detail, especially the orbit of Mercury, we cannot use Newton’s conception of gravity, but must lean on Einstein’s. Of course, Newtonian mechanics is fruitful in lots of applications, and consequently it makes sense to study it, but note that the relativity theories display its limits. GPS navigation, e.g., needs both special and general relativity. One could compare with the relation between Aristotelian and Newtonian mechanics. The concept force, e.g., are totally different, and Aristotle’s mechanics is qualitative while Newton’s is quantitative. Efforts were made, before Galileo entered the scene, to treat Aristotelian physics quantitatively, but they all failed. Thus, mathematical concepts evolve via explications, or reinterpretations, towards generality, and towards what seems to be the ‘correct’ one, while scientific concepts now and then undergo radical changes.

This leads to the third issue, the robustness of mathematical concepts. Mature mathematical concepts may have found there final formulation. They can be made precise in but one way. They have as such entered a mathematical discourse, and, as it seems, cannot be rejected. There are, of course, never any guarantee for consistency. A mathematical theory, and the concepts used in it, may lose interest and relevance, but this is not a question of rejection. What may happen with mathematical theories is that mathematicians end being interested in them. The theories may, e.g., be fully explored; they may even be forgotten. This may be compared with the introduction of new concepts. The concept set, as introduced by Cantor, lead to inconsistencies and thus had to be modified. Concepts of science, on the other hand, are not as robust as those of mature mathematics, they change with theory evolution, and this change can be radical. These three distinguishing features at the concept formation level between mathematics and science are important, and I will now turn to some consequences for mathematics education.

3 SOME CONSEQUENCES FOR MATHEMATICS EDUCATION

We have now seen how concept formation in mathematics and science can be analyzed as sequences of explications. A consequence of this analysis is that

---

14 This idea is explored in more detail in (Bennet & Sjögren, 2013).
we have an instrument that makes it possible to describe and understand differences between mathematics and science at the concept formation level, and how concepts find their positions in theories. In my opinion it is important that mathematics education makes these differences clear; at least they should be known, and reflected upon, by teachers of mathematics at all levels. What will follow in this section is, taking our analysis for granted, some suggestions for making the differences manifest. Since the analysis of concept formation above is historico-philosophical it is not especially remarkable that the examples below emphasize historical aspects. I will thus begin with a couple of examples from the history of mathematics, and how these can be fruitful in mathematics education.

Torricelli ‘proved’ that the area of a circle is $\pi r^2$ by the following argument.

Since every circle correspond uniquely to a line in the triangle, the area of the circle must equal that of the triangle, and this area is $\frac{2\pi r \cdot r}{2} = \pi r^2$. The argument is intuitive, easy to understand, but of course wrong; you cannot add ‘areas’ of lines; $\infty \cdot 0 \neq \pi r^2$. At the same time it may prepare for arguments using infinitesimals. The circles can be replaced with rings, and the lines with stripes, and via a limit process a correct argument can be formed. What has occurred here in the mathematical development is that a deeper understanding of the involved concepts has been gained; it is not that results have been overthrown. New explications have been developed, or to speak with Kitcher, there has been a reinterpretation of concepts.

Consider again the function concept discussed in section two. For mathematicians, this concept took approximately two centuries to develop into its modern form. It is then not astonishing that it is a difficult concept to understand for students. Also in this case a historic perspective and awareness would benefit teaching. One can begin with simple, causal connections between events, and mathematize this, perhaps in form of tables, and then go on to more complicated relations. Textbooks often define function as the set-theoretic notion, and then just exemplifies with the algebraic or the geometric one, identifying functions with algebraic expressions or curves. Such an exposition can be hard for a student to grasp. One may also note that Euler, who defined the function concept

---

15 On some examples of the use of history in mathematics education, see (Man-Keung, 2000). The importance of history in mathematics education is also emphasized by Guershon Harel (2008).
algebraically as analytical expressions, wouldn’t have accepted, e.g.,

\[ f(x) = \begin{cases} \sqrt{x}, & \text{if } x \geq 0 \\ \sqrt{-x}, & \text{if } x < 0 \end{cases} \]

as a function since it isn’t one algebraic expression. This is a way of describing functions that often raises problems for students. What is of importance here is that one cannot stop with functions as empirical connections. One has to discuss how this concept can be generalized and made precise, and also to keep the conceptions apart. Both these examples highlight the process of concept formation in mathematics, and are in this way connected to the issues in the introduction.

Then, concerning the first difference mentioned in the introduction, I think it is important in mathematics education to discuss both how mathematical truths may be arrived at, and how they can be proved. The conclusion in the analysis of concept formation is that concepts of mathematics often have an empirical origin, but propositions involving these concepts cannot be empirically proved; there are no ‘empirical’ proofs of mathematical truths. In science, on the other hand, the ultimate reason for regarding a proposition as true must be observation or experiment.

As a simple example, consider the sum of the angles in an Euclidean triangle. Measurements of the angles may lead to the guess that the sum equals 180°, and this conjecture can then be proved by ordinary arguments. Another example is ‘puzzle proofs’ of Pythagoras theorem. Cutting and pasting geometrical figures doesn’t provide a proof of the theorem, but can help in understanding, and lend ideas of how a proof can be carried out with algebraic or other means like Euclid’s beautiful proof of the theorem. This issue touches upon the problem of using visualizations in mathematics. Kajsa Bråting and Johanna Pejlare have a point when they argue that we have to know the actual mathematics in order to see it in a visualization (Bråting & Pejlare, 2008). This is perhaps not all there is to the problem. Consider the two functions

\[ f(x) = \begin{cases} x \cdot \sin(\frac{1}{x}), & \text{if } x \neq 0 \\ 0, & \text{if } x = 0 \end{cases}, \quad \text{and } g(x) = \begin{cases} x^2 \cdot \sin(\frac{1}{x}), & \text{if } x \neq 0 \\ 0, & \text{if } x = 0 \end{cases} \]

Even if a person is educated in mathematics (s)he can’t see neither that \( f \) is continuous, but not differentiable at 0, nor that \( g \) is both continuous and differentiable at 0 by looking at their graphs. For another example, consider the intermediate value theorem.16 Here an educated mathematician knows that you cannot see that the point set that constitutes the function satisfies the theorem, while a less educated person, regarding a function as a geometrical curve generated by a moving point, probably would regard the theorem as trivially true.

---

16 The theorem states that if a function \( f \) is continuous on a closed interval \([a, b]\), then \( f \) takes all the values between \( f(a) \) and \( f(b) \).
Of course, a picture could be very telling, and (Brown, 2008) contains several examples of ‘picture proofs’. Pointing at the importance of proof as the method of verification in mathematics, is not to argue for using a systematic representation of mathematics via the definition-axiom-theorem model. This is a method of presentation used to organize developed theories. It has been an ideal since Euclid’s Elements, but disregarding, e.g., Newton’s Philosophia Naturalis, it isn’t until the nineteenth century with the rigourization of mathematics that it has become the standard method of presenting mathematics.17

The function concept discussed in section two is an example of how mathematics in the process of concept formation gives rise to unique concepts. The idea that mathematical concepts can be made precise in only one way may be difficult to illustrate or make use of in teaching, but it should be possible to present concepts in a way that direct students towards more general versions of the concept, even the ‘correct’ one, and not to dead ends. For an example, consider subtraction; when trying to understand what \( 5 - 2 \) amounts to, one can say that we have to remove two objects from a pile of five. Then pondering over \( 2 - 5 \) seems to leave us with two options; it is impossible, or there will be zero objects left; and \( 5 - (-2) \) seems simply impossible. Here it should be clear that ‘removing’ objects is an acceptable operation, but it is limited, and it is not subtraction. To remove objects is a physical action, while subtraction is a mathematical operation. One step towards a more fruitful concept is to talk of ‘distance’ between numbers on the number line. This can take care of \( 5 - (-2) \), but not of \( 2 - 5 \). To come to grip with this example we need a still more general version; subtraction as the inverse operation of addition, and this can be defined algebraically as \( a - b = c \Leftrightarrow_{def} b + c = a \), or discussed geometrically with the number line and vectors, or rather a combination of both. In this way we move from the concept removing, via distance to inverse of addition to reach the concept subtraction. This can be understood as a sequence of more general explications, or with Kitcher as reinterpretations. It is important that these concepts are not presented as three totally different ones. One may also note that subtraction as the inverse operation of addition need not be more difficult to explain than the other two, and as a strength, it generates no new conceptual problems. As always in educational matters one has, of course, to consider the conceptual development of the student and use, in the given situation, appropriate explanations. Sometimes it may be fruitful even in elementary mathematics to be more abstract, since abstract reasoning may facilitate understanding of phenomena.18

---

17 See e.g. (Umland & Hersh, 2006).
18 See e.g. (Kaminski et al., 2008).
We have now seen via examples that mathematics teaching can benefit from history, and that an awareness of the concept formation process on behalf of teachers can make it possible to direct student towards correct concepts.

4 CONCLUDING REMARKS

The process of concept formation in mathematics and science can be described as sequences of explications of concepts which often has an empirical origin. Analyses of how concepts are formed point to a distinction between concepts of mathematics and concepts of empirical science; mathematical concepts have an robustness that concepts of science don’t have. It is important in mathematics education to make these differences between mathematics and science as clear and distinct as possible. One important step in accomplishing this is for teachers to have good insights in how concepts of mathematics have evolved, and how they are incorporated in mathematical theories; i.e., knowledge of the history of mathematics is important in mathematics teaching. In the teaching process it must be clear that mathematical entities are abstract. This in turn presupposes philosophical insights on behalf of teachers. A successful teaching and learning situation also presupposes dialogues in the classroom to facilitate concept acquisition on behalf of the students; silent mathematics is not a way towards understanding. It is also essential that mathematical propositions are deductively proved from propositions that are considered true; they are not empirical generalizations. In general a proof can be regarded as an answer to a ‘why’-question, and such questions must be answered seriously; i.e. proof must be provided if possible.

A final question that seems to be prompted by the text is whether a teacher of mathematics also must be a philosopher and a historian of mathematics. Obviously, the answer is “no”, since lots of mathematics teachers have been successful in teaching mathematics throughout the times without these insights. On the other hand, since to philosophize is to reflect upon your acting, a mathematics teacher with philosophical insights possibly may highlight other issues that clarify mathematics than, e.g., a mathematician trained in science would do. Also, the connections between mathematics and philosophy have been close since antiquity onwards. There is, furthermore, hardly any place in mathematics teacher education to make room for courses in both history and philosophy of mathematics in addition to courses in mathematics.

It is my conviction that the means of analysis and results that I have presented above are important in mathematics education, and ought to be incorporated in mathematics education as far as possible.
ACKNOWLEDGEMENT

Thanks to Fredrik Engström for valuable comments on an earlier version of the paper.

REFERENCES


THE SWEDISH PENAL LAW ON ACCOUNTABILITY AND SEVERE MENTAL DISORDER

Christer Svennerlind

1 INTRODUCTION

The Swedish Penal Code in force before 1965 stipulated that a person who commits a criminal deed under the influence of mental disease, mental deficiency or other mental abnormality of such a deep-going nature, that it must be considered to be on a par with mental disease, shall be exempted from punishment. In fact, the exemption was from every penal law sanction, and this due to exemption from criminal liability.

In 1965, the Criminal Code replaced the Penal Code. This change of code involved changes with regard to penal law ideology, part of which concerned mentally disordered criminal offenders. The long and somewhat unwieldy phrase just mentioned was kept though. However, it did no longer refer to mental conditions exempting from criminal liability. From then on, what it did exempt from was the particular penal law sanction imprisonment. In the Swedish legal literature and debate, this is commonly referred to as the imprisonment prohibition of the Criminal Code.

The work on this article has been generously funded by the Bank of Sweden Tercentenary Foundation.

---

1 There is a terminological nicety to heed here. In present Swedish penal law terminology, technically, the word “punishment” (in Swedish, “straff”) refers exclusively to either imprisonment or fine. The other penal law sanctions – such as conditional sentence, probation, etc. – are referred to by the term “sanction” (in Swedish, “påföljd”). The latter term is also used as a general term, and as such it denotes the two punishments as well. In the quoted versions of the Penal Code, all penal law sanctions are covered by the term “punishment”.

