Contents ▪ Obsah

Articles ▪ State

Discussions ▪ Diskusie

Book Reviews ▪ Recenzie

Organon F 21 (3) 2014: 282-302

Representations and Relations

PETR KOŤÁTKO

Institute of Philosophy. The Academy of Sciences of the Czech Republic
Jilská 1. 110 00 Prague 1. Czech Republic
kotatko@flu.cas.cz

RECEIVED: 03-02-2014 • ACCEPTED: 08-04-2014

ABSTRACT: The paper focuses on Uriah Kriegel's non-relational account of representation, based on the rejection of the widely shared assumption that "representing something involves (constitutively) bearing a relation to it". Kriegel's approach is briefly compared with another version of non-relational theory presented by Mark Sainsbury. The author discusses several reasons why the relational aspect of representation should stand in the center of our theoretical interest, despite the arguments of non-relationists. They concern (1) the origin of the very capacity to represent in our interactions with elements of our external environment; (2) the externalist arguments attempting to show that some of our states and acts are irreducibly embedded in our relations with external environment and these relations play an ineliminable role in the constitution of their content; (3) the fact that representations typically have conditions of satisfaction which relate the representing states or acts to the external world in such a way that if the conditions are not fulfilled, this counts as a representation-failure; (4) the fact that the representation ascriptions are often based relationally and the claim that two subjects think about the same often admits only relational interpretation. The author concludes by pointing to the wide variety of phenomena called "representation" and argues that there is no *a priori* reason to presuppose that all such cases admit, or even require a unified analysis.

KEYWORDS: Representation – proposition – truth-conditions – belief-ascriptions.

1. The relational account of representation rejected

Uriah Kriegel in Kriegel (2007) and Mark Sainsbury in Sainsbury (2012) have suggested their versions of the theory of representation which they

both label as *non-relational*: according to them, it does not belong to the nature of representation that it establishes a relation between the representing person or state or act and the represented entity. In both cases, the relational account of representation is introduced as an intuitively appealing view – which should be nevertheless given up in order to do justice to the fact that we can (and quite often do) represent non-existing entities. Uriah Kriegel demonstrates this conflict on a triad of apparently uncontroversial but jointly incompatible claims (cf. Kriegel 2007, 309):

(a) One can represent non-existents.
(b) One cannot bear a relation to non-existents.
(c) Representing something involves (constitutively) bearing a relation to it.

To avoid inconsistency, we should, as Kriegel suggests, reject (c), and hence abandon the relational account of representation.

Correspondingly, Mark Sainsbury in Sainsbury (2012, 127) presents a conflict between our intuitions concerning representation (here in the specific form of "thinking about") in two series of inferences leading to a contradiction:

(1) We are thinking about unicorns (A).
(2) We are thinking about something (from 1).
(3) There is something (or there are some things) we are thinking about (from 2).

(4) There are no unicorns (A).
(5) Hence there are no unicorns we are thinking about (from 4).
(6) Hence there is nothing we are thinking about (from 5).

The challenge exemplified by these series consists in the fact that the incompatible claims (3) and (6) seem to follow from indisputable assumptions (1) and (4). The solution is to give up (3) via rejecting the inference from (2) to (3). And this means to reject a general inferential principle which, like the claim (c) in Kriegel's triad, is presented as a summary of the relational account of representation:

(R)From "*x* represents *y*" infer "there is something such that *x* represents it".[1]

[1] The rejection of (R) corresponds to Quine's rejection of relational (transparent) construal of ascriptions of attitudes, like "Ralph wants a sloop". The relational construal

Clearly, Kriegel and Sainsbury suggest two ways of summarizing the same account of representation to be rejected: since accepting the claim (c) commits us to obeying the rule (R) in our inferences and accepting the rule (R) commits us to the claim (c).

The reason which should lead us to rejecting the relational account of representation, put in meta-theoretical terms, is in both cases the same: our theory of representation should reflect the fact that we can represent non-existing entities. It seems quite natural to conclude, with Kriegel and Sainsbury, that our general theory of representation should be indifferent to the existence or non-existence of what is represented. In other words, even if we know that *X* exists, the existence of *X* should not play any role in our explanation of what representing *X* consists in. Obviously, this is not the only way of doing justice to the possibility of representing non-existents. Another way is to admit that the term "representation" is used to refer to a variety of phenomena which need not require the same analysis, even if we insist that they have something in common.[2] We can decide to focus on particular cases we find important and explain their representational character from the specific frameworks within which they fulfill their functions – instead of trying to subsume them under some general (relational or non-relational) principle.

I will argue for the plausibility of this latter approach in the end of this paper. My main concern in the following chapters will be to point to several reasons why – despite the indisputable possibility of representing non-existents – the relational aspects of representation should stand in the center of our interest. But let me start with a few words about Uriah Kriegel's non-relational account of representation.

---

has the form: ∃x(x is a sloop & Ralph wishes that Ralph has x) and Quine rejects it because of the problems he finds in quantifying into referentially opaque contexts (contexts in which the substitutivity of co-referential terms *salva veritate* is not preserved) – cf. Quine (1956).

[2] "Having something in common" is not necessary for justifying the application of the same term in all such cases: its applicability can be very well based on a "mere" family resemblance between them, or between the corresponding meanings of the term (Wittgenstein famously demonstrated this on the word "game").

2. The adverbial account of representation

The question is what – if not the representation relation to some entity (understood as a standard relation with converse[3]) makes some act or state representational. As a reply, Uriah Kriegel suggests an adverbial account of representation. Cf. for instance:

> Your thought of Bigfoot does not involve constitutively a relation to Bigfoot, on the present account, but rather the instantiation of a non-relational property of representing Bigfoot-wise. This is why your thought can represent a non-existent, even though it cannot bear a relation to a non-existent. Thus an adverbial account of intentionality solves the problem of intentional inexistence. (Kriegel 2007, 315)

Again, the question arises, what does the representation function of this "representing Bigfoot-wise" consist in – what makes it the case that thinking or imagining or fearing Bigfoot-wise can count as thinking *about* Bigfoot or imagining Bigfoot or being afraid *of* Bigfoot. Kriegel offers the following reply: the acts of thinking Bigfoot-wise include the feature of *intrinsic phenomenal directedness,*[4] i.e. the "phenomenally constituted non-relational feature of being-directed-at-something" (Kriegel 2007, 322). If we want to explicitly indicate the intrinsic phenomenal directedness of a state which we would adverbially characterize as thinking *Bigfoot-wise*, we can say, with Kriegel, that this state is "*Bigfoot-ward-esque*."[5] The relation

---

[3] That means that *A*'s being in a relation *R* to *B* implies that there is some *B* which is in a converse relation to *A*. A.N. Prior's characteristics of intentional relations as "relations without converse" can be viewed as a version of the non-relational account of intentionality – in that sense, that e.g. *A*'s thinking about *B* does not establish a standard relation between *A* and *B*. Cf., e.g., Prior (1971, 136).

[4] "The picture we get is one where many conscious states involve something like *intrinsic phenomenal directedness*: some sort of phenomenally constituted non-relational feature of being-directed-at-something" (Kriegel 2007, 322).

[5] Perhaps it is worth pointing out that thinking or imagining in a way which is Bigfoot-ward-esque does not amount to thinking or imagining a state of affairs without particulars in the sense of Strawson's idea of feature-stating statements – cf. Strawson (1959, Ch. 7). Thinking that Bigfoot is somewhere in the vicinity is not thinking that a Bigfoot-feature is present somewhere in the vicinity and, obviously, it is also not thinking that a fictitious creature called "Bigfoot" is somewhere in the vicinity, but thinking that a real, physical monster is somewhere in the vicinity. Another thing is

between the notion of representing *Bigfoot-wise* and being *Bigfoot-ward-esque* is made clear, e.g., in the following pair of sentences:

> For phenomenally conscious representations of Bigfoot have a perfect candidate for constituting the property of representing Bigfoot-wise, namely, the property of being *Bigfoot-ward-esque*. That is to say, phenomenal directedness is a perfect candidate for constituting non-relational, adverbialized intentionality. (Kriegel 2007, 323)

Kriegel points out in Kriegel (2007, 335-336, note 51) that the phenomenally conscious representation, as understood by the adverbial theory, is not the same as the mode of presentation (Fregean *Sinn*) defined in opposition to what is represented (Fregean *Bedeutung*). Or with respect to another distinction (due to Brian Loar), it is not to be understood as *how-representation* in opposition to *what-representation*. What the adverbial theory suggests is a reduction of what-representation to how-representation;[6] or assimilation of reference to sense.

Here is the corresponding Sainsbury's formulation:

> What does 'London' refer to? London. What does 'Pegasus' refer to? Pegasus. This seems an unimpugnably correct answer, even in a context in which it is well-known that there is no such thing as Pegasus. A surprising moral: even reference, philosophers' preferred tool for describing

---

that we can think this in the *as if* mode, in other words, the modus of our thought can be make-belief, rather than belief.

[6] The distinction between what-representation and how-representation (and hence the function of this reduction) is sometimes blurred due to Kriegel's terminological fluctuation. For instance, Kriegel points out that in his account the unconscious representation of non-existents is based on a relation to a conscious representation and hence is relational, though the relation in question is not a relation to what is represented (Kriegel 2007, 309). In this formulation, "what is represented" clearly means an external object of representation (in the same sense in which a Napoleon's portrait is supposed to represent the actual Napoleon). In other contexts "what is represented" is clearly supposed to mean the content of the representation itself. Cf. Kriegel (2007, note 8): "If we want to individuate representations in terms of what is represented, then given that the representations of Hesperus and Phosphorus are different representations, we must say that what is represented in those representations is different. Yet what exists is clearly one and the same. The converse case is presented by water and twin-water, where what is represented is in some good sense the same, yet what exists is certainly different."

> word–world relations, is intensional, and so a non-relational notion. (Sainsbury 2012, 109)[7]

The reduction of what-representation to how-representation or of reference to its intensional parameter provides us with a universally applicable way of speaking about representation without presupposing any relation between the representing state or act and something existing independently on it in the external world. The relationalists should be expected to object that this, if presented as a general principle, is a reduction which makes the real nature of representation inaccessible to the theory. So, let us consider several possible forms of the *relationalist challenge*: objections presenting various reasons why we should think about representation in terms of relations between the representing subject and the external world, or why this dimension of representation is at least in some cases ineliminable.[8]

3. External determinants of content

Externalists like Burge, Putnam or Kripke have made well-known radically relationist claims about the constitution of the content of what we think or communicate. They have pointed out that our thoughts and communicative acts are embedded in our external relations with elements of our environment and that at least in some cases these relations play an ineliminable role in the determination of the content of our beliefs, desires, assertions, promises etc.[9] The externalist arguments focus on particular components of our thoughts or communicative acts which are supposed to

---

[7] An objection which immediately suggests itself concerns indexicals. When I say (with a pointing gesture) "This man is coming to kill me", will you insist that I have made a claim and expressed a thought about something (about a man!), even if in fact there is nobody in the direction in question? There should be no difference here from the Pegasus' case: my thought has an aboutness structure, in which nothing is missing, even though there is no man such that my thought could be about him. The direct reference theorists will argue that in such a case I do not express a complete proposition, although I may (falsely) suppose that I have made a singular statement about particular person and that I have a singular thought about that person.

[8] Mark Sainsbury opens this topic in Sainsbury (2012) when considering contexts in which the "wordly-side" of representations is in the center of our interest in belief ascriptions.

[9] Some of the following formulations are adopted from Kot'átko (2012).

mark gaps in the internal determination of their contents, and hence function as channels through which these external intervences into the content take place. These components include indexicals, proper names, natural kind terms and all those general terms which are subject to the division of linguistic and intellectual labour. John Searle has suggested an internalist reply to some of these arguments which I find efficient and generally applicable. The idea, as I would put it, is to take seriously the externalist claim that the subject himself relies on the external factors involved in the determination of the content of his thoughts and communicative acts and to include this reliance into the construal of content. Let me introduce some examples and then generalize them in a way which I find relevant for our present discussion.

(1) Tyler Burge, already in his early writings, argued for the priority of the beliefs which are irreducibly *de re*, which means that they include ineliminable (inconceptualizable) indexical components relating the beliefs directly to some elements of the believer's environment.[10] These *relations* are supposed to make the beliefs being *about* these elements. Consider a belief based on a visual experience in which the believer is directly (*en*

[10] Cf. Burge (1977, 51): "A *de re* belief is a belief whose correct ascription places the believer in an appropriate nonconceptual, contextual relation to objects the belief is about. The term 'nonconceptual' does not imply that no concepts or other mental notions enter into a full statement of the relation. Indeed, the relation may well hold between the object and concepts, or their acquisition or use. The crucial point is that the relation not be merely that of the concepts' being concepts *of* the object – concepts that denote or apply to it. For example, although concepts may inevitably enter into the acquisition of a perceptual belief, the believer's relation to the relevant object is not merely that he conceives of it or otherwise represents it. His sense organs are affected by it. Perceptual contact is, of course, not present in every *de re* belief. But it illustrates the sort of element independent of semantical or conceptual application that is essential to the notion."

Cf. Burge (1977, 51): "A sufficient condition for a belief to be *de re* (on the vague, 'neutral' epistemic construal, as well as on our favored semantical and epistemic construals) is for it to contain an analog of an indexical expression used deictically, and pick out a *re*. The first sentences that children actually use or understand are invariably keyed to their immediate, perceptually accessible surroundings. Attitudes that accompany such assertions are clearly *de re*. These developmental matters are closely related to the question of conditions for attributing language use and understanding. I shall argue that if an entity lacks *de re* attitudes, we would not attribute to it the use or understanding of language, or indeed propositional attitudes at all."

rapport) confronted with somebody coming from a distance in a swirling fog, without being able "to describe or image him in such a way as to individuate him fully" (Burge 1977, 55). Nevertheless, the perceptual relation itself is enough for the subject to be able to have a fully determinate belief about that man, which he can express e.g. by saying "This man is wearing a red cap".

In his reply to this externalist challenge, Searle suggests an internalist construal of the content of this belief, exploiting his reflexive account of perceptual beliefs. According to this account, the believer represents the object of his belief in terms of its causal relation to his experience. In our case the result will be: "The man causing this experience is wearing a red cap." It should be clear that the demonstrative "this" included in this construction need not disturb the internalist at all, since it plainly refers to something internal.

In other cases, the components of thoughts or communicative acts supposed to provide a space for the participation of external factors in the constitution of content are natural kind terms, proper names and those general terms which are subject to the division of linguistic labour. The connected externalist theories, and arguments in favour for them, are notoriously known: hence I will confine myself to the internalist re-interpretation of these cases in John Searle's style:[11]

(2) The internalist construal of the meaning of the term "water", accommodating Putnam's essentialism and his account of intension as "extension involving":

(a) *anything that shares the essence with the stuff which causes this experience* (given that the sample of water is picked out demonstratively);
(b) *anything that shares the essence with the stuff which satisfies D* (given that the sample of water is picked out by means of the description *D*).

(3) The internalist construal of the meaning of the name "Jan Novák", as uttered by particular speaker on particular occasion, accommodating (some aspects of) the causal theory:

the man baptized by the name "Jan Novák" at the beginning of the chain to which this utterance belongs.

[11] The following suggestions are either adopted from or inspired by the analysis presented in Searle (1983, Ch. 9). Details and discussion can be found in Koťátko (2012).

(4) The internalist construal of the content of a belief, accommodating the believer's deference to an expert concept he has not mastered:

Bert: "I have arthritis in my thigh."
(a) *I have in my thigh the disease referred to by experts as 'arthritis'.*
(b) *I have in my thigh the disease which satisfies the expert notion of arthritis.*

In all these cases, the internalist construal of content is based on a principle which can be, with respect to our present discussion, put as follows: we should approach the thinking or communicating subject as not only *related* to elements of his external environment but also as *relying on* these relations and *exploiting them* in the articulation of the content of his thoughts and communicative acts. This kind of the construal of content does not *eliminate* the role these external relations play in the constitution of content, nor does it *bracket* their external character, nor does it *reduce* or *convert* them into the subject's *directedness* to them. The subject's *directedness* to these external relations and his reliance on them is presented here as a way in which he allows these external factors play a crucial role in the specification of content of his thoughts and communicative acts. What makes this account of the construal of content *internalist* is that it approaches the thinker or speaker himself as the source of articulation of his thoughts and communicative acts and hence as their real *subject*: it is the thinker or speaker himself who involves the external relations into the content specification. The role of the external relations in the determination of content is not any more interpreted as an internally unmediated *intervence* from the outside: the relations to the external environment play precisely the role assigned to them by the subject himself in the specification of the satisfaction conditions of his thoughts and communicative acts. In short, this suggestion concerning the construal of content is *internalist*, but it is *not anti-relationist*. It reinforces rather than eliminates the role of external relations in the constitution of content and of its representational functions.

4. Satisfaction conditions and their fulfillment in the external world

But even if we deny that the subject's relations with elements of his external environment participate in the constitution of content of his thoughts and communicative acts in the specific way just described, we face

a general relationist challenge. It consists in the plain pointing out that the subject himself experiences the directedness of his thoughts and communicative acts as something which relates him to the external world. For instance, he approaches the satisfaction conditions of his belief as something which is fulfilled in the external world. If it comes out that they are not fulfilled, the subject is disposed to evaluate this as a *failure*. The relationist will insist that we should respect the aspirations attached to the conditions of satisfaction of the subject's beliefs. That means to admit that they *relate* the subject to the external world (contrary to what Kriegel claims e.g. in Kriegel 2007, 322) and that this is how the subject himself approaches the content of what he believes. The relation, to repeat, is such that if the truth conditions of the belief (or satisfaction conditions of any other attitude) are not fulfilled, the subject evaluates this as a failure.

The same holds for the representation function of, let us say, the singular term "Megan" or of the corresponding element of the belief, specifiable e.g. by means of the description "the woman baptized with the name 'Megan' at the beginning of the chain which I am just joining" (cf. section 3.3). If it comes out that there is no individual meeting the satisfaction conditions of such a representation, the believer evaluates this as a failure – since the function with which he introduced that representation into the content of his belief (the role it has been designed to play) was to pick out an individual in the external world. These parameters of our beliefs are ineliminably relational since they establish relations between beliefs and the state of the world (rather than a concept of the state of the world or our thoughts about the world). At the same time it is phenomenal in the sense of being transparent to introspection (cf. Kriegel's criterion of being phenomenal – Kriegel 2007, 322). It would hardly make sense to deny that the believer takes himself as being related to the external world in such a way that if Megan does not exist, he is wrong about how things are in the world.

So, if we exclude the subject-world relations from our point of view, we are likely to miss the very nature of the phenomenal directedness of (many of) our states or acts. Fulfillment of the satisfaction conditions of representation cannot be reduced to distinctness or determinateness of the representing act or state nor to its being "filled in" with the required kind of experience.[12] Nothing can replace the facts in the role of the factors which make satisfaction conditions of representations fulfilled and hence

[12] In sense of Husserl's term "Erfüllung"; cf. e.g. Husserl (1974; 120, 151, 169, 176).

make the representation enterprise successful or accomplished.[13] And this is how the representing subject himself approaches the conditions of satisfaction of his performance – it is an ineliminable part of his aspirations. Even the most vidid experience counts *for him* as an *evidence* that the representation is successful rather than as that factor which *makes it* successful – since he is always ready to admit that the experience may deceive him.

Let me put the same point in another way. In his paper (2007, note 51) Kriegel quotes Horgan's and Tienson's characteristics of phenomenal directedness: "a conscious state's phenomenal directedness at a state of affairs is a matter of the state presenting apparent objects apparently instantiating apparent properties and apparently bearing apparent relations to each other". But this is at most one part of the matter: it belongs, on the *phenomenal* level, to the mode of presentation of objects that they are presented not *as* merely apparent but *as* real and *as* really bearing certain properties. If it comes out that the objects or their having certain properties are merely apparent, the subject evaluates this as a failure.

Kriegel (2007) draws the following lesson from what we have just quoted from Horgan and Tienson: "Talk of what is represented via phenomenal directedness in appearance terms may thus afford us a way of talking of what is represented without committing to any existents." Clearly, this does not apply to the attitude of the thinking subject himself. It is only the reporter about another subject's belief, wish, assertion etc. who can choose a kind of report which does not impose existential commitments *on him*.[14]

This point, concerning ways of belief ascriptions, deserves some attention in our present context. If we want to identify the content of the belief

---

[13] In particular, reference to the world and its facts cannot be replaced by any specification of experiences which would verify our claim that the conditions of satisfaction of some representation are fulfilled. No matter how carefully you specify the set of experiences e.g. in Bigfoot's case (experiences individuated purely phenomenally, i.e. without reference to their external causes – since our aspiration is to avoid any references to the external world), it can still be the case that you have all these experiences and Bigfoot does not exist. A complete match between a fact (e.g. existence of Bigfoot) and a set of experiences is an ideal limit we can only approximate.

[14] The availability of this kind of specification of the content of other people's beliefs is rather limited. Quite often the *de re* – i.e. relational – content specifications are the only option open to us, as I will point out later (in section 5).

which John expressed by saying "Megan is married", we have a choice between the *de re* and *de dicto* mode of belief ascription:

(a) *De re*: John believes about Megan that she is married.

Here the identification of the belief's content includes relating John to particular person as *the* person his belief is about. We commit ourselves to the existence of that person and leave open the way in which she is being represented in John's belief: even if it comes out that John does not know about Megan's being called "Megan", this will not make our ascription wrong.

(b) *De dicto*: John believes that Megan is married.

Here it is the other way round: in particular we do not commit ourselves to the existence of Megan.[15]

John Searle has argued that this kind of distinction is applicable only to belief ascriptions, i.e. to the way in which *we* identify the content of John's belief – while it would not make sense when applied to the content of John's belief itself. It would be plainly absurd to ask John whether he believes *about* Megan, an inhabitant of the external world, that she is married, or "merely" believes *that* Megan is married, without adopting any assumption concerning her existence in the external world. In his polemics with Quine, Searle has used this argument to show that the *de re* – *de dicto* opposition has only a limited distinctive function with respect to beliefs (cf. Searle 1983, 208 f.). Searle's conclusion may seem to be compatible with Kriegel's and Sainsbury's anti-relationist account of representation. Any singular belief includes certain mode of presentation of an individual: in this sense (as Searle has put it) we can say that all beliefs are *de dicto*. In some cases this mode of presentation picks out a certain individual in the world, or in other words, the conditions of satisfaction of a singular representation are met by something in the world – and in that case we can say that the belief is also *de re*, in addition to and *on the basis of* its being *de dicto*. But this is not the whole story. It is not enough to specify the con-

[15] In such a kind of ascription the Megan-representation introduced into discourse by the utterance of the name „Megan" is not used to represent a person but to identify certain component of John's belief. In Mark Sainsbury's terminology, representation is here just "put on display", rather than exercised. In Uriah Kriegel's terms we can put the same by saying that the *de dicto* ascription does not include any *Megan-ward-esque* representation: it just ascribes such a representation to John, as a component of his belief.

tent of a representation in terms of its satisfaction conditions, if we do not take into account that these conditions are approached by the believer *as fulfilled in the actual world* (i.e. if we do not properly unpack the term "satisfaction").[16]

The situation will not change if we analyze away the name "Megan" in a way adopted from Russell's theory of descriptions. For this purpose, let us (following Quine) eliminate the name by means of the predicate "is Meganic" (plus the apparatus of quantifiers, variables and logical connectives). Then we get the following Russellian-Quinean analysis of the proposition expressed by the sentence "Megan in married":

$\exists x$ (Meganic (x) & $\forall y$ (Meganic (y) $\rightarrow$ y=x) & married(x))

Here our commitment to something being the case in the external world (in particular to the existence of such and such individual) included in the application of the Megan-representation is made fully explicit in the existential quantification.[17]

Let me summarize the position I have argued for in this chapter by relating it to Uriah Kreigel's characteristics of the phenomenal directedness:

> To be sure, there is something perplexing about the notion of intrinsic phenomenal directedness. Is not saying that phenomenal experience presents us with the external world precisely saying that it is inherently relational? The short answer is *No*: to say that phenomenal experience presents us with the external world is to say that it is inherently *directed* at the external world, not that it is inherently *related* to the external world. The former would entail the latter only if directedness at the external world involved a relation to it. The claim made here is that there is a kind of phenomenal directedness that does not involve a relation to the external world. (Kriegel 2007, 322)

---

[16] Analogically, Michael Dummett has pointed out that it is not enough to specify the truth conditions of an assertion, in order to understand the assertive utterance in question. We have to add that these conditions are presented as fulfilled, in accordance with what Dummett calls "convention of assertion" – cf., e.g., Dummett (1973, 298). Similarly we do not understand the game of chess if we just specify what counts as winning: we have to add that it belongs to playing the game that the players want (or at least present themselves as wanting) to win.

[17] Since our formula represents a transcription of our original sentence in the Quinean "canonical notation", it is supposed to make explicit the ontological commitments imposed by the assertive use of that sentence on the speaker.

The relationist's reply may go as follows: The directedness of representation is (typically) directedness towards external world as that sphere in which the conditions of satisfaction of the intentional conscious state are to be fulfilled. If they are fulfilled, the directedness of the state hits its target, otherwise it misses it, is not accomplished (or consummated) and the aspiration of the intentional state fails. We can *identify* the content of the state without knowing the outcome (or score) of the kind just mentioned. But we will not understand the function of that content, its role in our mental life, if we do not take into account that it consists in specifying conditions of satisfaction and approaching them as being fulfilled in the external world. So the conditions of satisfaction of a representation relate the subject to the external world, more specifically, to certain parameter of the state of the world, identified e.g. by the question "Does Megan really exist?". Or, if you wish, to certain place in the structure of the world (certain instance of reality) identified by that question – a place which is either occupied by the fact that Megan exists or by the fact that Megan does not exist. The relation is such that something's being the case in the actual state of the world will make the subject's act or state in question satisfied.

5. The relational basis of belief ascriptions

An interpreter ascribing to an interpreted person beliefs, desires etc. (and thereby also their components like singular representations) often does so on the basis of relating that person's behaviour (linguistic as well as non-linguistic) to particular elements or aspects of his environment. He could not achieve his goal (to identify the contents of the interpreted person's attitudes in a way which will enable him to make sense of that person's overall behaviour) without respecting the subject's specific perspective. Akeel Bilgrami has attempted to incorporate this respect into the general externalist (and hence relational) principle of identifying conceptual components of contents of thoughts in the following way:

> (C) When fixing an externally determined concept of an agent, one must do so by looking to indexically formulated utterances of the agent which express indexical contents containing that concept and then *picking that external determinant for the concept which is in consonance with other contents that have been fixed for the agent.* (Bilgrami 1992, 5)

The holistic clause in the second part of this principle (in italics) is supposed to play the role of an individualist constraint imposed on the externalist determination of concepts. As Bilgrami emphasizes, this constraint is not supposed to function as a kind of an internalist filter: since the contents we have already ascribed to the subject in question are themselves composed of externally determined concepts. Bilgrami also points out that although the principle (C) applies to the concept ascriptions, the externalist position voiced in (C) is not restricted to the *epistemological* problem of detecting other subjects' concepts and propositional contents of their beliefs. The way of determination of concepts specified in (C) is supposed to reflect the *external constitution* of concepts which makes them public items.

The reference to Bilgrami has been meant as an example of a consequently relationist account of concept (and content) ascriptions which nevertheless includes a systematic respect to the subject's idiosyncratic position or point of view. No matter whether we accept Bilgrami's version of "individualist externalism", the fact that concept and content ascriptions have often relational basis is, I suppose, indisputable. And those who, like Akeel Bilgrami, believe that it is essential for contents of our thoughts that they are public items, cannot separate the question of constitution of contents from the basis on which we ascribe them to one another. Since for the contents of thoughts to be public is to be justifiably ascribable to other subjects.

Now let me proceed to a special case of ascriptions which I find particularly challenging for anti-relationists: namely claims of identity of the contents of two or more subjects' thoughts. Consider the following statement:

(1) John is thinking about Brigitte Bardot and so does Mary.

The anti-relationists are committed to reject the inference from (1) to:

(2) There is something such that John and Mary are thinking about it.

But it should be absurd for anybody to reject the inference from (1) to:

(3) John and Mary are thinking about the same.

The question (for anti-relationists) is: what precisely is supposed to be identical in John's and Mary's thought? Uriah Kriegel's reply should be that they both are representing some person BB-wise or that they both have BB-ward-esque thoughts (cf. section 2). But what does that feature shared by their thoughts consist in if we are not allowed to identify it relationally, i.e. by referring to the represented person? Nobody should deny that we

can justifiably claim (1) and (3) even if we have no reason to suppose that John and Mary share the same representation of BB (representation with the same conceptual or imaginary elements). Honestly speaking, such a sharing is in normal situations extremely unlikely – which does not prevent us from making claims like (1) and (3) quite frequently.

In general, it would be highly counter-intuitive (and it would contrast with our practice of belief ascriptions) to insist that there is a set of non-relationally specifiable conditions which have to be fulfilled in order to admit that somebody is thinking about BB. Let us imagine that John has heard about BB as the most powerful sex symbol of 60's, does not know anything about her present activities and has even forgotten her name, while Mary knows her as a fan of dogs and the most passionate admirer of Putin in France. And Jane has just heard conversation in which the name "BB" has been used and thinks that the person spoken about must have been a great film star. What justifies us in claiming that they all are thinking about the same person if the representations involved in their thoughts are so radically different? In Jane's case it will be some parasitic description like "The women referred to as 'BB' by my parents", in Mary's and John's case two totally different non-parasitic descriptions. I suspect that the only possible justification for our claim that John, Mary and Jane are thinking about the same person (or, if you wish, that their thoughts are BB-ward-esque) can be relational. It can hardly be anything else than the fact that these radically different modes of presentation are satisfied by (and hence pick out) the same person in the external world – since this is the only thing they have in common. Obviously, an anti-relationist will be right in insisting that in all these cases the thoughts can relate the thinkers to BB only because these thoughts are BB-ward-esque. But the other side of the coin is that these thoughts can be evaluated as being BB-ward-esque only on relational basis, namely because the singular representations they include pick out BB in the external world.

Now let us consider a case in which this relational principle of solving the question of identity or non-identity of objects of representations is not applicable. Both Tom and Ann believe that there is a president of Bhutan and that that person invented perpetuum mobile. Ann has a thought which she would express by saying "The president of Bhutan must be a genius" while Tom would express his thought by the sentence "The inventor of perpetuum mobile should do business rather than politics". Would we say that they are thinking about the same man? They would conclude so if

they have a conversation and both utter the sentences mentioned. If they have the beliefs we have ascribed to them, they would agree that both modes of presentation pick out the same person. But we would, I suppose, comment the way in which they are mistaken about the world by saying that the question whether they are thinking about the same man is pointless because there is no president of Bhutan and no inventor of perpetuum mobile. Here the fact that no *relation* of somebody's being represented by somebody has been established, has the consequence that the question of identity should be rejected: we are entitled to say that *things being as they are*, the question *in fact does not arise*. If you find this evaluation of the situation intuitively plausible, as I do, it should be considered as another challenge to the anti-relationist account of representation.

6. Modalities of representation

Let me summarize the reasons why I think that the relational aspects of representation should stand in the centre of interest of any theory which declares the aspiration to explain what is going on when our acts or states represent something:

(1) The very capacity to represent develops in our interactions with elements of our external environment and many representation acts either take place within these interactions or are based on them.

(2) Externalist arguments have drawn our attention to the fact that some of our states and acts are irreducibly embedded in our relations with external environment and these relations play an ineliminable role in the constitution of their content, including its representational functions. The proper internalist reaction on these arguments is, I believe, to include the subject's reliance on these relations (as involved in the constitution of content) into the internalist construal of content, without eliminating (or "bracketing") their external character.[18] Hence this defence of internalism cannot work as a defence of the non-relational account of representation against the externalist challenge.

(3) Representations typically have conditions of satisfaction which relate the representing states or acts to the external world in such a way that if the conditions are not fulfilled (in the external world), this counts as a rep-

[18] For detailed discussion see Koťátko (2012).

resentation-failure. This relation cannot be reduced to mere "intrinsic phenomenal directedness".

(4) The representation ascriptions are often based relationally and the claim that two subjects think about the same often admit only relational interpretation – since there is no reason to suppose that they exploit the same mode of presentation.

Nevertheless, we are invited by non-relationists to abstract from these things in order to get a general theory which will be applicable also to representations of non-existents. The question is whether we should aim at such a theory and whether we can hope that if we succeed in identifying a feature or a set of features present in all the cases of (what we are used to call) representation, this will help us to explain how the representation function works in these cases.

Here are some of the cases I have in mind:

(a) thinking about an object based on (and in reaction to) a direct perceptual contact with it;
(b) thinking about Cicero;
(c) thinking about Homer as a person whose existence is uncertain;
(d) thinking about Pegasus as a mythological creature;
(e) thinking about Emma Bovary while (and as part of) reading Flaubert's novel.

Let me, to get a maximal contrast, confront the first and the last case. We have already had opportunity to compare an externalist and internalist approach to an example of a perception based belief – one that the believer would express e.g. by saying "That man is wearing a red cup" (cf. section 3). In both cases the representation of the object of the belief has been construed as based on a direct perceptual relation to it. The internalist version is even more radically relational than the externalist one: in the former, the relation to the object is not only involved in the content determination. The internalist construal of the content includes *the believer's reliance* on his being in certain (namely causal) external relation with the object as on that factor which will make his belief being *about* that object. As you may remember, the belief in question has, according to Searle, the following structure: "The man causing this experience is wearing a red cap."[19]

[19] Or, if we unpack the demonstrative in Russellian way: "There is precisely one man causing this experience and whoever causes this experience is wearing a red cap."

Of course you may object that the believer can be hallucinating and then there is no relation between his experience and the external world such that it picks out the object of his belief. But this is a typical example of a radical failure, in which the believer is wrong both about the state of the world and, correlatively, also about the state of his own mind. The mechanism of representation on which the believer relies does not work: this is how he himself would evaluate the situation and this should be part of the description of the situation also from the point of view of the theory of representation. In other words, the theory should not be indifferent to the contrast between the hallucinatory and non-hallucinatory case, even if it is true that there is no difference between them detectable by introspection. If perception based representations rely on certain external circumstances and play the role they are designed to play only under those circumstances, we should try to explain their nature from the way in which they function *under these circumstances*.

Let me now apply the same principle to Emma Bovary case. I will attempt to explain what thinking about Emma as part of reading Flaubert's novel consists in from the function which thinking about Emma has in our getting access to the literary functions of Flaubert's text. In other words, the question concerns the requirements imposed by the literary functions of Flaubert's sentences (those including the name "Emma Bovary") on the reader: what kind of interpretation of the occurrences of such sentences within the literary text will allow the text to fulfill its literary functions for the reader? Here is the reply I am suggesting: the reader is supposed to interpret the occurrences of these sentences in the text of the novel as records of utterances of a real person, the narrator, in which the narrator speaks about that person, who has been in the actual world assigned the name "Emma Bovary" at the beginning of the chain (in Kripke's sense) to which these narrator's utterances belong.[20] It is important to add that the reader is supposed to approach Flaubert's sentences in this way in the *as if* (or: *make-belief*) mode. This includes that the reader supposes, in the *as if* mode, that "Emma Bovary" is a standard proper name used by the narrator with standard referential function based on the previous act of baptism and the chain anchored in it in sense of Kripke's causal theory. In this account, Emma Bovary is the person the reader has to assume (in the *as if* mode) as

[20] Defence of this approach can be found in Koťátko (2013).

a real correlate of the narrator's utterances – a correlate understood in sense of the referential relation established by the Kripkean chain.