2 Evidently, being exempted from criminal liability implies being exempted from penal law sanction, but not vice versa.

3 The Swedish name is “Brottsbalken”. Since the natural translation of “brott” is “crime”, “Criminal Code” seems a more appropriate rendering into English than “Penal Code”. In spite of perhaps going against the stream, I will henceforth use “Criminal Code”.

4 The Swedish name is “Strafflagen”.

5 Since the introduction of the Criminal Code, three versions of imprisonment prohibition have consecutively been in legal force.
In 1992, revisions were made once again of the regulation concerning mentally disordered offenders. Essentially, the wording of the regulation was changed in two ways. First, “under the influence of” was replaced by another phrase, which unfortunately is a bit difficult to translate into English when “under the influence of” is already used. Anyhow, the new phrase was intended to refer to a causal connection stronger than the one referred to by the former phrase. For want of anything better, perhaps “caused by” will do as a translation. Second, “mental disease, mental deficiency or other mental abnormality of such a deep-going nature, that it must be considered to be on a par with mental disease” was replaced by the more handy “severe mental disorder”\(^6\). It was not the unwieldy nature of the former that motivated the change. The new one is intended to express a narrower concept than that expressed by the former phrase.

The most recent revisions of the same regulation were made in 2008. One of these is the introduction of a requisite intended to make the imprisonment prohibition more restricted by comparison to the one it replaces. As indicated above, in Swedish penal law since 1965, accountability\(^7\) has not been a requisite for criminal liability. The revisions made in 2008 do not change that. However, the next step might be to reintroduce such a requisite. And if that is indeed made to happen, the wording may very well be the same that is now used for a slightly different purpose.\(^8\)

In what follows, the Swedish penal law regulation over time with regard to mentally disordered offenders is briefly surveyed and discussed.

\section*{2 BEFORE THE CRIMINAL CODE OF 1965}

Ch. 5 Sec. 5 of the Penal Code, as worded from 1946 to 1965, can be translated:

No one shall be held responsible for a deed, which he\(^9\) commits under the influence of mental disease, mental deficiency or other mental abnormality of such a deep-going nature, that it must be considered to be on a par with mental disease.

He who, through no fault of his own, temporarily has got into such a state that he was not in possession of his senses shall not be punished.\(^10\)

\(^6\)“Severe mental disorder” seems to be the most frequently used English translation. Another is “serious mental disturbance”, used in for example Ds 1999:36.

\(^7\)The Swedish term is “tillräknelighet”. “Accountability” is far from being a perfect translation. I suspect though that whichever English term is chosen, there will be connotations that bedevil its use.

\(^8\)The phrase in question can be translated “ability to understand the meaning of the act or to adjust the acting in accordance with such an understanding”.

\(^9\)Although only masculine pronouns are used in the Swedish laws quoted, women are also included.

\(^10\)This translation of Swedish law text is mine. The same holds for the ones that follow, as well as the translations made of other texts in Swedish. The Swedish original in this case runs:
In the first paragraph, there is no hint of why any of the three enumerated conditions – i.e., being mentally diseased, mentally deficient or mentally abnormal of such a deep-going nature, that it must be considered to be on a par with being mentally diseased – exempt from legal responsibility. In the second paragraph, there is such a hint: not being in possession of one’s senses. The content of the first paragraph is similar to that of the Durham rule, in force in the District of Colombia from 1954 to 1972, and still in force in New Hampshire. A formulation of the Durham rule is: “[A]n accused is not criminally responsible if his unlawful act was the product of a mental disease or mental defect.”\(^\text{11}\) Combined with the view that it is for psychiatry to decide whether the offender suffered from a mental disease or mental defect, this gives expression to what can be described as the medical model of legal insanity.\(^\text{12}\)

Another sense in which a penal law system can be described as implementing the medical model is if it does not insist on a causal connection between mental disease or defect and unlawful act, but still entrusts psychiatry the task of determining what is a mental disease or defect. Partly using the words just quoted: An accused is not criminally responsible if he, when committing the unlawful act, was suffering from a mental disease or mental defect.

If the version of Ch. 5 Sec. 5 of the Penal Code quoted above gives expression to the medical model in any form, this is so only in its first paragraph. The second paragraph seems to be of another kind. Maybe the legislator found it counterintuitive that a temporarily confused person had to be mentally diseased or defected to be exempted from criminal liability. It should suffice that he or she is simply confused, given that it is not of his or her own fault.

So far, it is the last version of Ch. 5 Sect. 5 of the Penal Code in legal force that has been presented. An earlier version of the same section can be translated:

> A deed, committed by someone who is insane, or deprived of the use of his intellect, owing to disease or weakness due to old age, shall be exempted from punishment.

---

\(^{11}\)Durham v. United States, 214 F. 2d 862, 875 (D.C. Cir. 1954).

\(^{12}\)Cf. (Moore, 2014). That it is for psychiatry to decide what is a mental disease is seen from Carter v. United States, 252 F 2d 608, 617 (D.C. Cir. 1957). Cf. also (Moore, 1984, p. 227), where New Hampshire judge Charles Doe is quoted: “Insanity […] is the result of a certain pathological condition of the brain […] and the tests and symptoms of this disease are no more matters of law than the tests or symptoms of any other disease in animal or vegetable life […]. What is a diseased condition of mind is to be settled by science and not by law.” Judge Doe was inspired by the views of Isaac Ray, whose 1838 book is an early plea for the medical model, cf. (Reznik, 1997, p. 26). New Hampshire adopted the product test, which is a version of the Durham rule, in the case State v. Jones, 50 N.H. 369 (1871). Doe had formulated it two years earlier in State v. Pike, 49 N.H. 399, 442 (1869).
Has someone, through no fault of his own, got into such a state of mental aberration that he was beside himself; the deed, which he commits in this unconscious state, shall be exempted from punishment.\textsuperscript{13}

Judging from the wording, in neither of its two paragraphs does this version give explicit expression to the medical model in any of its forms. Not more than a temporal connection between the abnormal state of mind and the deed seems to be sufficient though. This would make it differ from the version in force from 1946. However, even before 1946, a causal connection was required, notwithstanding the lack of any mention of it.\textsuperscript{14}

In 1942, a governmental committee, the Penal Code Commission,\textsuperscript{15} commented on the accountability rules. Among other things the committee asserted:

The wording the accountability rules were given in the Penal Code of 1864 has not been suited to form the basis for a firm and uniform legal use. Due to the combination of biological criteria (mental disease, physical disease, infirmity, mental derangement) and psychological (deprived the use of the intellect, lack the full use of the intellect, beside oneself),\textsuperscript{16} which has here come to use, it has not been possible in a concrete and for the application of the law unambiguous way to specify which mental states were intended [...]. The accountability rules therefore have turned out to be exceedingly elastic. During the time they have been in force, their applicability has repeatedly undergone changes, whereby the course of the development has been towards an expansion of the domain of the legal unaccountability. This expansion, which has been of far-reaching practical significance, has come about to such an extent,

\textsuperscript{13}The Swedish original:

Gärning, som begås av den, som är avvita, eller vilken förståndets bruk, genom sjukdom eller ålderssvaghet, är berövad, vare strafflös.

Har någon, utan egen skuld, råkat i sådan sinnesförvirring, att han ej till sig visste; vare ock gärning, den han i det medvetslösa tillstånd förövar, strafflös.

Quoted here is the result of editorial revisions made in 1890, when two sections were merged into one with the two former making up one paragraph each, and a spelling reform in the beginning of the twentieth century.

\textsuperscript{14}Cf. (SOU 1942:59, p. 83), where references are made to legal cases distinguishing between, on the one hand, causal influence and, on the other, mere temporal connection, and where the latter is considered insufficient for applying Ch. 5 Sec. 5 of the Penal Code then in legal force.

\textsuperscript{15}This committee worked from 1938 to 1956. One of its tasks was to analyse the accountability rules of the Penal Code. Some of its results were presented in its interim report (SOU 1942:59). Together with the final report of a parallel committee, (SOU 1953:17), its final report, (SOU 1956:55), formed the main basis of the Criminal Code of 1965.

\textsuperscript{16}Some of these criteria belong to Ch. 5 Sect. 6 of the Penal Code. While the fifth section stipulated exemption from criminal liability, the sixth section stipulated reduction of punishment; the latter due to reduced accountability. Until 1965, these two sections regulated what to do with mentally disordered criminal offenders.
which means, that the present application of the Penal Code Ch. 5 Sect. 5–6 has little in common with the original interpretation. Decisive for this development has first of all been the development of the psychiatric science and the increasing widening and deepening of the knowledge of disease phenomena, where abnormal states, which due to their symptoms, which are difficult to determine, or their slow course up to now have escaped attention, have now been considered.

When applying the accountability rules of the Penal Code, legal practice has essentially come to be determined by the medical experts. It is true that it has not been questioned, that the court should be completely free in relation to medical opinions presented to the court, instead jurists as well as physicians have stressed, that the expert opinion is only of an advisory nature [...]. However, concerning the psychological questions, which are hard to judge and often difficult for the not medically competent to understand, the courts have availed themselves only to a small extent of this possibility [to form their own opinion]. Owing to the way the forensic psychiatric investigations are organised, legal practice has in the main become dependent on the opinions, which at different times have been represented in the Board of Medicine.  

Though the quoted text is a bit elusive, I believe it is safe to say that we learn from it that the content of the accountability rules has changed over time; i.e., I presume their changed applicability should be correlated with a change in content. The changed content has led to an expansion of the extension of the rules. A cause of this, if not the main one, has been the continuous progress of psychiatry. In other words, psychiatry has over time supplemented the intension of the rules. So, if the committee is right in its conclusions, something similar to the medical model has been implemented, in spite of appearance. In the second paragraph of the last quotation, it is reported that it is the courts that apply the accountability rules, and, strictly speaking, that the psychiatric expert opinion is merely advisory. However, it is added that because of the complicated nature of these issues, the courts more or less always rely on the opinions presented by the psychiatric experts.

17 SOU 1942:59, p. 30. “Board of Medicine” is a translation of “Medicinalstyrelsen”. Until 1968, the Board of Medicine was the highest medical authority in Sweden.

18 To be exempted from criminal liability often, if not always, meant being turned over to a mental hospital. With an increase of the number of persons exempted the mental hospitals became somewhat overcrowded (SOU 1942:59, p. 40). My guess is that this is at least part of what is meant when the expansion of the extension is said to have been of far-reaching practical significance.
3 AFTER THE INTRODUCTION OF THE CRIMINAL CODE OF 1965

In 1965, the Criminal Code replaced the Penal Code. Though I think it is fair to say that some of the alleged consequences of this change have sometimes been exaggerated, the introduction of the Criminal Code marks a divide in the Swedish legal history. Essential parts of the change are the incorporation of special care among the penal law sanctions and the rejection of accountability as a requisite for criminal liability.

The legislative foundation of these changes is to be found in Ch. 33 Sects. 2–3 of the original version of the Criminal Code. These two sections are the nearest counterpart of Ch. 5 Sect. 5 of the Penal Code. The two sections run:

For a crime that someone has committed under the influence of mental disease, mental deficiency or other mental abnormality of such a deep-going nature, that it must be considered to be on a par with mental disease, no other sanction should be applied than being turned over to special care or, in cases specified in the second paragraph, fine or probation.

A fine should be imposed, if it is found suitable for preventing the defendant from committing further crimes. Probation should be imposed, in case such a sanction in view of the circumstances is found to be more suitable than special care; […].

If a sanction mentioned here ought not to be imposed, the defendant shall be exempted from sanction.

---

19 I am referring particularly to claims concerning the impact of a treatment ideology at the expense of that of general prevention, cf. (Svennerlind, 2009).

20 The second paragraph of Ch. 5 Sect. 5 of the Penal Code, concerning temporarily confused or unaccountable lawbreakers, has no counterpart in the Criminal Code. According to a statement in the government bill to the promulgation law of the Criminal Code, made by the Minister for Justice at the time, the second paragraph of the Penal Code section should continue to have legal force. A statement in a later government bill supports the claim that the second paragraph lives on, now as an unwritten rule. Notwithstanding that, the lack of an explicit, legislated counterpart of the second paragraph of the law section has resulted in an uncertainty concerning what actually is the legal situation in this respect. This uncertainty still remains (cf. Asp et al., 2010, pp. 400ff.).

21 Left out is a cross-reference to a section of another chapter.

22 Sect. 3 is the third paragraph of the quotation. In Swedish the wording of the two sections is:

För brott som någon begått under inflytande av sinnessjukdom, sinnesslöhet eller annan själslig abnormitet av så djupgående natur, att den måste anses jämställd med sinnessjukdom, må ej tillämpas annan påföljd än överlämnande till särskild vård eller, i fall som angivs i andra stycket, böter eller skyddstillsyn.

Till böter må dömas, om det finnes ändamålsenligt för att avhålla den tilltalade från fortsatt brottslighet. Skyddstillsyn må ådömas, därest sådan påföljd med hansyn till omständigheterna finnes lämpligare än särskild vård; […].

Finnes påföljd som här sagt sägs icke böra ådömas, skall den tilltalade vara fri från påföljd.
From this it can be concluded that committing a crime either under the influence of mental disease, mental deficiency or other mental abnormality of such a deep-going nature, that it must be considered to be on a par with mental disease, exempts from being sentenced to imprisonment. It does not exempt from being sentenced to other penal law sanctions.\footnote{The phrase "declared exempted from penal law sanction" might be used for a criminal offender who falls under the third paragraph, i.e., Section 3 of Ch. 33 of the Criminal Code. As said earlier, being exempted from penal law sanction does not imply being exempted from criminal liability.}

The special care, mentioned in the first paragraph, makes here its entry as a penal law sanction. According to what was the main rule, being handed over to this special care is the sanction that should be imposed. In the Swedish penal law system of today, this kind of care goes under the designation "forensic psychiatric care".

Interestingly, when the Criminal Code replaced the Penal Code, the same phrase\footnote{I.e., “mental disease, mental deficiency or other mental abnormality of such a deep-going nature, that it must be considered to be on a par with mental disease”} continued to be used. Since the issue of intent should be vital after 1965, one would expect there to be a change in its extension. This since intent was then, and still is, a general penal law requisite. As stated in Ch. 1 Sect. 2 Par. 1 of the Criminal Code of today: Unless otherwise stated, an act shall be regarded as a crime only if it is committed intentionally.\footnote{In Swedish: En gärning skall, om inte annat är särskilt föreskrivet, anses som brott endast då den begås uppsåtligen.} When persons who before 1965 would have been exempted from penal law sanctions, because of abnormal states of mind, become punishable, their states of mind become relevant in a way they were not before. \textit{En clair}, one would expect that, as a consequence of the intent requisite, those who, with the regulation in force before 1965, would have been exempted from penal law sanctions would be more numerous than those who, with the regulation in force after 1965, would be sentenced to special care; this because each and every member of the latter must have had the right kind of intent when breaking the law. However, judging from the claims of several legal scholars, the intent requisite does not have the expected effect on these relative numbers. The probable explanation of that being that in cases where the criminal offender is mentally disordered, the courts tend to make a more extensive interpretation of the intent requisite than they otherwise do. The reason being that the offender is considered to be in need of psychiatric care, and such care is conditional on intentional law breaking.\footnote{Cf. (SOU 1996:185, pp. 503 ff.); (Westin, 1994, p. 223); (Heckscher, 1985, p. 245); (Strömmerstedt, 1987, pp. 132 ff.).}

However that may be, already a few years after the introduction of the Criminal Code, a prevalent opinion was that the requisite – i.e., being mentally diseased, mentally deficient or mentally abnormal of such a deep-going nature, that it must be considered to be on a par with mental disease – had become over-generously construed. Too many psychopaths and personality disordered were...
included. The same had been part of what in 1946 had prompted the revision of the Penal Code. The kinds of psychopathy that were intended to be included in the extension of the requisite were then described as “[c]ertain exceptional cases of constitutional psychopathy, provided that the constitutional character anomaly was so pronounced that it must be considered to be constantly balancing on the verge of the psychotic.” The rather strict requisite outlined here had not been consistently applied in the medico-legal practice, or so the critics claimed. Also after the introduction of the Criminal Code, history repeated itself. The elusiveness of the third disjunct of the requisite, called “the equivalence requisite”, was the source of a patent lack of consensus among forensic psychiatrists.

In 1992, the concept of severe mental disorder was introduced as a remedy for the problems emanating from the concept adopted in 1946 and retained in 1965. No explicit definition was given of it though. In the government bill to the law there was instead a collection of examples of what should be classified as severe mental disorders.

As severe mental disorders should primarily be accepted conditions of psychotic character, consequently conditions involving deranged reality evaluation and with such types of symptoms as delusions, hallucinations and confusion. Moreover, in consequence of a brain lesion, a mental impairment of severe kind (dementia) with deranged reality evaluation and inability to orientate in life may result.

As severe mental disorder should also be accepted severe depressions involving contemplation of suicide. Furthermore, grave personality disorders with impulse breakthroughs (character disorders) should also be accepted as severe mental disorders, for example certain disabling neuroses and personality disorders with impulse breakthroughs of psychotic character.

Compulsory care should furthermore be actualized when a crisis reaction is of such a nature that the effect on the psychological functional level becomes so marked that it is of a psychotic kind.

As severe mental disorder should also be classified the alcohol psychoses, such as delirium tremens, alcoholic hallucinoses and evident conditions of dementia. The same holds for the psychoses that drug addicts can contract. Also in other situations when a drug addict is in a state of severe confusion and it is evident that his physical health or his life is in danger, compulsory care should be an option. In certain cases also a state of abstinence can be so grave that it under a short time must be described as a severe mental disorder. It goes without saying that a severe addiction that only has grave

---

physical complications should not lead to compulsory psychiatric care.\textsuperscript{29}

Not mentioned here, but referred to elsewhere in the same government bill, is the distinction between the \textit{kind} and \textit{degree} respectively of a mental disorder. It is stated that certain mental disorders are always severe with regard to kind, but need not be so with regard to degree. Schizophrenia is the example mentioned.\textsuperscript{30} Depressions are mentioned as mental disorders that need not be severe with regard to kind, but can be so with regard to degree. There is also a third possibility, which is being severe with regard to kind as well as to degree. No example is given though. Anyhow, according to the government bill, both kind and degree need to be weighted into the assessment of a mental disorder as severe.\textsuperscript{31}

It is evident from the four quoted paragraphs that psychoses are considered to be special when it comes to elucidate what is aimed at, or should be aimed at, by “severe mental disorder”. Giving rise to a deranged reality evaluation, manifested as delusions, hallucinations or confusion, is the primary condition to meet for a mental disorder to qualify as a severe mental disorder. Over time, the tendency has also been to identify severe mental disorder with psychosis. Nowadays, in the Swedish penal law system, in forensic psychiatric examinations and judicial decisions, schizophrenia is the diagnosis that is often, if not always, declared to be a severe mental disorder.\textsuperscript{32} Most certainly, schizophrenia is a psychosis.

More or less worldwide, psychoses have a somewhat special status. For example in Norwegian penal law, the term “psychosis” and its derivations are used, instead of “severe mental disorder” or anything like that. Sect. 44 Par. 1 of the Norwegian Penal Code stipulates: “A person who was psychotic or unconscious at the time of committing the act shall not be liable to a penalty.”\textsuperscript{33} To be noticed here, besides the use of “psychosis”, is the lack of any causal requisite. While the original Swedish Penal Code apparently lacked a causal requisite, the Norwegian Penal Code in fact does so.

4 \textbf{AFTER THE REVISIONS MADE IN 1992 AND 2008}

In 1995, the Supreme Court passed its sentence in the legal case NJA 1995 s. 48. The relevant section of the Criminal Code at the time, regulating what to do

\textsuperscript{29}(Prop. 1990/91:58, p. 86)
\textsuperscript{30}Judging from the quotation, dementia is also a severe mental disorder by kind.
\textsuperscript{31}(Prop. 1990/91:58, p. 87)
\textsuperscript{32}Cf. (Borgeke, 2012, pp. 304ff.); (RMV-Rapport 2013:1).
\textsuperscript{33}The translation is from (Moore, 2014). In Norwegian the wording is: "Den som på handlingstiden var psykotisk eller bevisstløs straffes ikke.” As seen here, the Norwegian Penal Code uses the term “bevisstløs”. The equivalent “medvetslös” was used in the Swedish Penal Code until 1946. In the Swedish case, the German “unbewusst” probably inspired the choice of term, cf. (Sondén, 1930, pp. 75f).
legally with severely mentally disordered criminal offenders, and which had been so since 1992, was Ch. 30 Sec. 6. Its wording was:

A person who commits a crime caused by a severe mental disorder may not be sentenced to imprisonment. If in such a case the court also considers that no other sanction should be imposed, the accused shall go free from sanction.\(^{34}\)

The perpetrator in the legal case NJA 1995 s. 48, second lieutenant Mattias Flink, was in a sense found to be severely mentally disordered. The proviso is motivated since the use of the term “severe mental disorder” is made difficult by various complications. One of these, perhaps the main one, being that, in contradistinction to “mental disorder”, the term “severe mental disorder” is a legal term. Due to the similarity between the two, this is easily missed, or misunderstood. Another complication is that it is marred by a certain ambiguity. The latter complication seems to have emerged due to the legal case NJA 1995 s. 48.\(^{35}\) Flink was found guilty as charged; for murder in seven cases and attempted murder in three cases. According to the Supreme Court, he fulfilled the conditions for being severely mentally disordered at the time of the crimes in the following sense. If he had been psychiatrically examined at the time of the crimes, he would have been found to fulfil the conditions for compulsory psychiatric care. When the concept of severe mental disorder was adopted in 1992, it was explicitly declared that it should be one and the same within the frame of the Criminal Code as within that of the Compulsory Mental Care Act. From that would follow then the conclusion that Flink should be exempted from imprisonment. Since he was found not to be in need of any psychiatric care at the time of the trial, to sentence him to forensic psychiatric care was excluded as well.\(^{36}\) The remaining penal law sanctions then left would be fine and probation. Thus, he would more or less go free, after having killed seven persons with an automatic carbine and wounded three more with the same weapon.

According to the Supreme Court, Flink had self-induced his mental disorder, by drinking hard liquor and reacting to that in a way he ought to have foreseen. Note though that according to one of the consulted psychiatric experts, Flink had been psychotic already before he allegedly self-induced his mental disorder. That would have blocked the self-induction charge. However, the Supreme Court dismissed that, saying a bit cryptically that “the inquiry [cannot] with necessary clarity be said to show whether Flink suffered from a severe mental

\(^{34}\)In Swedish:

Den som har begått ett brott under påverkan av en allvarlig psykisk störning får inte dömas till fängelse. Om rätten i ett sådant fall finner att inte heller någon annan påföljd bör ädömas, skall den tilltalade vara fri från påföljd.

\(^{35}\)Cf. (Borgeke, 2012, p. 308).

\(^{36}\)This follows from Ch. 31 Sec. 3 of the Criminal Code. It is a rather long section. It does not seem necessary to quote it here.
disorder before the emergence of the psychotic state or that he is now suffering from such a disorder”. In the end, Ch. 30 Sec. 6 of the Criminal Code was declared not to apply to Flink. Hence, the sentence was imprisonment for life.38

After this judicial decision, it became obvious that the term “severe mental disorder” has two slightly different meanings: one within the frame of penal law, the other within the frame of the legislation regulating compulsory psychiatric care. The former takes into consideration the aetiology of an occurrence of a mental disorder when determining whether it is a severe mental disorder or not. The latter does not do so.

In 2008, in the aftermath of the legal case NJA 1995 s. 48, revisions were made of Ch. 30 Sec. 6 of the Criminal Code. One aspect of the revisions is evidently to be seen in the light of that legal case. There is also another. I am referring to the call for reintroducing accountability as a general requisite for criminal liability. The revision made in 2008 is not more than a step though in the direction of carrying out the reintroduction of such a requisite. Hence, among the national penal law systems,39 the Swedish system continues to be of its own kind; i.e., a penal law system that does not include accountability as a requisite for criminal liability.40

Since mid-year 2008, the wording of Ch. 30 Sec. 6 of the Criminal Code is the following.