The fact that the literary function of sentences including the name "Emma Bovary" requires such a reading, can be commented so that the occurrences of this name in these sentences indicate an aspiration at the referential function in a relational (more specifically: Kripkean) sense and that the reader is supposed to accept this aspiration (in the *as if* mode) as fulfilled, as part of his accepting (in the *as if* mode) the truth aspiration of the sentence as a whole as fulfilled. If this is right, then this specific case cannot be taken as a counter-example to the relational account of representation. Rather, it should be approached as parasitic (in the way just described) upon reference in relational sense. The act of representation which takes place here does not relate the reader to any entity, and hence is clearly non-relational: but representation in relational sense is *supposed* (in the *as if* mode) by the reader to take place here and this assumption is required by the literary functions of the text.[21]

The crucial point is that all this (including the reference to the narrator) belongs to the way in which the reader's thoughts about Emma represent their object: the mechanism of representation has the complex structure just described, in other words, the function of representation includes (and depends on) the moves just mentioned. The reader, in his Emma-thoughts involved in his reading Flaubert's text, represents Emma *via* approaching (in the *as if* mode) occurrences of the term "Emma" as records of the narrator's utterances and ascribing them (in the *as if* mode) the function described above.

Let us compare this with the preceding example: the mechanism of representation which is at work in case of a belief based on a direct perceptual contact with its object and includes the subject's reliance on the causal relation between that object and his current experience. According to my opinion, the moral to be drawn from such confrontations is that if we want to explain how representations work in particular cases, we should resist

[21] A corresponding point from the author's (rather than interpreter's) perspective has been made by Saul Kripke: "...when one writes a work of fiction, it is part of the pretense of that fiction that the criteria for naming, whatever they are, are satisfied. I use the name 'Harry' in a work of fiction; I generally presuppose as part of that work of fiction, just as I am pretending various other things, that the criteria of naming, whatever they are, Millian or Russellian or what have you, are satisfied. That is part of the pretense of this work of fiction" (Kripke 2013, 17).

the temptation to subsume them under some general (relational or non-relational) principle. This would be just a way of avoiding the real work to be done: to analyze the cognitive or communicative contexts in which they are embedded, the functions they are designed to fulfill in these contexts, the mechanisms involved and the conditions of their doing properly their work.[22]

References

BILGRAMI, A. (1992): *Belief and Meaning*. Oxford: Blackwell.
BURGE, T. (1977): Belief De Re. *Journal of Philosophy* 74, 338-362.
DUMMETT, M. (1973): *Frege: The Philosophy of Language*. London: Duckworth.
HORGAN, T. – TIENSON, J. (2002): The Intentionality of Phenomenology and the Phenomenology of Intentionality. In: Chalmers, D. J. (ed.): *Philosophy of Mind: Classical and Contemporary Readings*. Oxford – New York: Oxford University Press.
HUSSERL, E. (1974): *Formale und transzendentale Logik*. Haag: M. Nijhoff.
KOŤÁTKO, P. (2012): Searle's Defence of Internalism. *Organon F* 19, Suppl. Issue 2, 93-105.
KOŤÁTKO, P. (2013): Fikce, skutečnost a radikální vyprávění. *Organon F* 20, č. 1, 72-96.
KRIEGEL, U. (2007): Intentional Inexistence and Phenomenal Intentionality. *Philosophical Perspectives* 21, 307-340.
KRIPKE, S. (2013): *Reference and Existence*. Oxford: Oxford University Press.
PRIOR, A. (1971): *The Objects of Thought*. Oxford: Clarendon Press.
QUINE, W.V.O. (1956): Quantifiers and Propositional Attitudes. *The Journal of Philosophy* 53, No. 5, 177-187.
SAINSBURY, M. (2012): Representing Unicorns: How to Think about Intensionality. In: Currie, G. – Koťátko, P. – Pokorný, M. (eds.): *Mimesis: Metaphysics, Cognition, Pragmatics*. London: College Publications.
SEARLE, J. (1983): *Intentionality. An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.
STRAWSON, P.F. (1959): *Individuals*. London: Methuen.

[22] An earlier version of this paper has been presented at the workshop on *Phenomenality and Internalism* (June 2013, Prague). I am grateful to Uriah Kriegel, Eros Corazza, Alberto Voltolini and other workshop participants for inspiring criticism.

Organon F 21 (3) 2014: 303-326

Quantificational Accounts of Logical Consequence II: In the Footsteps of Bolzano[1]

LADISLAV KOREŇ

Department of Philosophy and Social Sciences. Faculty of Arts. University of Hradec Králové
Náměstí Svobody 331. 500 03 Hradec Králové. Czech Republic
ladislav.koren@uhk.cz

RECEIVED: 06-09-2013 • ACCEPTED: 07-05-2014

ABSTRACT: Quantificational accounts of logical consequence account for it in terms of truth-preservation in all cases – be it admissible substitutional variants or interpretations with respect to non-logical terms. In this second of my three connected studies devoted to the quantificational tradition I set out to reconstruct the seminal contributions of Russell, Carnap, Tarski and Quine and evaluate them vis-à-vis some of the most pressing objections. This study also prepares the ground for my discussion of the standard model-theoretic account of consequence to be found in the concluding study.

KEYWORDS: Logical consequence – quantificational accounts – substitutions – interpretations – persistence.

1. Introduction

The idea that we could fruitfully explicate the notion of *C logically following from P* as truth-preservation under all admissible variations with respect to all non-logical elements in $P \cup C$ has proved extremely influential in the western logico-semantic tradition.[2] Attempts to explain consequence following this recipe are sometimes called *quantificational accounts*. My first

[1] My work on this study was supported by the grant GAČR, n. P401/12/P599.

[2] *P* and *C* represent the premise-set and the conclusion-set respectively.

study (see Koreň 2014) devoted to the quantificational tradition revolved around three contentions. First, quantificational accounts give pride of place to the formal dimension of consequence. Second, what marks them out from other approaches that likewise emphasise the formal aspect of consequence is the fact that, one way or another, they propose to explain away, or reduce, the modal element that resurfaces in informal glosses on consequence and related notions. Third, full-blooded accounts fitting this description seem to have emerged only with Bolzano. Bolzano, unlike his distinguished predecessors, was prepared to take the bold step of accounting for consequence and related logical ideas solely by generalizing over appropriate cases (admissible substitutional variants) without having to appeal to irreducibly modal notions.

In this follow-up study I am about to explore the quantificational tradition following in the footsteps of Bolzano.[3] I start with Russell's account of modal and logical notions as representing specific properties of propositional functions, not least because it provoked a principal objection due to Wittgenstein, the gist of which seems to pose a *prima facie* challenge also to modern substitutional and interpretational accounts devised for languages of mathematical logic. Thus I shall reconstruct two important *objections from overgeneration* in connection with the quantificational accounts of Carnap (cf. Carnap 1937), Tarski (cf. Tarski 1936) and Quine (cf. Quine 1970/1986) and their ramifications. I then go on to spell out what I consider the main residual worries, suggesting that they gesture towards the standard model-theoretic approach as a superior quantificational account that promises to assuage them. Whether this conjecture can be vindicated vis-à-vis a battery of heavyweight objections levelled by the modern critics of the model-theoretic account is a delicate issue,[4] whose treatment is left for the concluding part of my explorations.

2. In the footsteps of Bolzano

2.1. Russell on modal and logical notions

Bolzano's substitutional account reduces logical truth of a sentence *A* to the universal truth of a sort: truth under all admissible variations with re-

[3] I summarized some of them in the concluding section of Koreň (2014).

[4] Etchemendy (1990) has been the most influential critical voice.

spect to its non-logical elements. Bolzano thought that the same holds for the relation of *C logically following from P*.[5] He claimed that Aristotle's turn of phrase "results of necessity" occurring in his classic account of *deduction* can only be understood in terms of the *whenever*-connection between *P* and *C*.[6] And this connection he explained as truth-preservation under all admissible variations w.r.t. the set *V* containing all the non-logical elements occurring in $P \cup C$.[7] Whether or not he was right about Aristotle's intentions is debatable, to say the least. What is not debatable is the fact that Bolzano made a very intriguing proposal that has proved attractive to many philosophers since then.[8]

Interesting affinities can be found in Russell's explications of the modal notions of *possibility*, *impossibility* and *necessity*. In his widely read lectures on the philosophy of logical atomism Russell (1918/1919) argues that such notions do not apply to propositions, but to "propositional functions". The reason is that, once we read "*A* is possible" as "*A* is sometimes true" or "*A* holds in some cases", this indicates that we can make sense of *A*'s having cases or instances. Yet only something with undetermined elements can have cases or instances. Such things are propositional functions with variable elements whose values for various definite arguments replacing the variables are various determinate propositions. Or so Russell argued.

Russell then goes on to say that a propositional function $\Psi(x_1,...,x_n)$ is *possible* if it holds in *at least one* propositional instance of it, that is, if it yields a true proposition for at least one admissible substitution for all its free variables. Now this amounts to reducing possibility to the truth of an

---

[5] Bolzano (1837/1972) employed the term *deducibility* for the generic relative consequence-relation, that is, *C* following from *P* with respect to a set *V* of variable elements (not necessarily all and only the non-logical elements) occurring in $P \cup C$. Logical consequence requires the set *V* to contain all and only the non-logical elements occurring in $P \cup C$. See Bolzano (1837/1972, §29). The same applies, *mutatis mutandis*, to his notion of *logical analyticity* – cf. Bolzano (1837/1972, §148).

[6] Recall the *locus classicus*: "[...] deduction is a speech in which, some things having been supposed, something other than what has been supposed results of necessity from their being so" (Aristotle 1964, 24b18-22). Bolzano's gloss is as follows: "[...] the 'follows of necessity' can hardly be interpreted in any other way than this: that the conclusion becomes true whenever the premises are true" (Bolzano 1837/1972, §155, §§219-220).

[7] Henceforth, I use "w.r.t." to abbreviate "with respect to".

[8] Łukasiewicz (1957) is one influential commentator who agreed with Bolzano that Aristotle implicitly subscribed to this approach.

existential proposition $\exists x_1,...,x_n\Psi(x_1,...,x_n)$. In a similar spirit, $\Psi(x_1,...,x_n)$ is said to be *necessary* if it holds in *every* propositional instance, that is, if a true proposition results for every admissible substitution for all its variables. This, again, comes to reducing necessity to the truth of a universally quantified proposition of the type $\forall x_1,...,x_n\Psi(x_1,...,x_n)$.[9]

This is not yet Russell's account of specifically *logical necessity*. Russell thought that genuine logical truths are law-like propositions concerned with the real world, with which the process of abstraction and generalization reached as it were its "utmost limit". Thus logic, he famously said, "is concerned with the real world just as truly as zoology, though with its more abstract and general features" (Russell 1919, 169). It follows, according to Russell, that genuine logical truths are fully generalized propositions composed solely of logical elements together with variables of appropriate logical types, none of which refers to any specific contents of the world. And this complete abstraction from the specific contents of the world is what renders logical truths *formal*, hence topic-neutral (but we shall see shortly that there is another sense that Russell attaches to the notion of formality – i.e. the Wittgensteinian idea of truth by virtue of a logico-syntactic make-up alone, hence irrespectively of possible ways the world could be – that does not coincide with the former sense).

To clarify what this amounts to we should note that for Russell neither the proposition

(1) Oscar is a philosopher or Oscar is not a philosopher

nor its first-order universal closure

(2) $\forall$x(x is a philosopher or x is not a philosopher)

is strictly speaking logically true (necessary), since neither is purely formal in that both involve reference to a specific subject-matter. What, according to Russell, would qualify as a genuine logical truth (law) is the second-order universal closure with respect to all the topic-sensitive elements:

(3) $\forall X\forall$x(x is X or x is not X).

In a sense, however, we can say that (1) or (2) are logically true (necessary), though in a derived way, being specific instances of (hence deducible

[9] In an analogous manner, impossibility is defined as non-existence of a verifying propositional instance of a propositional function. See Russell (1919, 162).

from) the completely general logical law (3). In this way, then, logical truth can be reduced to the truth of a completely general proposition.

This, I take it, should remind us of Bolzano's substitutional account in that the universal closure of all (including, possibly, higher-order) variables has a similar effect as the talk about the truth of *all* admissible instances of a propositional form (e.g. *x is X or x is not X*) irrespectively of what non-logical elements of fitting types uniformly replace variable elements (cf. Corcoran 1973 and Sagüillo 2002).

Incidentally, Bolzano's and Russell's accounts of logical consequence are close too. Using the familiar conditional-manoeuvre, Russell reduces the relation of logical consequence between the (finite) premise-set $\{A_1, ..., A_n\}$ and the conclusion B to the truth of a universal closure of the conditional

If A_1^* and ... and A_n^*, then B^*,

where the starred letters stand for the corresponding propositional functions that do not contain any non-logical elements but only logical constants together with variables of appropriate types. For instance, given that we treat '=' as a fixed logical constant, to assert

"¬(17 = 6)" follows logically from "17 is prime and 6 is not prime"[10]

is a way of asserting something general about the logical propositional function

If X(x) and ¬X(y), then ¬(x = y).

In fact, it is something that we could express by its second-order universal closure

∀X∀x∀y(if X(x) and ¬X(y), then ¬(x = y)).

With this higher-order truth the process of logical generalization has finally reached its utmost limit.

2.2. Wittgenstein's principal challenge

According to Russell, then, logical consequence reduces to logical truth. And the latter reduces itself to a fully general truth of a sort – truth irrespectively of the specific referents of non-logical terminology.

---

[10] Russell's preferred idiom was: *A formally implies B*.

Wittgenstein famously complained that this Russellian conception of logical truth – as a fully general truth – insufficiently distinguishes logical truths from mere generalities that could be only accidental:

> The mark of a logical proposition is *not* general validity... An ungeneralized proposition can be tautological just as well as a generalized one. (Wittgenstein 1921, § 6.1231)
> Logical general validity, we could call essential as opposed to accidental general validity, e.g. of the proposition "all men are mortal". (Wittgenstein 1921, § 6.1232)

What Wittgenstein claims here is that generality is neither necessary nor sufficient for logical truth. As regards the first point, he would contend that the claim made by (1) is tautological if anything is. He thus attacks Russell's view, according to which the status of (1) as a logical necessity is at best derivative: it can be called logically true, being an instance of the completely general law expressed by the claim (3). But, by Wittgenstein's lights, for a proposition to qualify as a logical truth it must be a vacuous tautology holding independently of factual matters, accordingly enjoying *a priori* status, as any recourse to empirical evidence is out of question (cf. Wittgenstein 1921, §6.1, §6.11). Now, (1) is such a tautology, as its elementary truth-functional character testifies.

As for Wittgenstein's second point, what he had in mind is that completely generalized truths may well express only something very general about reality. Precisely because of that, however, the possibility is always open that they hold only accidentally in that the reality may just happen to possess this general structure (or feature) rather than a different one:

> Our fundamental principle is that whenever a question can be decided by logic at all it must be possible to decide it without further ado. (And if we get into a position where we have to look at the world for an answer to such a problem that shows that we are on a completely wrong track.) (Wittgenstein 1921, §5.551)

Thus, once we assign logical propositions a subject-matter – be it completely general or, perhaps, about peculiar logical objects of a sort – we have failed to separate them principally from empirical propositions, and, in particular, from *a posteriori* generalizations:

> All theories that make a proposition of logic appear to have content are false. [...] On this theory it seems to be anything but obvious, just as,

> for instance, the proposition, 'All roses are either yellow or red', would not sound obvious even if it were true. Indeed, the logical proposition acquires all the characteristics of a proposition of natural science and this is the sure sign that it has been construed wrongly. (Wittgenstein 1921, §6.111)

Their truth would thus depend on *how things are* so that to recognize them as true we would presumably have to check the facts to confirm whether it is this (general) way rather than any other (general) way. Yet this is completely misguided if, as Wittgenstein has it, logical propositions are distinguished from factual-empirical propositions precisely in that

> [...] one can recognize that they are true from the symbol alone, and this fact contains in itself the whole philosophy of logic. (Wittgenstein 1921, §6.113)
> That is to say, if we know the logical syntax of any sign language, then all the propositions of logic are already given. (Wittgenstein 1921, §6.124)

This formal dimension has a semantic counterpart. Holding (or not) irrespectively of how things are, logical propositions do not describe reality but determine the very structure of the whole logical space of combinatorial possibilities. There is therefore no genuine reference to the factual-empirical, hence no genuine subject-matter – not even a completely general one.

2.3. Russell's way of addressing Wittgenstein's challenge

Returning now to Russell, he was well aware of Wittgenstein's challenge. He tried to fix the problem – apparently influenced by Wittgenstein – by contending that logical truths are to be not just fully general but also *tautological* in the specific sense of being true in virtue of their logico-syntactic make-up, hence irrespectively of the possible ways the world could be. As he also put it, they are to be true *in virtue of form* (cf. Russell 1919, 197). Of course, this manoeuvre ignores Wittgenstein's first point that complete generality is not necessary for logical truth, given that propositions such as (1) are logically true. Yet it appears to make at least some progress with regard to the second objection that complete generality does not guarantee logical truth. Russell agrees it does not. He denies, for instance, that sentences like "There is at least one thing" are truths of pure logic, though they may be expressed in purely logical words.[11] What he

[11] E.g. formalized as $\exists x(x = x)$, where identity is treated as a logical symbol.

tentatively suggests, I submit, is that *generality* + *formality* could provide such a guarantee. His idea seems to be, first, that truly logical propositions are truly general (abstract, topic-neutral) laws. Second, since such logical laws are the (unempirical) source of validity of their (less general) instances, the later inherit from the former specifically logical necessity.