A person who has committed a crime caused by a severe mental disorder shall primarily be sentenced to another sanction than imprisonment. The court may sentence to imprisonment only if there are special reasons. When judging whether there are such reasons the court shall pay regard to

1. whether the crime is highly culpable,

---

37 The same consulted psychiatric expert claimed that Flick was at the trial still suffering from a severe mental disorder, and accordingly in need of psychiatric care. The “now” in the quotation refers to the time of the trial.

38 If the Supreme Court had found Flink to be temporarily deranged through no fault of his own, he should have been acquitted. This in accordance with the unwritten rule mentioned earlier. At least in theory, there was also the possibility of ruling that the crimes were perpetrated independently of Flink’s abnormal mental state. That would have left the Supreme Court free to sentence him to imprisonment – given that the other requisites, among which is the intent requisite, were also met.

39 The penal law systems of four states of the United States of America – viz., those of Idaho, Kansas, Montana, and Utah – do not allow the insanity defence. In that respect, they resemble the Swedish penal law system.

40 Notwithstanding that, Sweden in 2002 ratified the Rome Statute of the International Criminal Court. According to Article 31 of that Statute, unaccountability is a ground for excluding criminal responsibility: “In addition to other grounds for excluding criminal responsibility provided for in this Statute, a person shall not be criminally responsible if, at the time of that person’s conduct[,] [...] [t]he person suffers from a mental disease or defect that destroys that person’s capacity to appreciate the unlawfulness or nature of his or her conduct, or capacity to control his or her conduct to conform to the requirements of law.”
2. whether the defendant lacks or has a limited need for psychiatric care,
3. whether the defendant in connection with the crime has himself caused his condition by intoxication or by any other similar means, or
4. the other prevailing circumstances.

The court may not sentence to imprisonment, if the defendant as a consequence of the severe mental disorder has had no ability to understand the meaning of the act or to adjust his acting in accordance with such an understanding. This does not apply though if the defendant himself has caused his inability in the way described in the first paragraph.

If the court in cases referred to in the first or second paragraph finds that no other sanction ought to be imposed, the defendant shall be free from sanction. 41

The imprint of the law case NJA 1995 s. 48 on this law section is evident in its first paragraph. Obviously, the items of the list of presumption breaking reasons reflect the “needs” of the Supreme Court in that law case.

In the second paragraph, there is also a new requisite found. It is of a disjunctive form and shows similarities to various tests of legal insanity; e.g., the one proposed by the American Law Institute, variants of which have been adopted from the 1960’s and onwards by several American state jurisdictions, as well as that of the federal jurisdiction. After John Hinkley’s attempt on the life of Ronald Reagan, in 1981, the test was repealed though by several of these states, as well as by all the federal courts. As formulated by the American Law Institute in 1962, the test runs: “A person is not responsible for criminal conduct if at the time of such conduct as the result of mental disease or defect he lacks substantial capacity either to appreciate the criminality of his conduct or to conform his

41 In Swedish:

Den som har begått ett brott under påverkan av en allvarlig psykisk störning ska i första hand dömas till en annan påföljd än fängelse. Rätten får döma till fängelse endast om det finns synnerliga skäl. Vid bedömningen av om det finns sådana skäl ska rätten beakta
   1. om brottet har ett högt straffvärde,
   2. om den tilltalade saknar eller har ett begränsat behov av psykiatrisk vård,
   3. om den tilltalade i anslutning till brottet själv har vållat sitt tillstånd genom rus eller på något annat liknande sätt, samt
   4. omständigheterna i övrigt.

Rätten får inte döma till fängelse, om den tilltalade till följd av den allvarliga psykiska störningen har saknat förmåga att insä gärningens innebörd eller att anpassa sitt handlande efter en sådan insikt. Detta gäller dock inte om den tilltalade har vållat sin bristande förmåga på det sätt som anges i första stycket.

Om rätten i fall som avses i första eller andra stycket finner att någon påföljd inte bör dömas ut, ska den tilltalade vara fri från påföljd.
conduct to the requirement of law. Although similar to this test, the requisite of the second paragraph of Ch. 30 Sect. 6 of the Criminal Code is a requisite for the penal law sanction imprisonment, not for criminal liability.

Thus, as regards the imprisonment prohibition, it is no longer sufficient that the law-breaking act is caused by a severe mental disorder; the person must also lack the ability to understand the meaning of the act or to adjust his or her acting in accordance with such an understanding. If these abilities are intact, the court is permitted to sentence the perpetrator to imprisonment, even if the act is caused by a severe mental disorder. In other words, the imprisonment prohibition applies to whoever, through no fault of his or her own, and as a consequence of a severe mental disorder, lacked the ability to understand the meaning of the act or to adjust his or her acting in accordance with such an understanding. Evidently, implied here is that severe mental disorder need not lead to a lack of either ability to understand or to adjust one’s acting in accordance with such an understanding.

There are several governmental reports proposing that accountability be reintroduced as a requisite for criminal liability. A reasonable guess is that, as was the case when the Criminal Code replaced the Penal Code, the same phrase will be retained. This time the phrase is “has had no ability to understand the meaning of the act or to adjust his or her acting in accordance with such an understanding”. And it will then be referring to a requisite for being exempted from criminal liability, instead of, as now, a requisite for being exempted from the penal law sanction imprisonment.

Consider once more the accountability requisite of the Penal Code, the one last being in legal force: No one shall be held responsible for a deed, which he commits under the influence of mental disease, mental deficiency or other mental abnormality of such a deep-going nature, that it must be considered to be on a par with mental disease. As described earlier, more or less the same wording was retained when in 1965 the Criminal Code replaced the Penal Code. The difference being that instead of exempting, when fulfilled, from criminal liability, in its later version it exempted from the penal law sanction imprisonment. Let us classify a requisite as the latter one as being an imprisonment requisite. Recall also that the imprisonment requisite that came into legal force in 1992 was intended to be narrower than the one it replaced. It seems safe now to say that if the two had been accountability requisites, the application of the 1992 version would have exempted fewer from criminal liability than the former. The imprisonment requisite of 2008 is intended to be even narrower than that of 1992. Hence, transformed into an accountability requisite it would exempt fewer from criminal liability than the requisite of 1992 likewise transformed.

A conclusion to draw from the reasoning above is that the notion of accountability, based on the imprisonment requisite of 2008, is another than the one

---

43 The two latest are (SOU 2002:3) and (SOU 2012:17).
that was in legal force until the introduction, in 1965, of the Criminal Code. This conclusion is based on, what might be described as, transitivity reasons.

5 A COMMENT ON THE CONCEPT OF SEVERE MENTAL DISORDER

Ponder upon what was said earlier about the notion of severe mental disorder. Recall the collection of examples of what should be classified as severe mental disorders. The main emphasis is on psychoses and psychosis-like states. What is so special with those states? I suggest that they exemplify what being unaccountable is. Being in such states compromises one’s accountability. This makes sense of the fact that a severely mentally disordered lawbreaker is given psychiatric care, instead of being sent to prison. It also makes sense of the reluctance to accept forensic psychiatric care as a penal law sanction.

If what is hinted at in the previous paragraph is correct, it has an interesting implication. The implication I am thinking of is that it would be incoherent to hold a person severely mentally disordered but still accountable. Yet, that is what might become the position of the Swedish penal law, if the requisite of the second paragraph of Ch. 30 Sect. 6 of the Criminal Code now in legal force would be transformed into an accountability requisite.

REFERENCES


44 For a similar view, cf. (Malmgren et al., 2010, p. 70).
45 I thank Jan Almäng, Tova Bennet, Helge Malmgren and Susanna Radovic for valuable comments on an earlier version of this paper.


Prop. 1964:10. Kungl. Maj:ts proposition till riksdagen med förslag till lag om införande av brottsbalken m.m.

Prop. 1990/91:58. Om psykiatrisk tvångsvård m.m.

Prop. 1993/94:130. Ändringar i brottsbalken m.m. (ansvarsfrihetsgrunder m.m.)


SOU 1942:59. Strafflagberedningens betänkande angående strafflagens tillräknelighetsregler, sinnesundersökning m.m.


SOU 1956:55. *Skyddslag*.


PERSUASIVE DEFINITIONS AND SOCIAL EXTERNALISM

Anders Tolland

Starting on a personal note: The title of this paper in no way implies that persuasive definitions is something I associate with Christian; I cannot remember him ever using one. As for social externalism, I have, on rare occasions, tried to contradict him on logical matters, but most of the time I have had common sense enough to defer to his excellent authority.

We are all familiar with persuasive definitions. In latterday Swedish politics the expression “real jobs” [“riktiga jobb”] has been thrown around a lot. Or, to construct a really extreme example, we can imagine a North Korean ambassador saying: “Real democracy is when The Great Leader decides according to the true interest of the people.”

It has been claimed that all definitions are, by their nature, persuasive. Well, it might be true that most definitions are made as a part of attempts to convince others, but I believe that a more traditional, narrower, characterisation of persuasive definitions serves us better.

The *locus classicus* is Charles L. Stevenson’s *Ethics and Language*. He describes persuasive definitions as being used on expressions that have both (substantial) evaluative and descriptive content. A persuasive definition changes the descriptive content, but keeps the established (usually positive) evaluative content, and attaches it to the new descriptive content.

Stevenson wedds this to his particular theory of meaning, but we do not need to follow him there. Stevenson’s characterization probably works with most versions of descriptive/cognitive vs evaluative/normative meaning, and what is said here does not depend on any particular meaning theory.

We have all encountered persuasive definitions, and the most likely explanations of their being used at all is that they work sufficiently well to be worth attempting. The question is: How can persuasive definitions work? Why is not our reaction to them always like Alice’s, when Humpty Dumpty redefines

---

1 Cf. (Walton, 2005).
2 (Stevenson, 1944, Ch IX)
“glory”. Stevenson talks about the “inertia” of the evaluative content, but does not explain where this inertia comes from.

Another suggestion is that the descriptive content is vague. This is certainly true of some persuasive definitions. Essentially contested concepts like 'democracy', 'freedom', 'justice' are especially apt for something like persuasive definitions. But vagueness hardly is a necessary condition, and it does not give us the “mechanics” of persuasive definitions.

How persuasive definitions actually can work is in the end something to be answered by empirical research, not from the philosophical armchair, but this paper presents an hypothesis, viz. that social meaning-externalism can, to a substantial degree, explain how persuasive definitions work.

The most well-known variety of externalism about linguistic meaning is causal externalism. “Meaning” can, of course, mean a lot of things. We are here concerned with the meaning of referring expressions, and the central point is that the meaning of an expression is what decides what it refers to.

It is called “externalism” because it claims that reference is, to a large extent, decided by things that are external to the individual language user, i.e., what is in her mind/head is not sufficient to determine what the expression actually refers to.

Saul Kripke has applied this idea to the meaning of proper names; what they refer to is determined by the (proper kind of) causal chain from an original name-giving (e.g., a baptism) to the present usage. And Hilary Putnam has applied it to natural kind concepts like water, elm, gold, tiger. In Putnam’s version reference is fixed by actual, essential similarity to indexically indicated examples. For ’water’ our examples are the liquid that flows in rivers and lakes, that rains from clouds, etc, and to be water is to share the common essential nature of these examples, which according to our present theories is to be $H_2O$.

In the same paper Putnam introduces another kind of externalism about linguistic meaning, social externalism, and with it another notion, the linguistic division of labor.

To illustrate this Putnam’s example ‘molybdenum’ works fine on me too. Once all I knew about molybdenum was that it was ‘some kind of chemical stuff’. But even in this state of utter ignorance I could still talk about (refer to) molybdenum. If, for example, I had happened to overhear some people discussing what material to use to construct a certain item, I might have, for some odd reason, ventured the comment “It’s best to make it out of molybdenum.” In

---

3(Kripke, 1980)
4(Putnam, 1975)
5Molybdenum happens to be a natural kind, but that is coincidental; causal and social externalism are independent of each other.
all probability, this comment would only have rendered me stares of surprise and/or contempt, and the answer: “Out of molybdenum! You don’t know what you’re talking about.” This would have been perfectly true; I did not know what I was talking about. But I still actually would have talked about it, viz. about molybdenum. One of the reasons that what I said was so singularly stupid was that I, despite my ignorance, managed to refer to molybdenum (and suggest that it should be used).