Unfortunately, Russell's account is rather obscure in this crucial respect. Thus, having said that the form cannot be one of the constituents of the proposition whose form it is – otherwise, what would hold the form and the other constituents together? – Russell tentatively suggests that it might be a subject matter of another logical proposition so that

> [...] it is possible that logical propositions might be interpreted as being about forms. (Russell 1918/1919, 75)

Fully general logical propositions, recall, are not about specific things, properties or relations, but, presumably, they are not completely devoid of subject matter either. Russell sometimes talks as if the formality of a logical proposition consisted precisely in the fact that its subject matter is a logical form:

> [...] another way of stating the same thing is to say that logic (or mathematics) is concerned only with forms, and is concerned with them only in the way of stating that they are always or sometimes true – with all the permutations of "always" and "sometimes" that may occur. (Russell 1919, 199-200)

Yet he felt rather insecure about this – and not without reason. On the one hand, the passage confirms the analogy with Bolzano: a logical truth such as (3) says, in effect, that the form (F) *x is X or x is not X* holds in all instances, for all admissible values of the variables "X" and "x". On the other hand, how does the fact that (3) is about (F) show that it itself is true in virtue of form? Indeed, in virtue of what form? In virtue of (F), which is supposed to feature as a constituent in its subject matter? That seems confused, as, intuitively, (F) is not the form of (3), but of its instances such as (1). And Russell cannot say that (3) also displays (F), because he has maintained that no form of a proposition can be a constituent of its subject matter. So, particular instances of (3) could perhaps be said to be true in virtue of the form (F), but they are not fully general propositions; and while (3) is fully general, it is not true in virtue of (F).

Keeping the spirit of Russell's approach, we could tentatively suggest the following. The fully general proposition (3) is *sui generis* in that it dis-

plays the form such that the class of propositions with the same form is the class containing nothing but (3). In this way, we could maintain that (3) is true in virtue of its form, since the only proposition of this form is true. Moreover, (3) is fully general. So (3) can be deemed logically true, as it is both completely general and true in virtue of form.

Unfortunately, even this amended proposal fails to address Wittgenstein's challenge, because it does no separate purely logical laws from contingently true generalities expressed in a purely logical idiom. As Russell pointed out himself, the proposition such as "There is at least one thing identical with itself" – and, in general, cardinality statements – can be translated into the purely logical idiom à la *Principia Mathematica* (cf. Russell 1919, 203; see also footnote on the same page). Yet it seems to state something substantive and contingent about the way the world is, which need not hold in different ways the world could be. Accordingly, its truth cannot be recognized on *a priori* grounds, which epistemic quality Russell clearly expects genuine logical truths to possess.

I noted that Russell was aware of the difficulty and proposed to attack it with glosses inspired by Wittgenstein's conception of tautology. Without any real progress, as far as I can judge. Indeed, he confessed to be unable to explain in a satisfactory way his Wittgenstein-inspired notion of tautology as capturing one essential mark of (logical) *analyticity*, whose other side is *a priority*.[12]

In what follows I am about to show that modern quantificational accounts propose a somewhat different approach: instead of seeking a unified account of genuinely logical laws ultimately grounding all logical necessities, they aim to formulate a meta-theoretic account of logical consequence and truth in terms of truth-preservation under all admissible variations of a sort. But we will also see that they inherit some of the central difficulties that confronted Russell's approach.

---

[12] See Russell (1919, 205). Another serious trouble is that if the idea of reduction of all logical truths (necessities) to general logical laws is cashed out as their deductive encapsulation in the latter, it founders on Gödel's first incompleteness theorem (Gödel 1931), provided that deduction from logical laws is understood in the standard proof-theoretic sense.

3. From substitutions to interpretations

3.1. Carnap: substitutional account for formalized languages

Another important contributor to the quantificational tradition was Rudolf Carnap. In his monumental *Logical Syntax* he provided substitutional accounts of logical notions, though he proceeded in a reversed order. He proposes first to define logical consequence this way:[13]

> *B* is a logical consequence of $A_1,...,A_n$ iff *B* is a consequence of $A_1,...,A_n$ and either (1) *B* and $A_1,...,A_n$ are all logical sentences containing only logical expressions, or (2) B^* is a consequence of $A_1^*,...,A_n^*$, where B^* and $A_1^*,...,A_n^*$ are any admissible substitution-variants of *B* and $A_1,...,A_n$ respectively, obtainable by uniformly replacing all descriptive expressions occurring in the latter sentences with other descriptive signs of the same logical type.[14]

This definition itself presupposes the definition of the generic consequence-relation in terms of what can be obtained from a premise-set by means of repeated applications of certain transformation rules.[15] With this in hand, Carnap accounts for analyticity (logico-analytical truth) as follows:

> *A* is analytic iff *A* is a consequence of the empty premise-set and either (a) *A* contains only logical vocabulary, or (b) *A* is such that every sentence obtainable from it by uniformly replacing all its descriptive signs with other descriptive signs of the same logical type is true (cf. Carnap 1937, 181; see also Coffa 1991).

---

[13] For details see Coffa (1991), Creath (1998), Procházka (2006) and several essays in Wagner (2009), especially de Rouilhan (2009).

[14] I have simplified the definition, leaving out of account the case when the descriptive signs in premises or conclusion are defined. In that case, the conclusion logically follows from the premise-set iff the logical consequence relation (as defined above) holds between their variants in which all non-primitive terms are everywhere replaced by their *definiens*.

[15] Note, though, that the consequence-relation is infinitary in character: basically, it is explained in terms of what can be obtained from a premise-set by means of repeated applications of possibly infinitary transformation rules such as the so-called omega-rule allowing us to infer the universal closure $\forall xPx$ from the infinite premise-set that includes $P(n)$, for each given natural number *n*. Introduction of such infinitary ("indefinite") deductive means was an important element in Carnap's original way of accommodating Gödel's incompleteness results within his generously syntactic project, fully acknowledging their limitative force with regard to finitary ("definite") means of proof.

This is one of three definitions of logical properties given in Carnap's *opus magnum*.[16] It can be found in the important part on the general syntax, in which Carnap attempts to reconstruct all the key notions of logical consequence, analyticity, contradictoriness and determinacy – including, importantly, the partition of all terms into logical and descriptive – starting with the definition of the consequence-relation. The other two definitions were given for the so-called Language I and Language II respectively. For Language I, Carnap's recipe was much like the general one provided above: viz. analyticity was defined as a limiting case of logical consequence from the empty premise-set. However, for the much stronger LII (of the type-theoretic sort), he defined logical consequence as a limit-case of L-contradictoriness: B follows logically from $A_1, ..., A_n$ iff $\{A_1, ..., A_n\} \cup \{\neg B\}$ is contradictory. Importantly, Carnap's account of analyticity and contradictoriness made use of the method of valuation and was remarkably close to Tarski's celebrated procedure of defining truth via a recursive definition of satisfaction of open sentences by sequences of objects (of fitting logical types).[17] Basically, the idea was that analyticity of complex (including quantified) sentences can be systematically reduced, in a step-by-step manner, to analyticity of atomic sentences.

In Carnap's view, this procedure amounted to the definition of *logical truth* for sentences of Language II articulated in purely logical (logico-mathematical) expressions. A problem with this way of defining logical truth is that it does not neatly extend to sentences containing descriptive constants. Yet it was arguably Carnap's ambition to provide a general method of defining analyticity (and related notions of contradictoriness and consequence) also for descriptive languages of science (e.g. for physicalistic extensions of Language II). Carnap's proposal here was that in the specific case of a descriptive sentence A we san say that:

> A is analytical (logically true) iff A^* is analytical, where A^* is a sentence obtained (a) by replacing all descriptive constants of A uniformly by

[16] For further details cf. Coffa (1991) and Creath (1998). The definitions were meant to be given for an object-language in a more powerful (syntactic) meta-language.

[17] Tarski (1935). It can be said that Carnap (1937) defined truth for formalized logico-mathematical languages, though, unlike Tarski, he did not realize that much the same procedure can be used to define truth also for languages containing descriptive terms – cf. Coffa (1991) and Creath (1998).

> variables of appropriate types and (b) universally closing the matrix so obtained with respect to each free variable introduced.

The definitions of contradictoriness and logical consequence for such sentences would have to be modified accordingly. Here, again, Carnap's procedure is recognizably quantificational in its spirit. In the next section, however, we shall have an occasion to see that this definition might eventually require a non-substitutional rendering of universal quantifiers supposed to close the purely logical matrix.

3.2. The objection from persistence violation

To sum up what has been said so far: substitutional accounts hold that consequence obtains between a (possibly empty) premise-set and the conclusion if there is no counter-example having the same logical form, where logical form is determined by the fixed logical terms and the pattern of remaining non-logical elements that are treated schematically. There is, mind you, no trace of modality in the explanations. Indeed, the appeal of quantificational approaches lies in the fact that we seem to need only appropriate quantifiers plus the notion of plain truth in order to account for logical properties in terms that are not philosophically contentious. In addition, quantificational approaches appear to do justice to the powerful intuition that logical properties are formal in nature: *if they apply to something, they apply to anything of the same form*. Logical status is, in this specific sense, *exceptionless*.

That being said, substitutional strategy of the sort I have reviewed may intuitively overgenerate, as it hinges on the expressive capacity of the underlying language. As Etchemendy put it, it is a plausible condition on an adequate account of logical consequence (truth) that:

> The property of being logically true with respect to a given F [class of fixed logical terms] should persist through simple expansions of the language ... [and] the property of not being logically true should persist through contractions of the language. (Etchemendy 1990, 30)

The same holds, *mutatis mutandis,* for consequence. Reductive substitutional accounts of logical consequence appear to violate this adequacy condition, at least when they are framed in a linguistic framework.[18] Viewed

[18] Unlike Bolzano's (1837/1972) original account, which was designed for non-linguistic propositions and their component ideas, whose number as well as identity is independent of common languages, hence not constrained by their expressive limits.

from one direction, some arguments (sentences) in some language L can turn out valid (logically true) on the linguistic substitutional account, only because L does not have enough expressions to express a genuine counter-example – available in simple expressive expansions of L. From the opposite direction: some arguments (sentences) in L that are invalid (logically false) on the substitutional account, given that L has enough expressions to express a genuine counter-example, can become valid in its contractions (or sub-languages) lacking expressive resources to frame a counter-example.

Essentially this kind of objection was put forward by Tarski (1936) in his classic article on logical consequence. He says there that an adequate account of this notion should capture the following property called "the condition F":

> X follows logically from K only if X' is true whenever every member of K' is true, where X', K' differ from X, K only by replacement of all constants except logical constants (cf. Tarski 1936, 415).

Tarski then rushes to point out that this is a sufficient condition of X following logically from K, only if the underlying language has sufficient resources to designate any (type-theoretic or set-theoretic) object its quantifiers range over. But he deems this idea patently absurd, if only because there are many more set-theoretic objects than there are linguistic expressions.

3.3. Tarski's interpretational account circa 1936

Tarski's remedy was not to reject the basic idea behind the substitutional account but to improve on it so as to overcome the problem of persistence-violation. Like Bolzano, Russell or Carnap, he was arguably a supporter of a demodalized quantificational account of logical properties in terms of form, truth and generality. The challenge he faced was to find its proper formulation. Having argued that the F-condition is necessary but not yet sufficient for logical consequence, Tarski proposed what is nominally a model-theoretic account of logical consequence:

> The sentence X follows logically from the sentences of the class K if and only if every model of the class K is also a model of the sentence X. (Tarski 1936, 417)

Familiar as this account sounds Tarskian models are non-standard by contemporary standards. First, the definition was designed for type-theoretic logical systems common in the 1930s. Second, the consequence rela-

tion is defined for fully interpreted sentences, but their models are defined by detour through sentential functions obtained by uniformly replacing all their non-logical constants (in all their occurrences) by variables of fitting logical types. Thus, according to Tarski, starting with the sentence X and the class of sentences K, what we obtain via such uniform replacements will be the sentential function X* and the class of sentential functions K*. Whatever sequence of set-theoretic objects (of types appropriate to free variables of X*) satisfies X* is then called a model (or realization) of the sentence X.[19] Much the same can be said, *mutatis mutandis*, of the relation between K and K*: an arbitrary sequence of objects of fitting logical types that satisfies each sentential function in the class K* is called a model of that class of sentences.

Tarski thereby arrived at a definition of consequence that is remarkably close to Carnap's account introduced in Section 2.4. And his procedure is clearly a variation on the simple quantificational theme, as it reduces logical consequence (and logical truth) to plain truth plus generality. The chief difference is the consistent switch from substitutions to semantic valuations: what we uniformly vary are not extra-logical expressions but their set-theoretic values.[20]

It is worthy of comment that Carnap's account of logical notions for Language II could address the objection from persistence violation, owing to the fact that he was prepared to switch from substitutions to semantic values if need arises. This applies to his generalized definition of logical truth for Language II covering sentences with descriptive terms, in case one cannot reduce their analyticity just to analyticity of their substitutional variants but needs to quantify over semantic values proper. The need for this move was clearly pointed out to Carnap by Gödel in their correspondence on Carnap's early attempt to define analyticity (circa 1932), which was based on the substitutional reading of quantifiers. Gödel showed Carnap that his attempt was marred by circularity and proposed to fix the problem by treating second-order variables as raging over any property whatever defined over the individual domain (whether or not it can be

[19] Tarski's account of models of sentential functions builds on his celebrated recursive definition of satisfaction of sentential functions by (infinite) sequences of appropriate objects, given in Tarski (1935).

[20] Also, unlike Carnap, Tarski does not presuppose the generic consequence-relation defined as a closure upon inference rules (including rules, such as the omega-rule, that are infinitary in character).

named). Carnap incorporated this into his amended account in the *Logical Syntax*, where he stipulated the range of second-order variables to be the power set of the countable set of numerical expressions of Language II.[21] Much like Tarski, Carnap could in this way transcend the problem of persistence violation.[22]

4. One world is not enough: the objection from overgeneration due the fixed domain

Russell, Carnap and arguably also Tarski in the 1930s all had the idea that quantifiers are logical terms that cannot be varied via substitutions or interpretations. This is built into their logic in that individual variables are determined to range over one fixed domain of all individuals.[23] In fact, they all used to work with type-theoretic deductive frameworks devised to keep the spirit of the logicist reconstruction of the classical mathematics while avoiding Russell-type paradoxes.

---

[21] Thus note that Language II, compared to Language I, is a very powerful formalism capable to embody a substantive amount of classical mathematical reasoning. With it Carnap embraced the idea of higher-order logic as a framework for mathematics. Now, while in the first-order fragment of Language II there are enough names (numerals) for every object in the domain of first-order quantifiers, the substitutional strategy is out of place with regard to higher-order fragments of Language II, as there are not enough names for every object in the intended domain of higher-order quantifiers (say for all properties/sets defined over the set of all individuals). See the discussion in Awodey – Carus (2007).

[22] Viz. the richness of objects to be found in the transfinite type-theoretic hierarchy – of which Language II is a part – that could be used in possible semantic valuations.

[23] Over this domain, then, domains for higher-order variables are defined, if the language is, as was then usual, type-theoretic. Whether Tarski (1936) propounded a fixed-domain or variable-domain conception of models is a vexed question. The fact is Tarski did not say there that different interpretations of sentential functions can be based on different domains. The discussion of this historical issue would take me too far afield. Let me just say that I think that Etchemendy (1990) is right that Tarski's account does not involve variable domains of interpretations. Consult Mancosu (2006; 2010), Bays (2002) or Corcoran – Sagüillo (2011) for more detailed arguments in favour of the view that Tarski held a sort of fixed-domain conception of models throughout 1930s, which are considerably more sensitive to the subtleties of Tarski's position in the historical context. Bays (2002), Mancosu (2006; 2010) and Schiemer – Reck (2013) also show that Tarski had resources that allowed him to simulate effects of domain-variation so that he could frame close enough versions of important metatheorems (such as completeness, etc.).

The trouble with this view is that once the domain of quantification is fixed, logical properties would seem to depend on its size in a way that is intuitively problematic. A notorious example is the sentence $\exists x \exists y \neg(x = y)$, which is logically true, if the domain contains at least two things, but logically false, if it contains only one thing, because it does not contain any non-logical element to be varied. Accordingly, if the fixed domain contains two or more things, any inference that has this sentence as a conclusion is logically valid. Generally, an inference of the following type that contains only cardinality-sentences about the least number of existing things

$$\frac{\exists x_1 \ldots \exists x_n [\neg(x_1 = x_2) \wedge \ldots \wedge \neg(x_1 = x_n) \wedge \neg(x_2 = x_3) \wedge \ldots \wedge \neg(x_2 = x_n) \wedge \ldots \wedge \neg(x_{n-1} = x_n)]}{\exists x_1 \ldots \exists x_{n+1} [\neg(x_1 = x_2) \wedge \ldots \wedge \neg(x_1 = x_{n+1}) \wedge \neg(x_2 = x_3) \wedge \ldots \wedge \neg(x_2 = x_{n+1}) \wedge \ldots \wedge \neg(x_n = x_{n+1})]}$$

is logically valid or invalid depending on the size of the fixed domain. Thus, for $n = 2$ and the fixed domain containing just two objects, the inference is invalid, since, in that case, the premise is true and the conclusion false. But if $n = 2$ and the domain contains at least three things, the inference is valid, the premise and conclusion being both true.