How is this possible? Knowing only that it is some chemical stuff, how can I succeed in referring to exactly molybdenum? I have no information that pinpoints precisely this substance; I was not even acquainted with any examples of molybdenum – actually, today I am still not. So what determined the reference when I said “molybdenum”?

Putnam’s answer is linguistic division of labor. We are speaking a natural, common, language, which means, on the one hand, that one has to abide by rules that one has not constructed oneself, but, on the other hand, that one can avail oneself of some common resources of the language community.

I intended to use “molybdenum” according to the correct standard of my linguistic community; that, together with my minimal knowledge, is sufficient to make me able to refer to molybdenum. The reference is fixed, not by what I know, but by the competence of experts connected to my linguistic community (and we also rely on these experts for a substantial description of what molybdenum is). When I used “molybdenum”, intending to use it in the standard way, I implicitly deferred to these experts.

This is the linguistic division of labor. Reference-fixing meaning need not be possessed by each individual speaker. For many terms all that the everyday speaker needs is some minimal knowledge and implicit deferring to the language community and its experts. The ordinary user’s ability to refer is parasitic on the reference fixing competence of the experts.

This is also social externalism. “Externalism” because reference is to a large extent, fixed by things outside of the ordinary, individual language user; “social” because it is fixed by other persons that one defers to.

The theories of social externalism and a linguistic division of labor entail that there is a number of expression in rather wide use whose descriptive content and extension is, at best, only vaguely known to the ordinary member of the language community; the specifics are deferred to the experts.

My hypothesis is that persuasive definitions (or at least a substantial part of them) make use of this phenomenon. Persuasive definitions can be used on expressions where the descriptive content actually entertained by ordinary language users is very thin, or unclear, or only held with little conviction, and where these ordinary users are prepared to defer to experts.

In such a situation the one proposing the persuasive definition, the persuader, is making a claim – implicit or not – to be an expert, and to be introducing the
correct, specific, descriptive content of the expression. Unlike Humpty Dumpty, the persuader doesn’t attempt to change a clearly recognized descriptive meaning. He fills in gaps or uncertainties. And, unlike Alice with “glory”, the persuaded is prepared to accept this because it is a part of the meaning susceptible to deferring.

If this is to be an even prima facie credible account of persuasive definitions, the scope of linguistic division of labor obviously must be extended, and I do take it that linguistic division of labor is very widespread indeed. This cannot be proven here. I restrict myself to three short comments to support this position.

First, social externalism is not an all or nothing matter. The extent to which the meaning of a notion depends on deferring is a matter of degree, to a large extent in the case of ‘molybdenum’, to some, but a much lesser, extent for ‘water’.

Secondly, we try to speak a natural language correctly, and a natural language is a collective enterprise. The individual user rarely gets to decide on her own what an expression means. We are quite used to follow the lead of other speakers, even to accept being corrected by others.

Thirdly, I believe that the linguistic division of labor is just the linguistic aspect of a fundamental and quite pervasive phenomenon, that I would like to call the cognitive or epistemic division of labor. What you know, or have good reason to believe, is massively dependent on the testimony of others. Almost everything you know, you know by relying on others. A test of what you really know all on your own – a test not as metaphysically extravagant as Descartes’ deceiving demon (or The Matrix) – is to think about what would not be in doubt if you were to take seriously the possibility that you are on The Truman Show. Our fundamental dependence on others gives those that wish to engage in persuasive definitions room for manoeuvre.

The upshot of these points is that persuasive definitions are only a very small aspect of a huge phenomenon, the epistemic division of labor. The linguistic division of labor should be seen only as a part of this larger phenomenon, but, on the other hand, this connection implies that the parts of language affected by social externalism is wider than what is presented by Kripke and Putnam.

REFERENCES


PAIR GRAMMARS AND COMPOSITIONALITY

Dag Westerståhl

1 INTRODUCTION

In traditional Generative Grammar, or at least since Montague, a grammar generates well-formed syntactic expressions, and a semantics maps these onto meanings. If the mapping is a homomorphism, the semantics is compositional. Many modern linguists, however, take grammars to generate both expressions and meanings at the same time. More exactly, the grammar generates triples

\[ \langle e, X, m \rangle \]

where \( e \) is a string, \( X \) is a syntactic category, and \( m \) is a meaning.\(^1\)

Semanticists who prefer this format appear to view it as a more expedient version of compositional semantics. In Pauline Jacobson’s words: “[…]

the Generative-Semantics-style solution is a priori simpler [than LF approaches] in that it allows the compositional syntactic rules […] to be stated in tandem with the compositional semantic rules—and this in turn simplifies the statement of the latter.” (Jacobson, 2002, p. 602).

However, there really are two ways of taking syntax and semantics to be generated ‘in tandem’. One is the idea that to each syntactic rule corresponds a semantic rule; with a few extra assumptions, this becomes the traditional account of compositionality. The other is the idea that the grammar generates triples as above. It is not obvious that these are two ways of saying the same thing. In fact, *prima facie* the ‘triple’ approach looks more powerful; for example, it enables a simple symmetric account of synonymy and ambiguity. Also, it would seem to permit grammar rules of a kind not easily reducible to the more traditional format. But my focus in this note is on compositionality:

- What exactly does it mean that a ‘triple grammar’ is compositional?

\(^1\)Alternatively, \( e \) could be a phonological representation, and \( m \) could be a meaning representation (in some formal language) rather than a model-theoretic object.
• Is compositionality for ‘triple grammars’ just a notational variant of the classical notion of compositionality?

The general theory behind the new grammar format has not, to my knowledge, attracted much attention by working semanticists – with one exception. The exception is Marcus Kracht, who in (Kracht, 2007) and more fully in the book (Kracht, 2011) discusses (among other things) relevant notions of compositionality in great detail. Still, he does not really compare these to the classical idea of compositionality, so he doesn’t address the second question above. To do that is the object of the present note.

To begin I will, following (Kracht, 2011), forget about syntactic categories and consider pair grammars, which generate expression–meaning pairs. As Kracht also notes, such pairs can be seen as abstract versions of de Saussure’s signs, combining a signifiant and a signifieé. But what is Saussurean compositionality?

Kracht (2011) reserves the term compositionality in the pair format for just one of several related notions that he defines. The idea is simple. Traditionally:

(PC) A grammar + semantics is compositional if the meaning of a complex expression is determined by the meanings of its immediate parts and the mode of composition.

And here is a Saussurean reformulation:

(PC²) A pair grammar is compositional if the meaning of a complex sign is determined by the meanings of its parts and the mode of composition.

Every sign has a unique meaning (i.e. its second element), so (PC²) makes good sense. But shouldn’t the first element of a complex sign, or perhaps both elements, also be thus determined? Kracht has different terms for these notions of compositionality, and we will see how they are related.

Kracht’s account is mathematically precise, so for comparison I need an equally precise account of traditional compositionality. It is natural to choose the one in (Hodges, 2001). In fact, the two have many similarities. Both are mathematical, both rely on partiality rather than primitive syntactic categories, and both use term algebras to represent derivations of expressions.²

²But not (Kracht, 2007).
³Kracht argues that this is motivated for principled reasons, and indeed proves a theorem showing how syntactic categories can, under certain assumptions, be dispensed with.
⁴There are many aspects of Kracht’s work I will not deal with. In particular, he is not only interested in abstract grammar formats and notions of compositionality, but also in their concrete implementations. This leads him to impose strong constraints both on what meanings are and on what kind of string manipulations a grammar is allowed to perform. Those constraints are irrelevant, however, to the points made here.
A partial algebra is a structure

\[ E = (E, f^E)_{f \in \Sigma} \]

where \( E \) is a non-empty set and for each function symbol \( f \) in the signature \( \Sigma \), \( f^E \) is an \( n \)-ary partial function on \( E \) for some \( n \). One may allow 0-ary functions symbols denoting elements of \( E \), but in the present linguistic context, it is convenient to collect these elements in a set \( A \subseteq E \) of atoms, and write

\[ E = (E, A, f^E)_{f \in \Sigma} \]

where the function symbols in \( \Sigma \) have arity \( \geq 1 \). Starting from \( A \) and repeatedly applying the functions yields a subset \([A]\) of \( E \). \( E \) is generated if \([A] = E\).

For example, \( E \) could be a set of strings, generated from atomic strings with concatenation and perhaps other operations. Then \( E \) really is a grammar, and the functions \( f^E \) are the grammar rules. But we will want to look at different algebras of the same signature. One such algebra is the corresponding partial term algebra \( GT_E \). Terms record how elements of \( E \) are generated from \( A \). If

\[ e = f^E(a, g^E(h^E(b, c))) \]

where \( a, b, c \in A \), then the term

\[ f(a, g(h(b, c))) \]

consisting of the function symbols and parentheses shown, records this particular derivation. If \( e \) can be derived in another way, say, \( e = p^E(a', q^E(b')) \), where \( a', b' \in A \) and \( p, q \in \Sigma \), then a distinct term

\[ p(a', q(b')) \]

records that derivation. Formally, we define the set \( GT_E \) of grammatical terms of \( E \), simultaneously with a mapping \( val \) from terms to their ‘values’ in \( E \):

1. If \( a \in A \) then \( A \in GT_E \) and \( val(a) = a \).
2. If \( f \in \Sigma \) is \( n \)-ary, \( t_1, \ldots, t_n \in GT_E \), \( val(t_i) = e_i \) for \( i = 1, \ldots, n \), and \( f^E(e_1, \ldots, e_n) \) is defined, then

\[ f(t_1, \ldots, t_n) \]

belongs to \( GT_E \), and \( val(f(t_1, \ldots, t_n)) = f^E(e_1, \ldots, e_n) \).

If \( E \) is generated, the mapping \( val \) is a surjective homomorphism, since

\[ val(f(t_1, \ldots, t_n)) = f^E(val(t_1), \ldots, val(t_n)) \]

\( GT_E \) is the domain of the partial term algebra.
\[ \text{GT}_E = (\text{GT}_E, A, f^{\text{GT}_E})_{f \in \Sigma} \]

where \( f^{\text{GT}_E} \) is now the syntactic operation that yields the term \( f(t_1, \ldots, t_n) \) from the arguments \( t_1, \ldots, t_n \) (when defined).

The crucial difference between \( E \) and \( \text{GT}_E \) is that terms are uniquely readable, while the same element of \( E \) may be derived in different ways. Also, the subterm relation, i.e. the constituency relation between terms, is clear. I use the following notation: if a term \( s \) in \( \text{GT}_E \) is written

\[ s[t_1, \ldots, t_k] \]

this indicates that \( t_1, \ldots, t_k \) are disjoint subterm occurrences of \( s \). \( s[u_1, \ldots, u_k] \) is then the result of replacing each \( t_i \) by \( u_i \) in \( s \); a result which may or may not be in \( \text{GT}_E \).

With these concepts in place, we now quickly present Hodges’ and Kracht’s notions of grammar and language. In what follows, \( E \) will be a set of entities called expressions, and \( M \) a set of entities called meanings. One may think of expressions as strings, but in fact the nature of expressions as well as meanings is immaterial for my purposes in this note.

3 GRAMMARS AND SEMANTICS

3.1 HODGES STYLE GRAMMARS AND SEMANTICS

A grammar for \( E \) is a partial algebra \( E = (E, A, f^E)_{f \in \Sigma} \) generating \( E \). The functions \( f^E \) can be thought of as grammar rules, and the atoms as lexical items.

Since expressions may be structurally ambiguous (derivable in different ways; cf. old men and women), meanings are assigned to terms. Thus, a semantics for \( E \) is a partial function \( \mu \) from \( \text{GT}_E \) to \( M \).

To deal with lexical ambiguity, i.e. an atom \( a \) in \( A \) having several meanings \( m_1, m_2, \ldots \), one can modify the term algebra slightly: introduce corresponding atomic terms, say, \( \overline{a}_1, \overline{a}_2, \ldots \) with \( \mu(\overline{a}_1) = m_1, \mu(\overline{a}_2) = m_2, \ldots \), and \( \text{val}(\overline{a}_1) = \text{val}(\overline{a}_2) = \ldots = a \). But note that this changes the signature. If we want the signature to remain the same, the only way is to start with a grammar that (by supplying enough atoms) has no lexical ambiguities.

Corresponding to \( \mu \) we have the partial equivalence relation

\[ t \equiv_\mu u \iff \mu(t) \text{ and } \mu(u) \text{ are both defined, and } \mu(t) = \mu(u), \]

on \( \text{GT}_E \). This is the notion of \( \mu \)-synonymy.

3.2 KRACHT STYLE PAIR GRAMMARS

An interpreted language \( L \) is simply a subset of \( E \times M \), i.e. a set of Saussurean signs. Put differently, \( L \) is a relation between expressions and meanings. We
assume that \( \text{dom}(L) = E \) (every expression has at least one meaning) and \( \text{range}(L) = M \) (every meaning is the meaning of at least one expression).

In this format one can account for ambiguity and synonymy without involving derivations or terms. Each sign has a unique meaning (and a unique expression). An expression \( e \) is ambiguous if there are distinct \( m, m' \) such that \( \langle e, m \rangle, \langle e, m' \rangle \in L \). For example, we can have ambiguous lexical items, such as \( \langle \text{bank}, m_1 \rangle, \langle \text{bank}, m_2 \rangle \) in \( L \), as well as (structurally) ambiguous complex strings: \( \langle \text{old men and women}, m \rangle, \langle \text{old men and women}, m' \rangle \). Dually, two expressions \( e \) and \( e' \) are synonymous if there is \( m \) such that \( \langle e, m \rangle, \langle e', m \rangle \in L \).

A grammar for \( L \) is just like a grammar for \( E \), except that now pairs are generated. Thus, a pair grammar is an algebra

\[
L = (L, A_L, f^L)_{f \in \Sigma}
\]

where \( A_L \subseteq L \) is a set of atomic pairs, each \( f^L \) is a partial function from \( L^n \) to \( L \), for some \( n \geq 1 \), and \( L \) is generated from the atoms by the \( F^L \). 5

4 COMPOSITIONALITY

4.1 CLASSICAL COMPOSITIONALITY

In the Hodges style framework, where \( \mu \) is a semantics for \( E \), there is essentially only one rendering of the idea that the meaning of a complex expression is determined by the meanings of its (immediate) parts and the mode of composition. It has two formulations.

1. \( \mu \) is compositional iff for each \( f \in \Sigma \) there is an operation \( r_f \) such that if \( f(t_1, \ldots, t_n) \in \text{dom}(\mu) \), then

\[
\mu(f(t_1, \ldots, t_n)) = r_f(\mu(t_1), \ldots, \mu(t_n))
\]

The operation \( r_f \) is partial: we take \( r_f(\mu(t_1), \ldots, \mu(t_n)) \) to be undefined when \( f(t_1, \ldots, t_n) \not\in \text{dom}(\mu) \).

2. A partial equivalence relation \( \equiv \) on \( \text{GT}_E \) is compositional if, whenever \( s[t_1, \ldots, t_k] \) and \( s[u_1, \ldots, u_k] \) are both in \( \text{dom}(\mu) \), and \( t_i \equiv u_i \) for \( i = 1, \ldots, k \), we have

\[
s[t_1, \ldots, t_k] \equiv s[u_1, \ldots, u_k]
\]

(1), but not (2), presupposes the Domain Principle (DP): \( \text{dom}(\mu) \) is closed under constituents (subterms). An easy inductive argument shows:

---

5Kracht calls the grammar an interpreted grammar. He makes a distinction I ignore here: if you allow the partial functions \( f^L \) to be defined and have values in \( E \times M \), or (as here) only in \( L \). In the former case, non-expressions (and non-meanings) can fall under grammar rules, as when an English speaker ‘knows’ how to express the past tense of a nonsense verb.
Fact 1 (Hodges).
If DP holds, $\mu$ is compositional if and only if $\equiv\mu$ is compositional.

4.2 FUNCTIONAL SAUSSUREAN COMPOSITIONALITY

Let $L = (L, A_L, f^L)_{f \in \Sigma}$ be a Kracht style pair grammar. In (Kracht, 2007), compositionality was taken to mean that the grammar generates expressions and meanings independently of each other. But as (Kracht, 2011) observes, this can be refined: meanings can be generated independently of expressions whereas the generation of expressions might depend on both elements of the signs, and vice versa for expressions. This leads to the definitions below.

I will use the intentionally ugly terms meaning-compositional, expression-compositional, and 2-compositional for Kracht’s three main notions of compositionality, for which he uses instead autonomous, compositional, and independent. My usage is for expository purposes only and not something I recommend. But, first, it is good to bear in mind that all three are notions of (pair grammar) compositionality. Second, I don’t want to prejudge the issue of which one of them, if any, deserves to be named plain ‘compositionality’. I come back to this issue at the very end.

Let $\pi_1, \pi_2$ be projection functions for pairs:

$$\pi_1((e, m)) = e, \quad \pi_2((e, m)) = m$$

Definition 2.

$L$ is meaning-compositional if for each $f \in \Sigma$ there is an operation $r_{2f}$ such that for $\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle \in L$,

a. if $f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle)$ is defined, then

$$\pi_2(f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle)) = r_{2f}(m_1, \ldots, m_n)$$

b. if $f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle)$ is undefined then $r_{2f}(m_1, \ldots, m_n)$ is undefined.

$L$ is expression-compositional if for each $f \in \Sigma$ there is an operation $r_{1f}$ such that the conditions above hold with $\pi_1$ instead of $\pi_2$, and $r_{1f}(e_1, \ldots, e_n)$ instead of $r_{2f}(m_1, \ldots, m_n)$.

Definition 3.

$L$ is 2-compositional if for each $f \in \Sigma$ there are operations $r_{1f}$ and $r_{2f}$ such that for $\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle \in L$,

a. if $f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle)$ is defined, then

(i) $\pi_1(f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle)) = r_{1f}(e_1, \ldots, e_n)$


---

6There is no established terminology. As noted, (Kracht, 2007) uses ‘compositional’ for what I call ‘2-compositional’ and (Kracht, 2011) calls ‘independent’.
(ii) $\pi_2(f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle)) = r_2f(m_1, \ldots, m_n)$

b. if $f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle)$ is undefined then either $r_1f(e_1, \ldots, e_n)$ or $r_2f(m_1, \ldots, m_n)$ is undefined.

Note that the a and b conditions together entail that if $L$ is meaning-compositional (expression-compositional) then $f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle)$ is defined iff $r_2f(m_1, \ldots, m_n)$ is defined, and furthermore that if $L$ is 2-compositional then $f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle)$ is defined iff $r_2f(m_1, \ldots, m_n)$ and $r_1f(e_1, \ldots, e_n)$ are both defined. So it is not the case that $L$ is 2-compositional iff it is meaning-compositional and expression-compositional. That would have been the case if we had dropped the b conditions; I adopt Kracht’s use of the prefix semi for these weaker notions:

**Definition 4.**

$L$ is semi-2-compositional (semi-meaning-compositional, semi-expression-compositional) if it is as in the above definitions minus the b conditions.

A first observation is that the definability requirements which are part of expression-compositionality and meaning-compositionality are very strong, and perhaps unrealistically so. Consider, for example, cases where an ambiguous expression has meanings belonging to different categories, e.g. can as a noun and can as a verb. Suppose

$\langle \text{can}, \text{[n-meaning]} \rangle, \langle \text{can}, \text{[v-meaning]} \rangle, \langle \text{the}, \text{[det-meaning]} \rangle \in L.$

Then there should be a binary $f \in \Sigma$ such that

$f^L(\langle \text{the}, \text{[det-meaning]} \rangle, \langle \text{can}, \text{[n-meaning]} \rangle) = \langle \text{the can}, \text{[np-meaning]} \rangle$

By expression-compositionality, $f^L(\langle \text{the}, \text{[det-meaning]} \rangle, \langle \text{can}, \text{[v-meaning]} \rangle)$ is also defined (since $r_1f(\text{the}, \text{can})$ is defined), so for some $m$,

$f^L(\langle \text{the}, \text{[det-meaning]} \rangle, \langle \text{can}, \text{[v-meaning]} \rangle) = \langle \text{the can}, m \rangle$

But it is hard to see what $m$ should be, or indeed why any meaning at all should be given to the can when can is a verb. Note that if $L$ is 2-compositional, this problem disappears, for then $r_2f([\text{det-meaning}],[\text{v-meaning}])$ can safely be assumed to be undefined for those arguments, from which it follows that $f^L(\langle \text{the}, \text{[det-meaning]} \rangle, \langle \text{can}, \text{[v-meaning]} \rangle)$ would be undefined too.

If one finds situations like the one just described intolerable, then expression-compositionality and, dually, meaning-compositionality are slightly defective notions. Then we could either restrict attention to 2-compositionality, or remove the definability requirements, that is, use the semi-versions instead.
4.3 SUBSTITUTIONAL SAUSSUREAN COMPOSITIONALITY

Substitution versions of pair grammar compositionality are not discussed by Kracht, but it is natural to ask if they exist. They do. We shall need the following properties of pair grammars:

(3) a. \( L \) is **right-centered** if for all \( f \in \Sigma \), if \( f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle) \) is defined and \( \langle e'_1, m_1 \rangle, \ldots, \langle e'_n, m_n \rangle \in L \), then \( f^L(\langle e'_1, m_1 \rangle, \ldots, \langle e'_n, m_n \rangle) \) is also defined.

b. \( L \) is **left-centered** if for all \( f \in \Sigma \), if \( f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle) \) is defined and \( \langle e_1, m'_1 \rangle, \ldots, \langle e_n, m'_n \rangle \in L \), then \( f^L(\langle e_1, m'_1 \rangle, \ldots, \langle e_n, m'_n \rangle) \) is also defined.

c. \( L \) is **confluent** if for all \( f \in \Sigma \), both \( f^L(\langle e'_1, m_1 \rangle, \ldots, \langle e'_n, m_n \rangle) \) and \( f^L(\langle e_1, m'_1 \rangle, \ldots, \langle e_n, m'_n \rangle) \) are defined, \( f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle) \) is also defined.

Lemma 5.

If \( L \) is 2-compositional (**meaning-compositional**, **expression-compositional**), then it is confluent (**right-centered**, **left-centered**).

**Proof.** We check the first case; the others are similar. Let \( f, r_1f, \text{ and } r_2f \) be as in Def. 3, and suppose that \( f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle) \) and \( f^L(\langle e'_1, m_1 \rangle, \ldots, \langle e'_n, m_n \rangle) \) are defined. By Definition 3a, \( r_1f(e_1, \ldots, e_n) \) and \( r_2f(m_1, \ldots, m_n) \) are both defined. Thus, by 3b, \( f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle) \) is defined.

Now let \( GT_L \) be the set of (grammatical) terms over \( L \) as in section 2 (with \( L = (L, A_L, f^L)_{f \in \Sigma} \) in place of \( E \)), and let \( val \) the corresponding surjective homomorphism from \( GT_L \) to pairs in \( L \). We define two (total) equivalence relations on \( GT_L \). For \( p, q \in GT_L \),

(4) \( p \equiv^i_L q \) iff \( \pi_i(val(p)) = \pi_i(val(q)), i = 1, 2 \)

The following substitution conditions are then natural. For \( i = 1, 2 \),

**Definition 6.**

\( \text{Subst}(\equiv^i_L) \) If \( p[q_1, \ldots, q_n] \) and \( p[q'_1, \ldots, q'_n] \) are both in \( GT_L \), and \( q_i \equiv^i_L q'_i \) for \( 1 \leq i \leq n \), then

\( p[q_1, \ldots, q_n] \equiv^i_L p[q'_1, \ldots, q'_n] \).

**Proposition 7.**

1. \( L \) is 2-compositional iff it is confluent and \( \text{Subst}(\equiv^1_L) \) and \( \text{Subst}(\equiv^2_L) \) hold.

   Also, \( L \) is meaning-compositional (**expression-compositional**) iff it is right-centered (**left-centered**) and \( \text{Subst}(\equiv^2_L) \) \( \text{Subst}(\equiv^1_L) \) holds.

2. \( L \) is semi-2-compositional iff \( \text{Subst}(\equiv^1_L) \) and \( \text{Subst}(\equiv^2_L) \) hold. \( L \) is semi-meaning-compositional (**semi-expression-compositional**) iff \( \text{Subst}(\equiv^2_L) \) \( \text{Subst}(\equiv^1_L) \) holds.