What this consideration seems to show is that, on the assumption that the quantifier is a fixed term picking out the universe of all existing individuals (or a fixed portion of it), quantificational accounts make logical properties dependent on the extra-logical fact of how many things there are in the fixed domain.[24] That is to say, any true cardinality sentence is logically true, and any false cardinality sentence is logically false, since there is nothing to be varied. Yet cardinality sentences, even if non-contingently true or false, do not seem to be true or false on purely logical grounds.[25]

---

[24] With the possible exception that the domain is to be non-empty, though even this is over the heads of proponents of free logics. Already Russell (1919, 203, n. 1) was uneasy about this specific assumption, which was derivable from the axioms of *Principia Mathematica*.

[25] Such systems were based on an infinite individual domain, but this could hardly be otherwise, if the aim was to reconstruct the body of classical mathematics in them. This old-fashioned project is deemed *passé* today. Indeed, one tends to think that the axiom of infinity compromised it from the start. That said, it is to be noted that this was not the received view in the late 1920s and early 1930s. Thus Carnap and Tarski in the 1930s – both with sympathies to the general logicist idea – were prepared to treat (at

This raises the following problem for all quantificational accounts discussed so far: it seems that logical properties should persist not just under contractions and expansions of the non-logical vocabulary, but also under possible contractions and expansions of the domain of quantification.

5. Quine's parsimonious approach: substitutional account vindicated?

Quine (1970/1986) is the last proponent of the quantificational approach whose views merit our attention. Interestingly enough, he had a sort of response to both aforementioned overgeneration objections, based on his version of the substitutional approach.[26] He restricted his substitutional account to languages regimented in the austere first-order idiom (lacking primitive identity-symbol, individual terms and function symbols), for which he distinguished five accounts of logical truth and consequence: in terms of *structure*, *substitutions*, *models*, *proof* and *grammar*. The first account states that *A* is a logically true sentence if it is true by virtue of its logical structure alone. Since *A*'s structure is revealed via replacing its predicates by schematic letters, Quine says that we can just as well define logical truth and consequence in the second way:[27]

---

least sometimes) issues of cardinality as logical issues. So the discussion in this section is to be read in this light. Note, however, that Russell himself (1919, 202–203) viewed the axiom of infinity as problematic (along with the infamous axiom of reducibility) on the ground that it seems to be an extra-logical postulate. In his classic review of logicism Carnap (1931) had voiced a similar worry with respect to the traditional logicist program; yet, somewhat later, in Carnap (1937) he did not seem to be worried about assuming the axiom of infinity (pursuing already his quasi-syntactic and pluralist approach to logic without any clear-cut boundary between logic and mathematics).

[26] See Quine (1970/1986, 53-56). Compare also Quine (1950/1982, Ch. XIII). Quine mentions that the completeness theorem shows that, in case of first-order languages, logical truth and consequence can be adequately (extensionally correctly) approached in terms of proof and provability. But he does not prefer the proof-theoretic approach, on the ground that, unlike the general substitutional approach, it is arbitrary to the extent it depends on the choice of this or that proof-system.

[27] See Quine (1970/1986, Ch. 4). Simple open sentences play in Quine's regimented language the role of terms in Bolzano's unregimented language. Further measures are to be taken to avoid possible collision of variables. For a discussion see McKeon (2004).

(i) A^* is a logically valid schema iff all its instances obtained via admissible substitutions – of (open) sentences for its simple sentential schemata – turn out to be true sentences.
(ii) A is a logically true sentence iff it is such an instance of some logically valid schema A^*.
(iii) B follows logically from A_1, ..., A_n iff "If A_1 and ... and A_n, then B" is a logically true sentence.

The third account, which is Quine's version of the model-theoretic account in terms of set-theoretic interpretations of non-logical elements over varying domains, goes roughly like this (cf. Quine 1970/1986, 51-52):

(i) Open sentence $S(A^*)$ is a set-theoretic analogue of the schema A^* iff $S(A^*)$ is obtained by uniformly replacing every simple sentential schema of the type $P(x_1,...,x_n)$ that occurs in A^* by a corresponding set-theoretic construction of the type $(x_1,...,x_n) \in \gamma$, γ being a variable ranging over sets.
(ii) Model-sequence $M = (D, \alpha, \beta,...)$ of (n-dimensional) sets defined over the domain-set D satisfies the set-theoretic analogue $S(A^*)$ iff $S(A^*)$ comes out true when D is assigned as the range of its individual variables and the sets α, β,... are assigned (in the requisite order) to the set-variables occurring in $S(A^*)$.
(iii) Model-sequence $M = (D, \alpha, \beta,...)$ satisfies the schema A^* iff M satisfies its set-theoretic analogue $S(A^*)$.
(iv) A^* is a logically valid schema iff every model-sequence M satisfies A^*.

The clauses for logical truth and consequence may then remain the same as in the previous account.

Once we have the two accounts in place, the objection from persistence-violation can be reformulated: due to a lack of expressive power on the part of the object-language it might happen that the first account would count some sentences as logically true (having no substitutional counter-examples to them) that have set-theoretic counter-models. This, Quine admits, holds for expressively impoverished regimented languages. However, he has an argument that in the case of a canonical first-order language expressively adequate to elementary arithmetic there is no shortage of substitutional counter-examples *vis-à-vis* set-theoretic counter-models. So his substitutional account of logical truth does not overgenerate with respect to the account of logical truth spelled out in terms of set-theoretic interpretations.

What he set out to show is that the following equivalence holds for an arbitrary first-order schema S:

> S is true under all admissible substitutions iff no admissible set-theoretic interpretation is a counter-model of S.

He notes that the right-to-left direction of the equivalence

> If no admissible set-theoretic interpretation is a counter-model of S, then S is true under all admissible substitutions

holds due to a version of Gödel's completeness theorem for first-order logic (see Gödel 1930). The theorem states that every S that is true under all admissible set-theoretic interpretations (so has no counter-model) can be derived in a visibly sound first-order proof-system such that every provable schema is assured to have only true substitution-instances. Quine's argument for the left-to-right direction

> If S is true under all admissible substitutions, then no admissible set-theoretic interpretation is a counter-model of S

draws on two crucial meta-theorems. Let us first reformulate this conditional equivalently as follows:

> If some admissible set-theoretic interpretation is a counter-model of S, then S is false under some admissible substitution.

According to the Löwenheim-Skolem fundamental theorem, if a first-order schema has a model (counter-model) at all, it has a countable model (counter-model) in the domain of positive integers. So what we have to substantiate is this:

> If S has a countable counter-model in the domain of positive integers, then S is false under some admissible substitution.

Quine points out that Hilbert and Bernays showed that countable models or counter-models can be expressed by elementary number-theoretic open-sentences (available, recall, in Quine's preferred language – see Hilbert – Bernays 1934). Such open-sentences are exactly what one would need to frame falsifying substitution-instances of S if S had a countable counter-model. So we have:

> If S has a countable counter-model in the domain of positive integers, then S is false under some admissible substitution of open-sentences of elementary number theory.

Put all this together by chaining of implications and you have proved the left-to-right direction of Quine's equivalence.[28]

Incidentally, Quine also had a response to the second overgeneration objection, though he did not formulate it explicitly himself.[29] The clue lies in the fact that he did not treat the identity sign "=" as a primitive logical symbol of his canonical language (as is usual in first-order languages with identity) but as a defined symbol that expresses *indiscernibility with respect to all n-adic predicates* of the object-language. Consequently, sentences containing only quantified individual variables, sentential connectives and '=' are purely logical only by appearance, that is, until we replace the defined identity sign '=' with its definiens involving descriptive *n*-adic predicates of the object-language.[30] Once so expanded, it can be shown that cardinality sentences like $\exists x \exists y \neg (x = y)$ admit of substitutional counter-examples.[31] In this way, Quine could sustain his contention that, for a sufficiently expressive first-order language at least, his parsimonious substitutional account is all one needs.

Quine's vindication of his substitutional account raises a couple of questions. First, one could worry that his proposal makes logical properties dependent on substantive matters alien to logic, for an elementary number theory is needed to provide an adequate theory of sentences (finite strings) and substitutions, which is in turn equivalent to the theory of finite sets. I doubt, though, that this worry would have bothered Quine. Granted, his explication of logical properties brings in its own ontological commitments. But this is as it should be, by Quine's lights, because any theoretical ac-

---

[28] However, Boolos (1975, 52-53) argues that Quine's argument is correct for logical truth but, for logical consequence, his proof of extensional adequacy goes through only if the-premise set of an argument is finite (or at least arithmetically definable).

[29] Shapiro (2000, 339) suggests this Quinean way out and McKeon (2004) details it.

[30] Quine (1970/1986, 63) calls this procedure "exhaustion of combinations".

[31] That said, Quine also argues that the basic laws of reflexivity, transitivity and symmetry of identity are logical truths in his substitutional sense, even though "=" is a defined binary predicate. Cf. Quine (1970/1986, 63-64). McKeon (2004) discusses some technical problems with Quine's idiosyncratic treatment of identity. Perhaps the most important philosophical point that many commentators have mentioned is that Quine makes identity of individuals relative on a richness of languages, since what is indiscernible with respect to all non-logical *n*-place predicates of one language, may well be distinguishable with respect to all non-logical predicates of a richer language. And that sounds counter-intuitive: Quine's '=' does not express identity after all.

count (the model-theoretic included) is bound to make some ontological commitments. And Quine is quick to remind us that the ontological costs of his substitutional approach are modest compared to the rival model-theoretic account formulated in terms of varying set-theoretical valuations (including varying domains of such valuations):

> [...] it renders the notions of validity and logical truth independent of all but a modest bit of set theory; independent of higher flights. (Quine 1970/1986, 56)[32]

Second, and more importantly, Quine's vindication of the substitutional account relies on fundamental meta-results for first-order predicate logic and it does not therefore carry over to languages regimented in higher-order predicate calculi. This would not have bothered Quine, who contended – partly on independent grounds – that second-order predicate calculus is a set theory in sheep's clothing, and hence no pure logic (see Quine 1970/1986, 66). Still, if one does not feel comfortable with Quine's fairly restrictive view of the realm of pure logic, one would have to seek another idea.

6. Conclusion

Where does this leave us? In light of the discussion so far, two reasonable desiderata on plausible accounts of logical properties have emerged that can be spelled out as follows:

(1) logical properties of (sets of) sentences had better persist under subtractions and expansions of the non-logical vocabulary;
(2) they should also persist no matter what sequence of values of appropriate types we assign to their non-logical elements – whatever possible domain of application those values may come from.

---

[32] Note that Quine's strategy of assuring extensional adequacy of his substitutional account already assumes that the first-order object-language is rich enough to embed elementary number theory. A related complaint could then be that, unlike the model-theoretic account assuming domain-variation, Quine's account makes logical properties dependent on the immanent ontological assumptions of the object-language itself. Quine, I guess, could retort that even the model-theoretic rival account makes some (indeed, much heavier) assumptions, albeit at the meta-level.

Only in this way, one might argue, could we hope to do justice to the intuition that logical validities as well as truths are topic-neutral and hence should not depend on substantial assumptions, be they empirical or mathematical. The problem is that the quantificational strategies fail to meet one or both of those desiderata.

When we set aside the objection from overgeneration due to the persistence violation as something that interpretational accounts eventually allow us to overcome, the real source of trouble seems to lie in the fact that all quantificational accounts so far reviewed treat quantifiers in a misguided way: viz. as rigidly ranging over one fixed domain, irrespectively, as it were, of the context of application. This, then, prevents them from doing justice to the intuition that specifically logical consequences and truths – conceived of as formal and topic-neutral in nature – are sensitive to the logico-semantic profile of sentences but insensitive to the specific contents associated with their descriptive vocabulary (if any) or specific domains they may be applied to.

That we can do better than this is the idea driving the standard model-theoretic account of logical consequence and related notions, which explicitly allows domains to vary across admissible semantic interpretations of language. I would eventually argue that this is the most promising quantificational approach on the market, not least because it provides formally rigorous explications of logical properties – relative to a principled account of the semantic behaviour of certain traditionally distinguished logical constants – that make room for fruitful metatheoretical comparisons between the semantic and the deductive side of logic.

This is a delicate issue – and one that has provoked much controversy recently – that deserves a separate discussion, because a mutated version of Wittgenstein`s principal challenge to Russell`s account is in a way still with us. Thus, if logical relations and properties are to be construed as formal and topic-neutral, it would seem that they should not be contingent on substantial truths. Yet even the model-theoretic approach appears to make them contingent on substantive matters, this time in the form of specific background set-theoretic assumptions.

This and closely related issues – such as what, in general, we can reasonably expect from fruitful meta-theoretical explications of central logical notions such as consequence – will be addressed in the concluding part of my explorations of the quantificational tradition.

Acknowledgments

I am grateful to Jaroslav Peregrin, Karel Procházka and James Edwards for corrections and instructive comments on earlier drafts of the article.

References

AWODEY, S. – CARUS, A.W. (2007): Carnap's Dream: Gödel, Wittgenstein, and *Logical Syntax*. *Synthese* 159, No. 1, 23-45.

BAYS, T. (2002): Tarski on Models. *Journal of Symbolic Logic* 66, No. 4, 1701-1726.

BOLZANO, B. (1837/1972): *Theory of Science*. Translated and edited by R. George. Oxford: Basil Blackwell. Translation of selected parts of *Wissenschaftslehre. Versuch einer ausführlichen und grösstentheils neuen Darstellung der Logik mit steter Rücksicht auf deren bisherige Bearbeiter*. 4 Vols. Sulzbach: J. E. v. Seidel.

BOOLOS, G. (1975): On Second-Order Logic. *Journal of Philosophy* 72, 509-527. Reprinted in: Boolos, G. (1998): *Logic, Logic, and Logic*. Harvard: Harvard University Press, 37-54.

CARNAP, R. (1931): Die logizistische Grundlegung der Mathematik. *Erkenntnis* 2, 91-105.

CARNAP, R. (1937): *Logical Syntax of Language*. London: Kegan Paul.

COFFA, J.A. (1991): *The Semantic Tradition from Kant to Carnap*. Cambridge: Cambridge University Press.

CORCORAN, J. (1973): Meanings of Implication. *Dialogos* 9, 59-76.

CORCORAN, J. – SAGÜILLO, J.M. (2011): The Absence of Multiple Universes of Discourse in the 1936 Tarski Consequence-Definition Paper. *History and Philosophy of Logic* 32, No. 4, 359-374.

CREATH, R. (1998): Carnap's Move to Semantics: Gains and Losses. *Vienna Circle Institute Yearbook* 6, 65-76.

DE ROUILHAN, P. (2009): Carnap on Logical Consequence for Languages I and II. In: Wagner, P. (ed.): *Carnap's Logical Syntax*. Palgrave MacMillan.

ETCHEMENDY, J. (1990): *The Concept of Logical Consequence*. Cambridge (Mass.): Harvard University Press.

GÖDEL, K. (1930): Die Vollständigkeit der Axiome des logischen Funktionenkalküls. *Monatshefte für Mathematik und Physik*. English translation in: Gödel, K. (1986): *Collected Works. Vol. I. Publications 1929 – 1936*. Feferman et al. (eds.), Oxford: Oxford University Press, Oxford, 102-123.

GÖDEL, K. (1931): Über formal unentscheidbare Sätze der Principia Mathematica und-verwandter Systeme I. *Monatshefte für Mathematik und Physik* 38, 173-198. English translation in: Gödel, K. (1986): *Collected Works. Vol. I. Publications 1929 – 1936*. Feferman, S. et al. (eds.), Oxford: Oxford University Press, Oxford, 144-195.

HILBERT, D. – BERNAYS, P. (1934): *Grundlagen der Mathematik*. Vol II. Berlin: Springer.

KOREŇ, L. (2014): Quantificational Accounts of Logical Consequence I: From Aristotle to Bolzano. *Organon F* 21, No. 2, 22-44.

ŁUKASIEWICZ, J. (1957): *Aristotle's Syllogistic from the Standpoint of Modern Formal Logic*. Oxford: Clarendon Press.

MANCOSU, P. (2006): Tarski on Models and Logical Consequence. In: Ferreiros, J. – Grayand, J. (eds.): *The Architecture of Modern Mathematics*. Oxford: Oxford University Press, 209-237.

MANCOSU, P. (2010): Fixed- versus Variable-Domain Interpretations of Tarski's Account of Logical Consequence. *Philosophy Compass* 5, 745–759.

MCKEON, M. (2004): On the Substitutional Characterization of First-Order Logical Truth. *History and Philosophy of Logic* 25, 195-214.

MORSCHER, E. (2012): Bernard Bolzano. In: Zalta, E. (ed.): *Stanford Encyclopedia of Philosophy*. Available at http://plato.stanford.edu/entries/bolzano.

PROCHÁZKA, K. (2006): Consequence and Semantics in Carnap's Syntax. In: Kolman, V. (ed.): *From Truth to Proof*. Miscellanea Logica (vi), 77-113.

QUINE, W.V.O. (1950/1982): *Methods of Logic*. 4th ed. Cambridge (Mass.): Harvard University Press.

QUINE, W.V.O. (1970/1986): *Philosophy of Logic*. 2nd ed. Cambridge (Mass.): Harvard University Press.

RUSSELL, B. (1918/1919): The Philosophy of Logical Atomism. *Monist* 28, 495-527; *Monist* 29, 32-63, 190-222, 345-380. Reprinted in Russell, B. (1956): *Logic and Knowledge*. London: Allen and Unwin, 177-281.

RUSSELL, B. (1919): *Introduction to Mathematical Philosophy*. New York: Macmillan.

SAGÜILLO, J.M. (2002): Conceptions of Implication. *Logica Trianguli*, 41-67.

SCHIEMER, G. – RECK, E. (2013): Logic in the 1930s: Type Theory and Model Theory. *The Bulletin of Symbolic Logic* 19, No. 4, 433-472.

SHAPIRO, I. (2000): The Status of Logic. In: Boghossian, P. – C. Peacocke (eds.): *New Essays on the A Priori*. Oxford: Clarendon Press, 333-366.