The proof of this is not difficult and I omit it here.
5 TRANSLATING BETWEEN THE FORMATS

If we could associate with any pair grammar \( L = (L, A_L, f^L)_{f \in \Sigma} \), where \( L \subseteq E \times M \), a Hodges style grammar \( E_L = (E, A, f^E)_{f \in \Sigma} \) and a semantics \( \mu_L \) for \( E_L \) with values in \( M \) such that

a. \( L \) is compositional (in one of the Saussurean senses) iff \( \mu_L \) is compositional;

b. \( \langle e, m \rangle \in L \) iff there is \( t \in GT_{E_L} \) such that \( val(t) = e \) and \( \mu_L(t) = m \),

that would be an indication that the two formats are just notational variants of each other. But this is not so. Unless \( L \) is compositional (in some sense), there seems to be no way to even begin defining \( E_L \) or \( \mu_L \).

In the other direction, if we have a grammar \( E = (E, A, f^E)_{f \in \Sigma} \) and semantics \( \mu \) for \( E \) with values in \( M \), it may seem that we would at least have the expression composition functions of a corresponding pair grammar at our disposal. But actually, unless \( \mu \) is compositional, it is not clear that we can even define a corresponding function on pairs.

So it seems that the best we can hope for is a mapping from compositional (in some sense) pair grammars to syntactic algebras with compositional semantics, and vice versa. To substantiate this claim and make it precise, we need to look at facts about translations between the two formats.

5.1 FROM 2-COMPOSITIONAL PAIR GRAMMARS TO SYNTACTIC ALGEBRAS WITH COMPOSITIONAL SEMANTICS

Let \( L = (L, A_L, f^L)_{f \in \Sigma} \) be a 2-compositional pair grammar, so that for each \( f \in \Sigma \), the operations \( r_1f \) and \( r_2f \) are given. Let

\[
AE_L = \{ e : \exists m (e, m) \in A_L \}
\]

Then

\[
E_L = (E, AE_L, r_1f)_{f \in \Sigma}
\]

is a syntactic algebra, i.e. \( E \) is generated from \( AE_L \) via the \( r_1f \).

Proof. If \( e \in E \) there is \( m \) such that \( \langle e, m \rangle \in L \). If \( e \not\in AE_L \) then \( \langle e, m \rangle \not\in A_L \), so there are, since \( L \) is generated, \( f \in \Sigma \) and \( \langle e_i, m_i \rangle \in L \) such that \( \langle e, m \rangle = f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle) \), where \( e_1, \ldots, e_n \in E \). By 2-compositionality,

\[
e = \pi_1(f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle)) = r_1f(e_1, \ldots, e_n)
\]

However, \( L \) and \( E_L \) need not have the same signature, since the number of atoms may differ. We have (where \( |X| \) is the cardinality of \( X \)):

\[
|AE_L| \leq |A_L|
\]
To see this, choose for each $a \in AE_L$ a unique $j(a) \in M$ such that $\langle a, j(a) \rangle \in A_L$.

To obtain $|AE_L| = |A_L|$, and thus have the same signature for the two algebras, we must assume that there are no atomic ambiguities in $L$.

Next, we wish to find a semantics $\mu_L$ for $E_L$ which corresponds in a suitable way to the meaning of signs in $L$. Since $\mu_L$ should apply to terms in $GT_{E_L}$, structural ambiguities in $L$ are not a problem. And we just assumed that atomic ambiguities have been eliminated. But that doesn’t completely eliminate lexical ambiguity. There might be an atom $a \in AE_L$ which, in addition to its unique atomic meaning $j(a)$, also had a non-atomic meaning, i.e. $\langle a, m \rangle \in L - A_L$ for some $m$. So we need to strengthen the assumption just made to the requirement there are no lexical ambiguities in the following sense:7

\[ \text{(7) } L \text{ has no lexical ambiguity iff } \langle e, m \rangle \in A_L \text{ implies } \forall m' (\langle e, m' \rangle \in L \Rightarrow m' = m). \]

Thus, each $a \in AE_L$ has a unique meaning $\mu_L(a) \in M$. So when forming $GT_{E_L}$ we can identify the atomic terms with the atomic expressions, as in section 2. Then we define the complex (grammatical) terms in $GT_{E_L}$ in the usual way, simultaneously with the surjective homomorphism, say, $val^E$, from terms to $E$.

Next, we inductively extend the function $\mu_L$ to (some) terms $t$ in $GT_{E_L}$, in such a way that for each $t \in dom(\mu_L)$,

\[ \langle val^E(t), \mu_L(t) \rangle \in L \]

Suppose $t = f(t_1, \ldots, t_n)$, and that $\langle e_i, m_i \rangle \in L$, where (induction hypothesis) $val^E(t_i) = e_i$ and $\mu_L(t_i) = m_i$. Since $f(t_1, \ldots, t_n)$ is well-formed, $r_1f(e_1, \ldots, e_n)$ is defined, say, $r_1f(e_1, \ldots, e_n) = e$. If $r_2f(m_1, \ldots, m_n)$ is also defined, say $r_2f(m_1, \ldots, m_n) = m$, we let $\mu_L(t) = m$; otherwise $\mu_L(t)$ is undefined. Since $L$ is 2-compositional, $f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle)$ is defined in the former case, and $f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle) = \langle e, m \rangle \in L$.

**Proposition 8.**

If $L = (L, A_L, f^L)_{f \in \Sigma}$ is 2-compositional and has no lexical ambiguity, then $\mu_L$ is a compositional semantics for $E_L$, and the function

\[ h(t) = \langle val^E(t), \mu_L(t) \rangle \]

is a surjective homomorphism from $GT_{E_L}$ to $L$. Thus, for $\langle e, m \rangle \in E \times M$, $\langle e, m \rangle \in L$ iff for some $t \in GT_{E_L}$, $val^E(t) = e$ and $\mu_L(t) = m$.

**Proof.** The first claim is already proved: by construction, $r_{1f} = r_2f$ is the composition operation corresponding to the syntactic operation $r_{1f}$. Note that the

---

7We need not assume here that distinct atoms have distinct meanings. Also, note that we do not assume the condition for meanings corresponding to (7) since e.g. brother and male sibling should be allowed to have the same meaning. This is an asymmetry between expressions and meanings.
construction also yields that \( \text{dom}(\mu_L) \) is closed under subterms. That \( h \) is a homomorphism is seen as follows:

\[
    h(f(t_1, \ldots, t_n)) = \langle \text{val}^E(f(t_1, \ldots, t_n)), \mu_L(f(t_1, \ldots, t_n)) \rangle \\
    = \langle r_1 f(\text{val}^E(t_1), \ldots, \text{val}^E(t_n)), r_2 f(\mu_L(t_1), \ldots, \mu_L(t_n)) \rangle \\
    = f^L(\langle \text{val}^E(t_1), \mu_L(t_1) \rangle, \ldots, \langle \text{val}^E(t_n), \mu_L(t_n) \rangle) \\
    = f^L(h(t_1), \ldots, h(t_n))
\]

Surjectivity follows by an easy inductive argument.

\( h \) doesn’t have to be injective: we could have two distinct terms with the same surface form and the same meaning. For example, in some versions of Montague Grammar, there are (at least) two ways to derive

\[(8) \quad \text{John likes every student.}\]

but the meaning is the same.

The assumption that \( L \) has no lexical ambiguity is necessary in Proposition 8: such an ambiguity would require \( \mu_L \) to ‘choose’ one of the meanings, and then the other(s) would violate the last claim in the proposition. In this sense, pair grammars are better equipped than the classical syntactic algebra + semantics approach to meaning.

5.2 WHAT IF L IS ONLY EXPRESSION-COMPOSITIONAL?

The construction in the preceding subsection used full 2-compositionality of \( L \), i.e. both the syntactic and the semantic composition functions. Now it is clear that if we did not have expression-compositionality, we could not form the syntactic algebra \( E_L \), so there would be no expression terms to assign meanings to. But suppose we know that \( L \) is expression-compositional. Then we can form \( E_L \) and \( GT_{E_L} \) as before. What about \( \mu_L \)?

Assume, for simplicity, that the atoms cause no problems, i.e. that each atom (atomic term) \( a \) gets a unique meaning \( \mu_L(a) \) in \( L \). Then we can still extend \( \mu_L \) to complex terms by induction, even though the operations \( r_2 f \) are not available.

For suppose \( t = f(t_1, \ldots, t_n) \), and that \( \langle e_i, m_i \rangle \in L \), where (induction hypothesis) \( \text{val}^E(t_i) = e_i \) and \( \mu_L(t_i) = m_i \). As before, since \( f(t_1, \ldots, t_n) \) is well-formed, \( r_1 f(e_1, \ldots, e_n) \) is defined, say, \( r_1 f(e_1, \ldots, e_n) = e \). By expression-compositionality, this fact alone entails that \( f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle) \) is defined, say that \( f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle) = \langle e, m \rangle \). Let \( \mu_L(f(t_1, \ldots, t_n)) = m \).

A few remarks are in order.

1. It follows that

\[
    \exists t \in GT_{E_L}(\text{val}^E(t) = e \& \mu_L(t) = m) \Rightarrow \langle e, m \rangle \in L
\]
but we have no guarantee for the implication in the other direction, since we have insufficient 'control' over when the functions \( f^k \) are defined.

2. By construction, \( \mu_L \) is a total semantics, i.e. defined for all terms in \( GT_{E_L} \), in contrast with the case when \( L \) was assumed to be 2-compositional.

3. But \( \mu_L \) need not be compositional. If \( L \) is not semi-meaning-compositional there will be a situation where for some \( f \in \Sigma \), \( f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle) = \langle e, m \rangle \), \( f^L(\langle e'_1, m_1 \rangle, \ldots, \langle e'_n, m_n \rangle) = \langle e', m' \rangle \), and \( m \neq m' \). If there are \( t_i, t'_i \) such that \( val^E(t_i) = e_i, val^E(t'_i) = e'_i \), and \( \mu_L(t_i) = \mu_L(t'_i) = m_i \), then \( \mu_L \) is not compositional. But in general we don’t know that such \( t_i, t'_i \) exist. In section 6 we will see a ‘real’ example where \( \mu_L \) is not compositional.

4. As noted in section 4.2, expression-compositionality can seem problematic, and that one might instead retreat to semi-expression-compositionality, where the definability requirement is dropped. Then the definition of \( E_L \) and \( GT_{E_L} \) remains unchanged, and the definition of \( \mu_L \) still works, but \( \mu_L \) is now, perhaps more realistically, partial: in the inductive step above, we let \( \mu_L(t) = m \) if \( f^L(\langle e_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle) \) is defined and equal to \( \langle e, m \rangle \); undefined otherwise.

5. The situation is completely symmetric if we instead start from (semi-)meaning-compositionality (and assume that there are no atomic synonymies).

5.3 FROM SYNTACTIC ALGEBRAS AND COMPOSITIONAL SEMANTICS TO COMPOSITIONAL PAIR GRAMMARS (AND BACK)

Now let a grammar \( E = (E, A, f^E)_{f \in \Sigma} \) be given, with its corresponding set \( GT_E \) of grammatical terms and surjective homomorphism \( val : GT_E \to E \), and a compositional semantics \( \mu \) for \( E \) with values in \( M \).