TARSKI, A. (1935): Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica* 1, 261-405. English translation in: Tarski, A. (1983): *Logic, Semantics, Metamathematics*. 2nd ed. Indianapolis: Hackett, 152-267.

TARSKI, A. (1936): Über den Begriff der logischen Folgerung. *Actes du Congrés International de Philosophie Scientifique* 7, 1-11. English translation in: Tarski, A. (1983): *Logic, Semantics, Metamathematics*. 2nd ed. Indianapolis: Hackett, 409-420.

WAGNER, P. (2009): *Carnap's Logical Syntax*. Palgrave Macmillan.

WITTGENSTEIN, L. (1921): *Tractatus Logico-Philosophicus*. Transl. by C. K. Ogden. London: Routledge.

Organon F 21 (3) 2014: 327-337

Expressive Completeness in Brandom's *Making It Explicit*[1]

RADEK OCELÁK

Institute of Philosophy and Religious Studies. Faculty of Arts. Charles University in Prague
Náměstí Jana Palacha 2. 116 38 Prague. Czech Republic
radioc@seznam.cz

RECEIVED: 06-01-2014 • ACCEPTED: 05-05-2014

ABSTRACT: In this paper, I focus on the notion of expressive completeness in Robert Brandom's *Making It Explicit*. For Brandom as a normative pragmatist, a theory of meaning is expressively complete if it specifies a human practice that is sufficient to confer on expressions conceptual contents so rich that the very conferring practice can be described by means of these expressions. I put the notion of expressive completeness in contrast with the related, but non-identical notion of self-referentiality of a semantic theory. Further, I examine the position of the concept in Brandom's philosophical project: I assess the justification Brandom provides for his claim of expressive completeness of the presented theory, and I outline the consequences he can draw for his overall project provided that expressive completeness is achieved. Whether it is actually achieved, remains however an open question.

KEYWORDS: Expressive completeness – normative pragmatism – Robert Brandom – self-referentiality – theory of meaning.

[1] I am grateful to Martin Stokhof and Jaroslav Peregrin for their valuable comments. All remaining flaws are my own responsibility. This study was written within the Programme for the Development of Fields of Study at Charles University, No. P13 Rationality in human sciences, sub-programme Knowledge and Normativity. Also, I owe thanks to the Dutch Nuffic for their support of my past studies in the Netherlands.

1. Introduction

In philosophizing about language meaning, one condition on a successful theory of meaning suggests itself quite naturally: the condition of self-referentiality, or the theory's meaningfulness according to its own standards. More specifically, the theory should correctly account (whatever that might involve) for meaning of all those expressions by means of which it is formulated. Probably the most famous failure in this respect is Wittgenstein's: the penultimate remark of *Tractatus* openly reflects the paradoxical character of the presented theory, namely that it appears senseless when measured by the very standards it puts forth.[2] Self-referentiality is certainly not the only criterion of assessing theories of meaning. For pragmatic reasons we might prefer, e.g., one that lacks self-referentiality but assigns meanings to a broad range of expressions, to a self-referential theory with poor coverage of other expressions than those employed in its own formulation. Still, there is a clear theoretical appeal to self-referentiality: unless a theory of meaning satisfies this condition, it is in principle always in need of *another* theory of meaning (cf. Scharp 2010, 265).

Robert Brandom in his *Making It Explicit* (henceforth *MIE*; see Brandom 1998) sets himself the ambitious goal of developing his theory of meaning up to the point at which something called *expressive completeness* is reached. Although this notion, I think, despite some misleading formulations does not coincide with that of self-referentiality, it is closely related. In comparison with the seemingly more straightforward ideal of self-referentiality, I here intend to clarify the notion of expressive completeness (section 2) and its position in Brandom's philosophical project (sections 3 and 4).[3] In the latter, my focus will be on how Brandom hopes to justify the claim of expressive completeness of his theory, what he uses it as a premise for, and whether expressive completeness, as opposed to genuine self-referentiality, suffices for his aims. Concerned primarily with the significance of expressive completeness for Brandom, I will not provide a verdict on whether expressive completeness is actually achieved in his book. I will

---

[2] Garver (1996) argues that the later Wittgenstein's philosophy, unlike *Tractatus*, is a success in this respect. I object to this claim elsewhere – see Ocelák (forthcoming).

[3] That is, the project of *Making It Explicit*. In the present paper, Brandom's more recent work (see Brandom 2008, in the first place) and its possible relevance to the issues at hand are left aside.

indicate, though, how much still needs to be shown in order to provide the answer, at any rate an affirmative one.

2. Expressive completeness

Brandom is a normative pragmatist. His theory of meaning thus comes in the form of a (normative) specification of human practices that are, purportedly, sufficient to confer appropriate conceptual content on various expressions engaged in these practices. (Brandom is also a rationalist who endorses a most intimate relation between thought and speech. Where many others would talk about the *meaning* of an expression, Brandom prefers to characterize it semantically through determining its *conceptual content*, something which is at the same time attributable to certain intentional states and performances.) In Brandom's view, the most explanatory account of an expression's meaning consists in a refined specification of how a practice is normatively structured so that it employs that expression in the particular role it does.

The official ambition, then, is to have the theory elaborated up to the point where it becomes expressively complete (see Brandom 1998, xx, xxii, 641). This is achieved when the practice the theory specifies is sufficient to confer on expressions conceptual contents so rich that the very conferring practice can be described by their means.

Some of Brandom's formulations suggest that this aim is (at least in his pragmatic setting) identical to what I called a theory's self-referentiality above. (As a matter of fact he even talks about "self-referential expressive completeness"; see Brandom 1998, xxii.) For instance: "[...] they [i.e., the participants of the specified practice] will be able to express the theory offered here. [...] the project eats its own tail [...] presenting an explanation of what it is to say something that is powerful enough to explain what it itself is saying" (Brandom 1998, xx).[4] But these two ideals need not involve the same, and I want to argue that they do not in Brandom's case. What we primarily ascribe expressive completeness to seems to be a vocabulary rela-

[4] Cf. Laurier (2005, 142), too: "[...] it should be expressively complete, in the sense of including an account of the conceptual resources that are needed in order to formulate this very account of conceptual content." Also, Scharp (2010, 265): "Hence, the entire theory of meaning [...] can be formulated by the members of the extended practice."

tive to its instituting practice.[5] But it is of course not necessary that this practice will be specified in terms of the vocabulary it institutes. We can specify a practice (calling this specification our theory of meaning) that confers conceptual content such that the resulting idiom is powerful enough to describe the practice, while being quite distinct from our own expressive resources (which we have employed in our theory). Such a theory will have achieved the goal of expressive completeness. (That is to say that the *vocabulary* instituted by the described practice will be expressively complete. In a derived sense though, we may ascribe expressive completeness also to the theory.) However, the theory is not truly self-referential, for it does not directly account for the meaning of all of its own expressions. Let me call this feature *weak* self-referentiality.

I believe that this, and not genuine self-referentiality, should be seen as the ambition of *MIE*, or in any case is a more realistic one. The focus should be on the content, rather than on the formulation, of our theory of meaning: the instituted vocabulary should suffice to specify the same practice our theory specifies, be it in different terms. Even the *MIE*'s expression I quoted above (cf. Brandom 1998, xx) can be perhaps read in this content-focused manner, since *to say something* is ambiguous between emphasis on the expression and on the content. Also other formulations allow the interpretation in favor of stating the theory's content rather than the theory itself:[6] "[...] the theory should specify practices sufficient to confer on the various locutions considered all the *kinds of* content required to state the theory itself" (Brandom 1998, 116). Elsewhere: "[...] the scorekeeping practices that confer conceptual content on the *fundamental sorts of* explicitating vocabulary used in stating the theory [...]" (Brandom 1998, 641). In these quotations, I italicized the qualifications which indicate that the practice need not confer content on the actual terms employed in our theory, but rather on something with an analogical function.

The two most obvious options are that the instituted vocabulary is completely distinct from the theory's language, or that the former is a reduced version of the latter. The second is in my opinion the case with *MIE*, and only in this sense can Brandom reasonably aspire to expressive

---

[5] Illustrated by, e.g.: "What is being claimed is the expressive completeness of the regimented ascriptional idiom, over a certain domain" (see Brandom 1998, 613).

[6] Where stating the theory's content precisely amounts to describing the content-conferring practice.

completeness. On the seven hundred pages of *MIE*, incomparably richer vocabulary is made use of than that which is explicitly introduced via specifying the instituting practices. It is only the weak self-referentiality that is meaningfully in question in *MIE*. (That is fortunate: I doubt anyone would be willing to read a book as thick as *MIE*, composed exclusively of *if… then…*, *believes that*, *is entitled to* and a handful of other regimented locutions that are explicitly introduced by Brandom.)

Maybe we could seek evidence in *MIE* for a position slightly closer to genuine self-referentiality, characterized by an implicit claim that the instituted vocabulary is about as powerful as the book's own idiom in general, not only with respect to specification of the instituting practice (cf. the quotations from Brandom 1998, 116, and 641). But given how tremendously richer the actual language of *MIE* is, this position would be hard to defend. In the following I will try to show that for Brandom's purposes even weak self-referentiality can do well; although it is yet another question whether it is actually achieved in *MIE*.

3. Justification for the claim of expressive completeness

What is Brandom's justification for the claim that his theory of meaning is expressively complete, in the sense of weak self-referentiality as defined above? According to the picture presented in *MIE*, a full discursive practice consists of two layers. First, the basic assertional practice, which confers conceptual content on the non-logical vocabulary, and second, the practice whereby broadly logical expressions are introduced, which is that of making explicit the features and proprieties of the basic practice. This logical superstructure is what makes "merely rational" participants of the basic practice, who are already in possession of simple conceptual content, into "logical creatures", who are able to express their discursive acting in speech, dragging it thus in the space of reasons. Henceforth, I will talk about the *basic practice* instituting the *basic vocabulary*, and about the *full vocabulary* (consisting of both the basic and the logical vocabulary) being instituted by the *full (discursive) practice*.[7]

[7] Brandom's layered picture of discursive practice, mainly in the formulation of the more recent book *Between Saying and Doing*, is outlined in more detail and criticized in Lauer (2012).

Now, this assumed picture involves that participants of the full discursive practice have at their disposal means for making explicit, or specifying, the *basic* practice. But for expressive completeness it is necessary that a practice establishes vocabulary in terms of which that practice *itself* can be specified. By the same assumption, participants of merely the basic practice are not capable of such specification. Brandom therefore needs to show that the conceptual resources instituted by the full discursive practice are rich enough to make explicit this very practice, not just the basic one. Scharp in Scharp (2010, 264) draws attention to this as the point of Brandom's appeal to the notion of *expressive equilibrium*, traced back to Frege's *Begriffsschrift*, which is somewhat hidden in the body of *MIE*.[8]

According to Brandom (1998, 114; cf. also Scharp 2010, 264), an expression is in an expressive equilibrium if the appropriate inferences instituting its content can be made explicit by use of that very expression. The idea seems to be that the practice which institutes, e.g., the conditional *if… then* as explicitating some proprieties of the basic assertional practice itself needs nothing more involved than this conditional to be made explicit. (Note that no harmful circle is involved: the conditional-instituting practice does not presuppose an explicit introduction; it can go on even without ever being made explicit.) The practice conferring the conceptual content on conditionals could be, for instance, made explicit as follows: "Asserting 'if A, then B' amounts to undertaking such a commitment *if* the commitment to A is added, *then* B is undertaken as well." Of course, nothing at all can be made explicit *solely* by means of *if… then*. So the condition for expressive equilibrium should be specified to the effect that it should be possible to explicitate the instituting practice using only the involved expression itself, *together with* some other limited vocabulary; and in Brandom's layered picture it seems natural to appoint the whole (broadly) logical vocabulary to this role.[9] After all, such interrelatedness of logical locutions would correspond with his need to have *all* of his logical expressions in an expressive equilibrium. Less will not do, for then there would be a logical expression (i.e. an expression needed in making the basic practice explicit) whose instituting practice is not subject to specification by partici-

[8] Otherwise, questions of expressive completeness are primarily discussed in the preface and the conclusion chapter of Brandom (1998).

[9] The normative expression *commitment*, which I used in the specification above, is also understood as logical in Brandom (1998).

pants of the full discursive practice, and expressive completeness would thus be lost.

Brandom quite clearly subscribes to the need of expressive equilibrium for the whole logical vocabulary, but unfortunately the matter is not explicitly discussed when various logical locutions are introduced further in the book. So it is hard to see, and would require a properly detailed discussion to determine, to what extent expressive equilibrium, as a necessary condition for expressive completeness of the proposed theory of meaning, is satisfied in *MIE*. Just as it would require a proper discussion to say – another condition, not independent though – how correct Brandom's treatment of logical expressions is as such, regardless of the ideal of expressive equilibrium. Needless to say, such a broad assessment cannot be supplied in this short essay.[10]

Apart from possible critique from within the two-layered picture, there is a serious challenge to this very conception, labelled *the Layer Cake Picture*, by Lauer (2012), who develops an argument introduced by Laurier (2005). Lauer argues, still rather internally to Brandom's overall project, against the claim that the basic, pre-logical practice assumed in *MIE* can be regarded as autonomously discursive. The core of his reasoning is the following. Agents of the basic practice, without any (broadly) logical vocabulary, conceived as discursive scorekeepers, *a fortiori* lack attitude-ascriptional locutions. But these are necessary in order for them to be capable of attributing discursive *attitudes*, rather than merely *statuses*, to other agents. (By Brandom's own commitment in Brandom 1998, 639-640, an agent cannot implicitly attribute a scorekeeping attitude unless she is able to explicitly ascribe that attitude.) And without the ability to keep "double books" on a single agent, to distinguish between what that other agent *acknowledges* commitment to and what he *is* in fact committed to, one lacks a grasp of the objective character of conceptual content. However, the objective, or representational, aspect of conceptual content is taken as a necessary condition for discursivity in *MIE*. It follows that the basic, pre-logical

[10] The main contribution of Scharp (2010) is a particular argument against expressive completeness of Brandom's theory, building on his treatment of the predicate "true" and his commitment to a Kripkean approach to the Liar's paradox. Brandom (2010) accepts the critique, but claims his account of truth in Brandom (1998, Ch. 5) not to be a proper part of his theory of meaning. The notion of truth is not a logical notion appealed to in making explicit the basic practice, he says, so chapter 5 could be left out of *MIE*, with the rest of the project still aspiring for expressive completeness.

practice is also pre-discursive, even if it can stand on its own, and that an autonomous *discursive* practice presupposes at least some of the locutions that are regarded as (broadly) logical by Brandom.

This deconstruction of the layered picture of discursive practice seems convincing to me, and makes Brandom's claim for expressive completeness again more obscure. On one hand, disrupting the neat inner structure does not imply that the full discursive "cake" cannot be expressively complete as a whole. On the other, it is now harder to substantiate why it should be. As I have just shown, in *MIE* the ambition depends on the accessibility of the expressive equilibrium for the introduced logical vocabulary. But even Brandom's notion of the logical is undermined now. More specifically, it cannot be the case any more that there is a layer of expressions which only serve to make explicit the implicitly present proprieties of an independently conceivable practice so that a fully discursive practice is set up. For the implicitly present proprieties of a practice that lacks objectivity of content are insufficient, even when made explicit.[11] Once we accept Laurier's and Lauer's attack on the layered picture, the way to expressive completeness via expressive equilibrium, which was at least sketched in *MIE*, is further blurred for us. First we need to define the logical anew, only then we can claim expressive equilibrium for it.

4. The role of expressive completeness

Detached from the question, whether and how expressive completeness can be achieved, it remains for us to see the significance of this ideal for Brandom's project; to see for what purpose expressive completeness even in the sense of weak self-referentiality can serve Brandom well. While expressive completeness presupposes the expressive equilibrium in *MIE*, it itself appears to be a prerequisite for another equilibrium, the *interpretive*, introduced on the last pages of *MIE*'s Conclusion. Or rather for one particular case of it. Interpretive equilibrium is achieved when members of a community are able to attribute to one another the same attitudes they are adopting themselves (cf. Brandom 1998, 642). In the basic practice, as conceived

[11] Lauer (2012) finally suggests a version of expressivism about logical expressions void of the Layer Cake Picture, under the head of *dialectical expressivism*. But the idea is rather vague and I do not think it can help with the issue of expressive completeness.

in *MIE*, this is not the case, because there the agents adopt scorekeeping attitudes towards others, themselves unable to attribute to them further scorekeeping attitudes above "mere" statuses. On the contrary, the full discursive practitioners are capable of explicit (even iterative) embedding of attributions, and they can therefore attribute any attitude they adopt (cf. Brandom 1998, 642-643).

Admittedly, this last case of interpretive equilibrium does not depend on expressive completeness. But a different case does. The claimed expressive completeness of his theory enables Brandom's glamorous (or disappointing – cf. Rosen 1997, 168) move at the end of *MIE*, a punchline of the whole story. It allows him to let the external and internal interpreting perspective on a fully discursive community coincide. With expressive completeness, discursive practitioners are themselves capable of specifying the practice by which their conceptual resources are instituted; that is, capable of making explicit the implicit proprieties of the instituting practice. So, they are capable of attributing to each other all the statuses and attitudes that *we* (external interpreters who formulate our theory of meaning *qua* specification of the full discursive practice) are capable of attributing to them. In this interpretive equilibrium between them and us, "[e]xternal interpretation collapses into internal scorekeeping. [...] It is recognizing them as us" (Brandom 1998, 644).

And this collapse is a most welcome thing for Brandom, the rationalist. It saves the irreducible normativity of any discursive practice by referring it to our own norms, which cannot be stated once for good. "There is never any final answer as to what is correct; everything, including our assessments of such correctness, is itself a subject for conversation and further assessment, challenge, defense, and correction" (Brandom 1998, 647). Also, the collapse of perspectives seems sufficient to mitigate a particular unease we might have, as external interpreters of a community, concerning the objectivity of their content. Namely, does any genuine objectivity arise on the basis of the distinction between one's acknowledged and "real" commitments, given that this distinction is always made from a particular scorekeeping perspective and is therefore just a matter of two different *attitudes* of the scorekeeper? That is how Brandom's story goes. But once we recognize our interpreting perspective as internal to the community, there is no point for us in denying "real" objectivity: it is *us* (among others), who are to make the distinction, and we are bound to hold *our* commitments dear.

Note, finally, that all it takes to achieve this ultimate goal, the collapse of the external and the internal perspective, is a theory of meaning that is expressively complete in the sense of weak self-referentiality. For this purpose, a central one indeed, it does not matter that the more intuitive desideratum of full self-referentiality is not met. The agents of the specified practice need not talk the same language as we do. Yet they form with us, as long as an interpretive equilibrium is in place, an important "we": that of rational and logical creatures.

I hope to have clarified what good news expressive completeness of the theory presented in *MIE* would mean – if only after having shown how far such news is from being confirmed.

5. Summary

I have contrasted the ideal of full self-referentiality of a theory of meaning with the notion of theory's expressive completeness, for which full self-referentiality is not required, and I have argued that it is only the latter, or a weak self-referentiality, that can be reasonably seen as the ambition of Brandom's *Making It Explicit*. I have shown that Brandom's claim for expressive completeness is based on the idea of an expressive equilibrium, which he hopes to have achieved for his broadly logical vocabulary. However, such an achievement is by no means evident in *MIE*. Moreover, a recent serious challenge to Brandom's underlying picture of discursive practice makes it even more unclear how expressive completeness could be reached by way of an expressive equilibrium. Whether or not it is achieved in *MIE*, I have shown what a desired goal expressive completeness is for Brandom's overall pragmatist and rationalist project, allowing the ultimate coincidence of the external and the internal interpreting perspective on a discursive community.

References

BRANDOM, R. (1998): *Making It Explicit. Reasoning, Representing and Discursive Commitment*. Cambridge – London: Harvard University Press.

BRANDOM, R. (2008): *Between Saying and Doing: Towards an Analytic Pragmatism*. Oxford – New York: Oxford University Press.

BRANDOM, R. (2010): Reply to Kevin Scharp's "Truth and expressive completeness". In: Weiss, B. – Wanderer, J. (eds.): *Reading Brandom. On Making It Explicit*. London – New York: Routledge, 357-359.

GARVER, N. (1996): Philosophy as Grammar. In: Sluga, H. D. – Stern, D. G. (eds.): *The Cambridge Companion to Wittgenstein*. Cambridge: Cambridge University Press, 139-170.

LAURIER, D. (2005): Pragmatics, Pittsburgh Style. *Pragmatics and Cognition* 13, 141-160.

OCELÁK, R. (forthcoming). Giving Expression to Rules: Grammar as an Activity in Later Wittgenstein. To appear in *Human Studies*.

ROSEN, G. (1997): Who Makes the Rules around Here? *Philosophy and Phenomenological Research* 57, 163-171.

SCHARP, K. (2010): Truth and Expressive Completeness. In: Weiss, B. – Wanderer, J. (eds.): *Reading Brandom. On Making It Explicit*. London – New York: Routledge, 262-275.

Organon F 21 (3) 2014: 338-362

The Exploding 'Ought'[1]

DAVID BOTTING

Institute of Philosophy of Language. Universidade Nova de Lisboa
Av. de Berna 26 – 4° piso. 1069-061 Lisboa. Portugal
davidbotting33@yahoo.co.uk

RECEIVED: 22-02-2014 • ACCEPTED: 23-05-2014

ABSTRACT: In this paper I wish to discuss so-called principles of inheritance and the familiar claim that it leads to deontic paradoxes. By combining two such paradoxes it will be shown that inheritance amounts to a principle of explosion: supposing that in the actual world there is at least one thing that one ought to do, almost anything is something one ought to do. I will then attempt to qualify the principle of inheritance so as to avoid this and other paradoxical results.

KEYWORDS: Broome – deontic paradoxes – inheritance – ought – paradox of the Good Samaritan – Ross's paradox – Wedgwood.

1. Introduction

There are eight sections of this paper. In this, the first section, I will set out the structure of the paper. In the second section I will lay out a number of inheritance principles and explain how and in what way they each support inheritance. In the third section I will lay out the axioms of deontic logic and show that if we use these as the semantics of 'ought' then

[1] I am currently involved in two projects sponsored by the Portuguese Foundation for Science and Technology (FCT): "Is moral reasoning essentially dialogical?" (FCT-funded individual post-doctoral project SFRH/BPD/77687/2011) and "Argumentation, Communication and Context" (FCT-funded institutional project in ArgLab PTDC/FIL-FIL/110117/2009).

inheritance turns out to be a fairly obvious consequence of the first two axioms. In the fourth section I will describe paradoxes that inheritance seems to generate, in particular one I christen a Principle of Explosion for oughts. This is the case against inheritance. In the fifth section I will discuss why inheritance is an attractive principle and why we should want to preserve it if we can. This is the case for inheritance. In the sixth section I will discuss, but reject, the possibility of explaining away these paradoxical results as being not as paradoxical as they initially seem. In the seventh section I will try to break the deadlock between the cases for and against inheritance by putting forward new principles for transmitting oughts from ends to means. The qualified inheritance supported by these principles should avoid the problematic paradoxes, though it will leave some problems unresolved. In the eighth section I will lay out my conclusions.

2. The principles

Below are a few examples that express roughly the same idea of inheritance:

> Inheritance$_C$: p semantically entails q ⇒ "It ought to be the case that p" semantically entails "It ought to be the case that q" (cf. Cariani 2009, 1).
>
> Inheritance$_W$: If one ought to E, and it is necessary that (one E's ⊃ one M's), then one ought to M (cf. Setiya – Way n.d., 27).
>
> Inheritance$_S$: If X objectively ought to do A, and to do A X must do B, it follows that X objectively ought to do B (cf. Schroeder 2009, 234).

These are largely the same; however, there may be differences hidden in the modal terms used.

By "to do A X *must* do B" Schroeder (2009, 234) seems to mean nomic necessity or something similar, for he talks of the relation of B to A as a relation of means to ends. While accepting the possibility of alternatives, Setiya – Way (n.d., 3) are much more explicit about what they mean by "necessity" in "it is necessary that (one E's ⊃ one M's)", favouring an epistemic construal where P is epistemically necessary if and only if P is true at all candidates for the actual world not ruled out by the relevant body of information. Clearly, this rules *in* all the logical consequences of the relevant body of information. Lastly, "p semantically entails q" says that there is no model in which p is true and q is false.

All three principles have the result that inheritance is closed under entailment, though they differ in what those entailments can be from. Inheritance$_C$ seems the most modest, for only the entailments of *p* alone inherit from *p* that it ought to be done. Inheritance$_W$ seems the least modest, for all the entailments of the relevant body of knowledge will qualify as epistemically necessary. Schroeder is less forthcoming, but we might suppose that he means whatever can be entailed from the initial conditions of the actual world and its causal laws; then, inheritance$_S$ is closed under entailment in the same way and degree that deductive-nomological explanation is closed under entailment.

Similar in spirit is:

> Inheritance$_B$: ((S requires of N that p) & (p ∈ q) is logically valid) ∈ (S requires of N that q) (Broome 2007, 19).

S is here the source of the requirement and N the agent. This principle says that requirements that a source may generate are likewise closed under entailment, since if (p ∈ q) is logically valid then this means that p entails q. This shares inheritance$_C$'s modesty, entailment being from *p* alone rather than from *p* conjoined with other true propositions (a relevant body of information).

3. The axioms

Inheritance is often supported by appeal to the semantics of 'ought,' or at least to the semantics of the *deliberative* 'ought.' Wedgwood (2006, 137) puts it like this:

> [T]he semantic value of the practical or deliberative 'ought' is determined by the role it essentially plays ... in *practical reasoning* or *deliberation* ... given by the following rule: Acceptance of the first-person statement '$O_{\langle me,t\rangle}(p)$' ... commits one to making *p* part of one's plan about what to do at *t*.

To commit to making *p* part of one's plan is for one's plan to be a proposition that logically entails *p*. If this is a biconditional (as it seems to be), it follows pretty quickly that any *q* within the deductive closure of the plan is a proposition for which '$O_{\langle me,t\rangle}(q)$' is true, that is to say, it is something I ought to do at *t*, and given that *p* is in the deductive closure of the plan

any *q* logically entailed by *p* will *eo ipso* be likewise within the deductive closure of the plan. Thus, ($O_{\langle me,t\rangle}(p)$ and (p ∈ q) is logically valid) ∈ $O_{\langle me,t\rangle}(q)$ turns out to be to be a materially valid inference; to deny the consequent while accepting the antecedent is simply to misunderstand what the deliberative 'ought' means.

Wedgwood (2006, 144-148) spells out these consequences, which turn out to be the axioms of von Wright's original deontic logic. These are:

1) If *p* and *q* are logical equivalents, $O_{\langle me,t\rangle}(p)$ and $O_{\langle me,t\rangle}(q)$ are logical equivalents.
2) If $O_{\langle me,t\rangle}(p \wedge q)$ then $O_{\langle me,t\rangle}(p)$ and $O_{\langle me,t\rangle}(q)$.
3) If $O_{\langle me,t\rangle}(p)$ and $O_{\langle me,t\rangle}(q)$ then $O_{\langle me,t\rangle}(p \wedge q)$.
4) If *p* is logically false, then $O_{\langle me,t\rangle}(p)$ is logically false.
5) If *p* is logically true, then $O_{\langle me,t\rangle}(p)$ is logically true.

Although he does not formulate an inheritance principle explicitly, it is easy to see that it follows from the principles above. If ($p \in q$) is logically valid then *p* and ($p \wedge q$) are logical equivalents, so if $O_{\langle me,t\rangle}(p)$ is true then from (1) above $O_{\langle me,t\rangle}(p \wedge q)$ must also be true, and from (2) above $O_{\langle me,t\rangle}(q)$ will be true. To deny inheritance, then, is to deny that one of these first two principles is true.

4. The paradoxes: the case against inheritance

However, these intuitive principles of inheritance are often held up to be false because of certain counter-examples. These are the deontic paradoxes.

Broome denies that inheritance_B is true because of Ross's paradox. This makes use of the fact that any proposition entails the disjunction of itself and any other proposition, true or false. In short, the rule of ∨-introduction guarantees that whenever *p* is entailed so also is $p \vee q$ for any *q*. Therefore, if $O_{\langle me,t\rangle}$(*I post the letter*) then $O_{\langle me,t\rangle}$(*I post the letter or burn it*). But then I can do something I ought to do by burning the letter, despite the fact that by burning it I cannot post it. This is deeply counter-intuitive (cf. Broome 2007, 20).

Does Broome reject (1) or (2) above? In using a possible-worlds semantics for requirements it seems that he is committed to the same axioms that Wedgwood is committed to, since any possible world in which I post

the letter is a world in which I post the letter or burn it; possible-worlds semantics seem to support automatically the deductive closure of requirements. Broome denies this by making a distinction between the property and a code.

Suppose that the requirement is a rational requirement; then, rationality is its source and being rational is the property. The things that rationality requires are the code and by complying with the code the agent instantiates the property, i.e., he is being rational to the extent that he complies with what rationality requires of him. There may be other properties that the agent, in having the property (e.g., being rational) must necessarily have (e.g., being alive) and propositions that must necessarily be true (e.g., I post the letter or burn it) but these are not part of the code. "Not all propositions that are necessary conditions for having the property need be in the code," says Broome (2007, 15), indicating that (1) above is an axiom only for the property and not the code; logical equivalents and implicata are not substitutable *salve veritate* into the context of a deontic operator when these are being used to express codes.

If so, why is Ross's paradox still a problem? In the property sense it is true, Broome seems to say; the world in which I post the letter is the same world in which I post the letter or burn it, and by occupying this world I am doing what I ought to do or what I am required to do. Something like this seems to be Wedgwood's position also when he says that if we bear in mind its truth-functional meaning, this is not counter-intuitive at all, explaining away the fact that it seems counter-intuitive on the Gricean grounds that it is less informative than what we should say, viz., "I ought to post the letter" (cf. Wedgwood 2006, 149-150). Wedgwood freely admits that this amounts to a kind of *principle of explosion* for oughts:

> *First Principle of Explosion for Ought*: "If there is anything that you ought to do, then whatever you do, you do something that you ought to do" (Wedgwood 2006, 150, ff. 23). In symbols, $O_{\langle me,t\rangle}(p) \in N$ instantiates the property of doing as he or she ought by making q true for any (or any compossible) q.

In other words, since $O_{\langle me,t\rangle}$("I post the letter") is true, I can do something I ought by doing something else, whether it is burning the letter or something quite irrelevant like scratching my finger. Note that it does not follow that $O_{\langle me,t\rangle}$("I burn the letter") or $O_{\langle me,t\rangle}$("I scratch my finger"), but it does follow that by doing these things I would instantiate the same deontic

properties and obey the same deontic axioms as I would if I posted the letter. Wedgwood sticks to his guns that this is not as counter-intuitive as it first appears – there are lots of things that you ought to do, and the problem is only with doing all of them.

Why, then, does Broome reject inheritance$_B$? Simply because it would be strange if the source issued a requirement that could be satisfied in principle by satisfying any other arbitrary proposition, or any proposition that is true in a possible world or even all the possible worlds in which the required proposition is true. It remains counter-intuitive that this can be even slightly rational, that I am satisfying a genuine requirement or doing something that I ought to do when I do this.

It is also counter-intuitive that this can be because of some completely unrelated 'ought.' As the First Principle of Explosion implies, it doesn't actually matter what the derived ought is derived from; as long as there is something that you ought to do, then there are all sorts of things that you ought to do, or all sorts of things by doing which one is to count as rational. The only requirement seems to be that there is some possible world in which they are all true.

Here is another paradox. For any true q, p implies $p \wedge q$ and $p \wedge q$ implies q. This gives us:

> *Second Principle of Explosion for Ought*: If there is anything that you ought to do, then any true proposition whatever is something that ought to be made true. In symbols, $O_{\langle me,t\rangle}(p) \in [q \supset O_{\langle me,t\rangle}(q)]$.

Now, perhaps we might accept the result (as axiom 5 above indicates) that if q is necessarily true then $O_{\langle me,t\rangle}(q)$ is true, but surely it is unacceptable for $O_{\langle me,t\rangle}(q)$ to be true whenever q and $O_{\langle me,t\rangle}(p)$ just happen to be true in the same world. Once again we can derive this result from the first two axioms of deontic logic: if q is true, then p and $(p \wedge q)$ are logical equivalents, so if $O_{\langle me,t\rangle}(p)$ is true then from axiom 1 above $O_{\langle me,t\rangle}(p \wedge q)$ must also be true, and from axiom 2 above $O_{\langle me,t\rangle}(q)$ will be true.

Together with the First Principle of Explosion, this second principle implies:

> *Combined Principle of Explosion*: Provided there is some p for which $O_{\langle me,t\rangle}(p)$ is true, any q that is true or is made true by acting is something that I ought to do because it is a way of satisfying $O_{\langle me,t\rangle}(p \vee q)$ (by the first principle) and because it satisfies $O_{\langle me,t\rangle}(q)$ which (by the second principle) must also be true.

I don't think that an attempt to explain away this result on pragmatic grounds works; it cannot be true, yet it is the direct result of the fact that inheritance is closed under entailment.

5. Derived requirements: the case for inheritance

Perhaps failure to be closed under entailment might be considered no great loss, but it has been noted (Wedgwood 2006; Cariani 2009; Broome 2007) that inheritance has some highly intuitive consequences. One of these is the ease with which one can appeal to inheritance to explain why one fails to do what one ought although one does not initially seem to violate a specific requirement. For instance, although there is no specific 'ought' prohibiting driving at double the speed limit, this is something that one ought not to do because one ought not to drive above the speed limit, and driving above the speed limit is necessary to drive at double the speed limit. Or in an example from Goble discussed at Broome (2007, 21-22), it is because one ought not to camp on public streets at all that one ought not to camp there on Thursday night. Inheritance guarantees this simply on the basis of deductive (arguably material) validity; there are specific oughts for these things because these subsist in the more inclusive ought. Broome responds that no such explanation is needed – when one camps on Thursday night one violates the code that requires one not to camp at all; we do not need the code to provide a new requirement for the specific case, the general requirement will suffice. The temptation to think otherwise is due to conflating the code sense with the property sense once more.

This contrasts with Cariani's view that the semantics of 'ought' should be such that, for instance, I would be correct to assert "I ought to drive at less than double the speed limit," and this because I would be correct to assert "I ought to drive at less than (or at) the speed limit." On the view that 'ought' is a propositional operator these are correct things to assert because O("I drive at less than the speed limit") is true and, because of inheritance$_C$, it follows from this that O("I drive at less than double the speed limit") is true. Although Cariani rejects the view that 'ought' is a propositional operator he takes it as a constraint on his semantics that it should support the correctness of these assertions (see Cariani 2009, 15). Broome's view translated into the language of 'ought' seems to say that we are not strictly speaking correct to assert that I ought to drive at less than double the

speed limit; we are saying something strictly false, because there is no 'ought' whose embedded proposition is "I drive at less than double the speed limit," or in other words, O("I drive at less than double the speed limit"), read now as "S requires of N that N drive at less than double the speed limit" is false.