We can assume that atomic terms have been added so that there aren’t any atomic ambiguities. More precisely, we assume that the set \( \overline{A} \) of atomic terms is well-behaved in the following sense:

\[
\overline{A} \text{ is well-behaved iff}
\]

\[a. \ t \in \overline{A} \text{ iff } val(t) \in A; \]
\[b. \ \text{if } t, t' \in \overline{A}, t \neq t', \text{ and } val(t) = val(t'), \text{ then } \mu(t) \neq \mu(t');^8\]
\[c. \ \overline{A} \subseteq dom(\mu)\]

Let

\[
(10) \quad \text{a. } L_{E,\mu} = \{\langle val(t), \mu(t) \rangle : t \in dom(\mu)\}
\]
\[
\text{b. } A_{t_{E,\mu}} = \{\langle e, m \rangle \in L_{E,\mu} : e \in A\}
\]

---

^8(9) a and b entail that (i) no grammar rule in \( E \) can have an atomic string as value, and (ii) although there may be ‘atomic synonymies’, they don’t hold between atomic terms with the same surface form—very natural assumptions.
For $f \in \Sigma$, we define a corresponding partial function $f^E,\mu$ from $(L_{E,\mu})^n$ to $L_{E,\mu}$. Take any $(\text{val}(t_1), \mu(t_1)), \ldots, (\text{val}(t_n), \mu(t_n))$ in $L_{E,\mu}$, and let
\[
f^E,\mu(\langle \text{val}(t_1), \mu(t_1) \rangle, \ldots, \langle \text{val}(t_n), \mu(t_n) \rangle) = \\
\langle \text{val}(f(t_1, \ldots, t_n)), \mu(f(t_1, \ldots, t_n)) \rangle
\]
if $f(t_1, \ldots, t_n) \in \text{dom}(\mu)$; undefined otherwise. To show that $f^E,\mu$ is well-defined, we need to verify:

\begin{enumerate}[label=(\arabic*)]
  \item If $\text{val}(t_i) = \text{val}(t'_i)$ and $\mu(t_i) = \mu(t'_i)$ for $i = 1, \ldots, n$, and $f(t'_1, \ldots, t'_n)$ is in the domain of $\mu$, then
    \begin{enumerate}[label=(a.), start=1]
      \item $\text{val}(f(t_1, \ldots, t_n)) = \text{val}(f(t'_1, \ldots, t'_n))$
      \item $\mu(f(t_1, \ldots, t_n)) = \mu(f(t'_1, \ldots, t'_n))$
    \end{enumerate}
  \end{enumerate}

But (11-a) follows from the definition of $GT_E$, and (11-b) follows from the compositionality of $\mu$. We have

\begin{equation}
L_{E,\mu} = (L_{E,\mu}, At_{E,\mu}, f^E,\mu)_{f \in \Sigma} \text{ is a pair grammar, i.e. it generates } L_{E,\mu} \text{ from } At_{E,\mu}.
\end{equation}

Proof. If $(e, m) \in L_{E,\mu} - At_{E,\mu}$, there is $t \in \text{dom}(\mu)$ such that $\mu(t) = m$ and $\text{val}(t) = e \not\in A$. Then $t \not\in \overline{A}$, so $t = f(t_1, \ldots, t_n)$ for some $f \in \Sigma$ and $t_1, \ldots, t_n \in GT_E$. If $\text{val}(t_i) = e_i$ and $\mu(t_i) = m_i$, this means that $(e, m) = f^E,\mu(\langle c_1, m_1 \rangle, \ldots, \langle e_n, m_n \rangle)$.

Moreover, $L_{E,\mu}$ has the same signature as $GT_E$, since

\begin{equation}
|\overline{A}| = |At_{E,\mu}|
\end{equation}

Proof. For $t \in \overline{A}$, let $k(t) = \langle \text{val}(t), \mu(t) \rangle$. Using the assumptions (9), one easily shows that $k$ is a bijection from $\overline{A}$ to $At_{E,\mu}$.

Since $\mu$ is compositional, there is for each $f \in \Sigma$ an operation $r_f$ such that if $f(t_1, \ldots, t_n) \in \text{dom}(\mu)$, then
\[
\mu(f(t_1, \ldots, t_n)) = r_f(\mu(t_1), \ldots, \mu(t_n)),
\]
and $r_f$ is undefined if $f(t_1, \ldots, t_n) \not\in \text{dom}(\mu)$. (We also assume that $\text{dom}(\mu)$ is closed under subterms.) So the definition of $f^E,\mu$ can be written

\begin{equation}
f^E,\mu(\langle \text{val}(t_1), \mu(t_1) \rangle, \ldots, \langle \text{val}(t_n), \mu(t_n) \rangle) = \\
\langle f^E(\text{val}(t_1), \ldots, \text{val}(t_n)), r_f(\mu(t_1), \ldots, \mu(t_n)) \rangle
\end{equation}

Also,

\begin{equation}
f^E,\mu(\langle \text{val}(t_1), \mu(t_1) \rangle, \ldots, \langle \text{val}(t_n), \mu(t_n) \rangle) \text{ is defined if and only if each of } f^E(\text{val}(t_1), \ldots, \text{val}(t_n)) \text{ and } r_f(\mu(t_1), \ldots, \mu(t_n)) \text{ is defined.}
\end{equation}
Proof. The ‘only if’ direction is clear. Conversely, if $f^E(val(t_1), \ldots, val(t_n))$ is defined, $f(t_1, \ldots, t_n) \in GT_E$. And if $r_f(\mu(t_1), \ldots, \mu(t_n))$ is also defined, $f(t_1, \ldots, t_n) \in \text{dom}(\mu)$, so $f^E,\mu(\langle val(t_1), \mu(t_1) \rangle, \ldots, \langle val(t_n), \mu(t_n) \rangle)$ is defined.

It follows that $L_{E,\mu}$ is 2-compositional, with $f^E$ and $r_f$ as the composition operations corresponding to $f^E,\mu$. This proves the first claim in part (a) of the next proposition. For (b), we will say that $(E, \mu)$ has no lexical ambiguity if (as in Hodges’ original formulation) each atom in $A$ has a unique meaning, i.e. each atomic expression is the value of a unique atomic term, so one can identify $A$ and $\overline{A}$. This also means that $\overline{A} \subseteq \text{dom}(\mu)$, so $\overline{A}$ is then trivially well-behaved.

Proposition 9.
(a) If $\mu$ is a compositional semantics for $E$ and $GT_E$ has well-behaved atomic terms, then $L_{E,\mu}$ is 2-compositional, and we have that $\langle e, m \rangle \in L_{E,\mu}$ iff for some $t \in GT_E$, $val(t) = e$ and $\mu(t) = m$.
(b) If $(E, \mu)$ satisfies the stronger requirement of having no lexical ambiguity, then if we apply the construction in section 5.1 to $L_{E,\mu}$, we get back what we started from, i.e. $E_{L_{E,\mu}} = E$ and $\mu_{L_{E,\mu}} = \mu$.

Proof. The 2nd claim of (a) is immediate from the construction. For (b), note:

\[(16) \quad L_{E,\mu} \text{ has no lexical ambiguity (in the sense of (7)).}\]

For suppose $\langle e, m \rangle \in At_{E,\mu}$ and $\langle e, m' \rangle \in L_{E,\mu}$. There are terms $t, t'$ such that $val(t) = val(t') = e \in A$, $\mu(t) = m$, and $\mu(t') = m'$. Since $(E, \mu)$ has no lexical ambiguity, $A = \overline{A}$, $t = t'$, and thus $m = m'$.

Since $L_{E,\mu}$ is 2-compositional, the construction in section 5.1 can be applied, and $E$, $GT_E$, $L_{E,\mu}$, and $E_{L_{E,\mu}} = (E, A^{E_{L_{E,\mu}}}, f^{L_{E,\mu}})_{f \in \Sigma}$ as defined there, all have the same signature. Now,

\[
A^{E_{L_{E,\mu}}} = \{ e : \exists m \langle e, m \rangle \in At_{E,\mu} \} \\
= \{ e : \exists m \exists t (e = val(t) \& m = \mu(t) \& e \in A) \} \\
= A
\]

The first two equalities hold by definition, and the third holds since every $e \in A$ is $val(t)$ for some $t \in \text{dom}(\mu)$. Furthermore, going from $(E, \mu)$ to $L_{E,\mu}$ we took the functions $f^E$ as the expression composition functions in $L_{E,\mu}$ (cf. (14)), and the construction in section 5.1 used those very functions for building the corresponding expression grammar. That is,

\[f^{L_{E,\mu}} = f^E, \text{ for } f \in \Sigma\]

This shows that $E_{L_{E,\mu}} = E$. Also, again by the lack of lexical ambiguity in $L_{E,\mu}$, we have

\[GT_{L_{E,\mu}} = GT_E\]
Next, by our constructions, the following holds for \( f \in \Sigma \) and \( t_1, \ldots, t_n \in \text{dom}(\mu) \), with \( \text{val}(t_i) = e_i \) and \( \mu(t_i) = m_i \), \( i = 1, \ldots, n \):

\[
f(t_1, \ldots, t_n) \in \text{dom}(\mu) \iff f^E(e_1, \ldots, e_n) \text{ and } r_f(m_1, \ldots, m_n) \text{ are defined}
\]

\[
\iff f^{E, \mu}((e_1, m_1), \ldots, (e_n, m_n)) \text{ is defined}
\]

\[
\iff f(t_1, \ldots, t_n) \in \text{dom}(\mu_{LE, \mu})
\]

Using this it is straightforward to show by induction that

\[
\mu_{LE, \mu} = \mu
\]

This completes the proof.

We remark that if we try to perform the construction of \( L_{E, \mu} \) from \( E \) and \( \mu \) without assuming that \( \mu \) is compositional, we could define \( L_{E, \mu} \) and \( Al_{E, \mu} \) as in (10), but the definition of \( f^{E, \mu} \) would not work, since (11) might fail. There seems to be no obvious way to define a suitable function on pairs in this case.

6 DISCUSSION

We saw that one can straightforwardly translate a 2-compositional pair grammar into a corresponding syntactic algebra and compositional meaning assignment, provided all lexical ambiguities have been removed. The translation works also if the pair grammar is merely expression-compositional, in which case the induced meaning assignment need not be compositional.

In the other direction, a syntactic algebra and compositional meaning assignment translates to a 2-compositional pair grammar. If there is no lexical ambiguity, this translation is an inverse of the translation from pair grammars to the standard format. But if we start with a non-compositional meaning assignment there seems to be no way to translate into the pair format.

I think we can tentatively conclude that the 2-compositional pair format is essentially equivalent to the standard compositional format, except for the (important) fact that the former can handle lexical ambiguity in a straightforward and intuitive way. Furthermore, we see that the weaker notions, here called meaning-compositionality and expression-compositionality, are new notions, without clear counterparts in the standard format. And, interestingly, that without compositionality there is no obvious way to compare the standard format and the pair format.

Are there situations where one would be content with meaning-compositionality, or expression-compositionality, but where 2-compositionality failed? This cannot really be answered until we have seen proposals for pair (or triple) grammars for serious fragments of some (natural or formal) language that satisfy such a combination of compositionality properties.\(^9\) I am aware of one relevant case

\(^9\)The examples given in (Kracht, 2011, ch. 5), are all 2-compositional.
in the literature, namely, the semantics for quotation given in (Potts, 2007). This grammar generates triples \( \langle e, X, m \rangle \); we focus on the \( e \) and \( m \) parts. Essentially, his quotation rule is the following unary function \( q^L \) on triples:

\[
q^L(\langle e, X, m \rangle) = \langle e, X, \langle e, X, m \rangle \rangle
\]

That is, the whole triple becomes the meaning of the quote.\(^{10}\) So we see immediately that this grammar is not meaning-compositional, let alone 2-compositional. But it is fairly clear that it is expression-compositional.

What should we make of this? Potts himself says that the grammar is compositional, but without specifying in which sense. It seems he was thinking of the general idea, shared by proponents of direct compositionality, see (Barker & Jacobson, 2007), that syntax and semantics are generated ‘in tandem’. As I noted in the Introduction, this could simply be construed as the basic idea of pair grammars, without any requirement of compositionality.

Intuitively, Potts’ semantics is non-compositional in the sense that substituting synonymous parts of a complex expression need not preserve meaning. From the various substitution versions (section 4.3) one sees that this holds whether we use meaning-compositionality or 2-compositionality, or translate into the standard format. Expression-compositionality just says that the expressions are generated independently of the meanings, but says nothing about the meanings themselves.

Finally, back to the issue of terminology. If you think (reasonably) that compositionality should have something to do with meaning, expression-compositionality is a separate idea. So there is a good reason to adopt Kracht’s label autonomy for it; it expresses a strong form of ‘autonomy of syntax’. What has not been made clear so far, however, is whether the distinction between meaning-compositionality (compositionality according to Kracht) and 2-compositionality (independence according to Kracht) for pair grammars has a useful role to play. This, it seems, requires further study.

ACKNOWLEDGEMENT

Thanks to Björn Haglund for spotting a number of small mistakes in an earlier version. The remaining ones, big or small, are mine.

REFERENCES


\(^{10}\)This is to take care of indirect and mixed quotation, where we need access not only to the string or phonological form quoted, but also its meaning.