True, Broome might want to say that it is truly the case that I ought to drive at less than double the speed limit and that what I should not strictly say is that I am required by the source of the requirement S to do this. But after appealing to linguistic data to support his claim that the source sense is the one we typically use – that we do not say, for instance, that morality requires us to be alive despite the fact that being alive is necessary for behaving morally (cf. Broome 2007, 15-16) – it seems fair to wonder whether Broome can afford to be so sanguine about the prospects of explaining it away on the grounds of confusion between the source sense and the property sense; it is not obvious that the oddness of saying "There is no requirement to drive at less than double the speed limit, but there is a requirement to drive at less than the speed limit," or "There is no requirement to drive at less than double the speed limit, but there is a requirement not to drive over the speed limit" is dispelled even restricting ourselves to the source sense. In fact, it is not obvious from what Broome says that there *is* any requirement not to drive over the speed limit, for he could give an analogous explanation of this as for camping on a Thursday. This amounts, as I have said already, to rejecting axiom (1) as applying to codes. But we can easily explain why there are such requirements by accepting inheritance$_B$, the requirements not to drive over the speed limit and to drive at less than double the speed limit following automatically from a requirement to drive at the speed limit or under.

So, at least some entailments seem to be a good thing. One disanalogy between 'good' inheritance and 'bad' inheritance, it might be thought, is that in the cases of 'good' inheritance the inheritance was from a genus to a species or from a more specific 'ought' to a less specific 'ought'; in other words, there was an intensional logical connection as well as an extensional logical connection. There is no such intensional connection between p and $p \vee q$ for arbitrary q or between p and $p \wedge q$ for true arbitrary q. This suggests one plausible way of qualifying inheritance. For instance, you could stipulate that inheritance is not closed under deductive entailment but under what Chisholm (1981) calls *conceptual entailment*: if a thinker cannot have P as the content of a belief without having Q as the content of the be-

lief, then P *conceptually entails* Q. Conceptual entailment differs from deductive entailment in the following ways: 1) P does not *conceptually entail* a conjunction of itself and a necessary truth R, e.g., P does not *conceptually entail* P & 2 + 2 = 4, and; 2) P does not *conceptually entail* P OR Q.[2] When P *conceptually entails* Q and Q *conceptually entails* P, P is *conceptually identical* to Q. This rules out successively, and quite neatly, the first and second principles of explosion above, and rules in what we want to rule in, for I cannot have as the content of a belief that I drive under 65 m.p.h. without having as the content of a belief that I drive under 100 m.p.h.

Relating this back to the axioms of deontic logic, we can modify axiom (1) to (1*): "If *p* and *q* are *conceptually identical*, then $O_{\langle me,t\rangle}(p)$ and $O_{\langle me,t\rangle}(q)$ are *conceptually identical* (or, perhaps, logically equivalent)." Note that this is no longer a possible-worlds semantics – which is extensional – but a much finer-grained semantics based on intensions. Now, if P *conceptually entails* Q then P *conceptually entails* P & Q and, trivially, P & Q *conceptually entails* P. Thus, if P *conceptually entails* Q then P is *conceptually identical* to P & Q. Then, by our new axiom (1*), $O_{\langle me,t\rangle}$(P) and $O_{\langle me,t\rangle}$(P & Q) are *conceptually identical*. Next, by an analogous version of (2) I will call (2*), it follows from $O_{\langle me,t\rangle}$(P & Q) that $O_{\langle me,t\rangle}$(P) and $O_{\langle me,t\rangle}$(Q). So, we can derive a version of inheritance on the basis of *conceptual entailment* in an analogous way as before.

However, an inheritance principle so defined is both too strong and too weak. It is too weak because it does not handle the deontic paradox of the Good Samaritan. If it is the case that I ought to help those in need, then it is logically necessary that there be someone in need, yet clearly one ought not to act so that somebody is in need in order to help them afterwards. Conceptual entailment does not seem to help here, because I cannot have as the content of a belief that I help someone without having as the content of a belief that there is someone to help. I am not introducing any new, arbitrary propositions here.

It is too strong because it rules out ordinary means-ends relationships. Suppose that I ought to review a paper, and in order to do so it is necessary that I accept the commission to review the paper. Intuitively, I ought to accept the commission, and inheritance_S explains why I ought to do this –

[2] P and Q are actually properties, rather than propositions. For Chisholm, when we believe something we attribute a property to ourselves rather than have an attitude towards a proposition. However, I will speak as if they were propositions.

the means inherits from the end for which it is a means the force (or a part thereof) of the 'ought'-claim. There is generally no logical connection between the end and the means, still less an intensional connection, although there is a necessary connection.

6. The bullets

We could save our new principle of inheritance if we could "bite the bullet" for these two problems. In this section I will discuss the possibility of doing precisely this, and note that this has been done in the philosophical literature.

Beginning with the Paradox of the Good Samaritan, the paradox is dissolved as long as we index the 'ought' operator to time, Wedgwood (2006, 150) says, it being the case after the one to be helped is in this situation that one ought to help him, but not before. This makes sense of our intuitions that it is not the case that the person ought to be in this situation, and even that he ought not to be in this situation, and yet, *given* this situation, helping him is what we ought to do. Properly indexed to time we do not get any conflict between the 'ought'-claims O("I help X") and O("X needs my help").

Cariani seems to be referring to much the same thing when discussing 'secondary obligations.' Suppose that Mary ought to turn in her paper by Friday. Then, at all the deontically best worlds available before Friday, Mary has turned in her paper. Saturday comes and Mary has not turned in her paper. Should Mary be punished for not handing in her paper on Friday? In none of the deontically best worlds is Mary punished, for in those worlds Mary handed in her paper. The solution, Cariani (2009, 10) says, is to update the possible worlds on Friday when Mary misses her deadline. Before Friday, Mary ought to turn in her paper by Friday, but after it is too late to hand in her paper and hence the worlds in which she does so are no longer accessible, Mary ought to be punished, without it being the case that Mary ought to have done whatever puts her in this situation. So, $O_{\langle \text{Mary, Friday} \rangle}$("Mary hands in her paper by Friday") is true but $O_{\langle \text{Mary, Saturday} \rangle}$("Mary hands in her paper by Friday") is false, whereas $O_{\langle \text{Mary, Friday} \rangle}$("Mary is punished on Saturday") is false but $O_{\langle \text{Mary, Saturday} \rangle}$("Mary is punished on Saturday") is true. Similarly, at the time that the person comes to be in need, there comes also the obligation to help. Worlds where the person

does not need help or where Mary has passed in her paper are no longer available; the modal claims can only be assessed against this updated background.

This seems to suggest optimism for biting the bullet. However, I am unconvinced. This still seems to amount to a rejection of inheritance$_C$; for the proposition "Mary ought to be punished for not handing her paper in by Friday" entails that "Mary did not hand her paper in by Friday", so if O("Mary is punished for not handing her paper in by Friday") then, by inheritance$_C$, O("Mary did not hand her paper in by Friday"), which implies that Mary ought to have done something for which she ought to be punished, that is to say, something she ought not to have done. Clearly, it cannot be the case that Mary ought to have done something that she ought not to have done. Similarly, I still do not find it obvious why the possible-worlds semantics (or even a more fine-grained semantics) does not license O("X needs my help") for the same indices as O("I help X"), and it seems to me that Wedgwood has answered a different question to the one asked. Granted that without backwards-causation or time-travel there is nothing I can actually do to bring about the situation that X finds himself in, but if I could then surely what I ought to do is not put X into that situation but, on the contrary, to make it so that X is not in that situation, and it seems not unreasonable to expect this to be grounded in the semantics.

Furthermore, even if we accept that it is *not* the case that Mary or X *ought* to be in the situation they are in, what it seems that we want to be able to say is that they *ought not* to be in that situation. That is to say, even if we deny that inheritance has the consequence that O("X needs my help") for the current time it does not itself have the consequence that O("X does not need my help"), either for the current time or earlier, when I may have been able to do something about it. Yet surely if it is true that I ought to help you it must be the case that you are in a situation that you ought not to be in and that I ought to stop you being in were it possible.

Perhaps it might be argued that it is not the semantics of the 'ought'-operator that should license the inference from an 'ought'-claim that is true at some later time to an 'ought'-claim that was true earlier, but it seems strange that inheritance should appear to give precisely the wrong answer. On the contrary, it seems to me that what makes the 'ought'-claim that I should help X true at t is the fact that it inherits from a true 'ought'-claim prior to t that I ought not permit X to be in the situation that I later ought to help him out of. The fact that I was not at that time in a position

to stop X from getting into that situation, and that I am not at the current time in a position to have stopped X from getting into that situation, are contingent features of the situation and not a matter of logic. I would say also that this is something that I ought to do at t even before t. Given that she does not hand in her paper on Friday, it is equally as true before Friday as after that I ought to punish her on Saturday. Time enters into the content of our obligations but not their structure; it is not, I dare to say, an index. The updating mechanism implies instead that one's obligations come into and out of existence.

So, I don't think the Paradox of the Good Samaritan is so easily disposed of.

What about inheritance from means to ends? We noted that the new inheritance principle did not license O(M) on the grounds of O(E) when M is the means to E, yet inheritance$_W$ and inheritance$_S$ certainly did. Are they right to do so?

It might be thought that the force of the 'ought'-claim should not be transmitted from ends to means after all. Suppose that, even if I accept the commission to write the review, I am so disorganized that I am actually unlikely to write the review.[3] It may then be the case that I ought not to

[3] This example is based on the following example of Setiya – Way (n.d., 9): "Professor Procrastinate ... and Professor Dispatch have equally strong reason to review a book. A necessary means to this is accepting the commission to review it. There are no side-benefits to accepting; the only reason to accept ... transmits from the reason to review. Dispatch is extremely likely to write the review, if he accepts. Procrastinate is extremely likely not to write the review, if he accepts. ... Dispatch and Procrastinate have the same reason to accept: namely, as much reason as they have to write the review. But surely Dispatch has more reason to accept than Procrastinate has. We would unhesitatingly advise Dispatch to accept, while being very reluctant to advise Procrastinate to do the same."

This is couched in the language of reasons; Setiya and Way would claim that Professor Procrastinate has reason to write the review (unless, possibly, he knows that there is no chance at all of his writing the review, in which case it is questionable whether this is something he ought to do in the first place) but less reason to accept the commission. They are doubtful whether it is possible to say that there is an all-things-considered reason to write the review but that it is not what one ought to do. Thus, this might be a case where Professor Procrastinate ought to achieve the end (write the review) and cannot do so without taking the necessary means (accepting the commission) but ought not to take the means, since by doing so he would be making it less likely that the review will get written than if he left it to Professor Dispatch. This

accept the commission, for by doing so I may knowingly be probably putting myself into a less deontically good world – a world in which the paper is not reviewed – than I would if I did not accept the commission and left it for someone else to review. It is because of situations like this that Setiya – Way (n.d.) reject inheritance$_W$ altogether. The new principle of inheritance may then be considered to be giving the right result.

The example here seems to me under-described. The fundamental 'ought'-claim involved (that often seems to be left out of the presentation) seems to be O("Someone reviews the paper"). If I am the only one who can review this paper, then it is necessary for the truth of "Someone reviews the paper" that I review the paper. Thus, inheritance should license O("I review the paper"). Next, it is necessary for the truth of "I review the paper" that "I accept the commission to review the paper", so inheritance should license O("I accept the commission to review the paper"). Inheritance$_W$ and inheritance$_S$ do license these 'ought'-claims, giving us:

O("Someone reviews the paper")
⇓
O("I review the paper")
⇓
O("I accept the commission to review the paper")

where ⇓ is inheritance on the grounds of (extra-logically) necessary relations between the propositional contents of the 'oughts'. The new principle of inheritance does not license either, although it does license O("Someone reviews the paper") on the grounds of its being *conceptually entailed* by O("I review the paper") – the converse of the inheritance above.

Now, if I am not the only one who can review the paper then "I review the paper" is not necessary for "Someone reviews the paper" so inheritance does not license O("I review the paper"). To get O("I review the paper") we must appeal instead to a principle based on sufficiency:

> *Means-Ends Transmission Principle*: If you have a reason to do A and doing B is a sufficient means to doing A, you have a reason to do B (cf. Way 2010, 224).

---

seems to be Setiya and Way's intuition: accepting the commission does not inherit from writing the review that it is something that ought to be done for someone in Professor Procrastinate's situation. For this reason, Setiya – Way (n.d., 17) deny that inheritance$_W$ is true.

Since my reviewing the paper is a sufficient means for making it such that someone reviews the paper, this principle means that I have a reason to review the paper. If we admit that at least sometimes (possibly when it is undefeated) this reason amounts to an 'ought'-claim, then I ought to review the paper, and inheritance guarantees the rest, giving us:

O("Someone reviews the paper")
↓
O("I review the paper")
⇓
O("I accept the commission to review the paper")

where ↓ amounts to transmission on the grounds of (extra-logically) sufficient relations between the propositional contents of the 'oughts'.

In both versions, it seems that I ought to accept the commission. To describe the case where it might seem that I ought not to accept the commission in line with Setiya and Way's intuitions we need to include the other agent explicitly, giving us:

O("Someone reviews the paper")	
↓	↓
O("I review the paper")	O("You review the paper")
⇓	⇓
O("I accept the commission")	O("You accept the commission")

This shows (by the double strikethrough) that O("I accept the commission") does not inherit any force from O("I review the paper"), the reason being that you are more likely to review the paper than I am and we cannot both of us review the paper and accept the commission, so that if I accept the commission I make it less likely that you will and consequently less likely that someone reviews the paper.

Note that this could be taken as a counter-example to the Means-Ends Transmission Principle as applied to 'oughts,' rather than inheritance, or perhaps not so much a counter-example as a case where the reason, being defeated, falls short of being one I ought to act on. The only thing that I ought to do is act in such a way that some sufficient means for the end is taken. So, what follows from O("Someone reviews the paper") is O("I review the paper or you review the paper or ...") for the disjunction of sufficient but non-necessary means. Now, it may seem odd that I ought that you do something, but there clearly are acts of omission that I can 'per-

form,' – e.g., not accepting the commission – to help make it such that you review the paper. This is so for any acts by doing which would make it impossible for you to do them or less likely that you do them, as it seems we are meant to assume in this example. To put the matter slightly differently, it could be said that reviewing the paper is something that *we* ought to do, although my contribution is entirely a matter of my *not* doing things. This is because it is not *my* writing the review that the deontically best world requires but only that the review be written. This is to reject Setiya and Way's intuition that I ought to review the paper, it then being no surprise that it is not the case that I ought to accept the commission. In fact, since I ought that you accept the commission, and we cannot both accept the commission, I ought *not* to accept the commission – a stronger result than it *not* being the case that I ought to accept the commission.

Perhaps some of the 'ought'-facts in the deontically best worlds are agent-relative – it is not enough that some state is achieved or some act is performed by someone, but that a particular agent perform that particular act. That is to say, we might drop the assumption that the fundamental 'ought'-claim is O("Someone reviews the paper") or O("The paper is reviewed") but is O("I review the paper") or O("The paper is reviewed by N"). Is it then still possible to claim, when I am very unlikely to write the review, that I ought not to accept the commission? I don't think so. My accepting the commission is necessary for my writing the review, and although you may be more likely to write the review, it is not the writing of the review but *my* writing of the review that is at issue, that is in the deontically best world. Setiya and Way's intuitions depend (not unrealistically) on it not mattering who actually reviews the paper as long as it gets reviewed, but in those cases I have argued that it is not the case for any particular person that they ought to review it, though it may well be true that collectively they ought to bring about its being reviewed.

This means that when there is a genuine 'ought'-claim for an end, the force of that 'ought'-claim should be transferred to necessary means. It remains a problem with the new principle of inheritance that it does not have this result. Along with the Paradox of the Good Samaritan, these bullets are not easily bitten.

7. The means

Instead, I propose we take Schroeder's more instrumentalist-looking principle inheritance$_S$ "If X objectively ought to do A, and to do A X must do B, it follows that X objectively ought to do B" and try to define what "to do A X must do B" means in such a way as to allow transmission from ends to means and from the more specific to the less specific. As a first approximation, I propose that it means that X A's *by* B-ing.

This simple linguistic test of the 'by'-locution seems to rule out what we want to rule out and rule in what we want to rule in. It is false to say "I posted the letter by posting the letter or burning it" and "I helped the needy by putting them in need" and true to say "I drove at less than 100 m.p.h by driving at less than 60 m.p.h" and "I reviewed the paper by (in part) accepting the commission."

The most complete analysis of the by-relation is in Goldman (1970, 43):

> *Act-token* A *level-generates act-token* A' *if and only if*
> (1) A *and* A' *are distinct act-tokens of the same agent that are not on the same level*;
> (2) *neither* A *nor* A' *is subsequent to the other; neither* A *nor* A' *is a temporal part of the other; and* A *and* A' *are not co-temporal*;
> (3) *there is a set of conditions* C* *such that*
> (a) *the conjunction of A and* C* *entails* A', *but neither* A *nor C* alone entails* A';
> (b) *if the agent had not done* A, *then he would not have done* A';
> (c) *if* C* *had not obtained, then even though* S *did* A, *he would not have done* A'.

The relation of *level-generation* is meant to be wider than the by-relation and is not exhaustive of the possible relations between act-tokens: act-tokens are *identical* if they are the exemplification by the same subject (however described) of the same property (where, unlike Goldman, I count those properties the same that are *conceptually identical* as defined above) at the same time; they are on the *same level* if they are not identical but differ only by having different concepts of their objects; one act-token is a temporal part of another when it is one of a series of actions, e.g., each separate action involved in tying a shoelace or changing a tyre; one act-token is co-temporal with another when they both need to be performed independ-

ently at the same time for another action, e.g., the act of jumping is co-temporal with the act of shooting, and these together *level-generate* the act of jump-shooting.

Condition (1) above specifies that act-tokens related by level-generation are not identical or on the same level, but are the act-tokens of the same agent and, because (2) specifies that neither act-token is subsequent to the other, at the same time. The rest of (2) specifies that act-tokens are not in either of the other two relations. Condition (3) says that an act-token A level-generates another A' when it is logically entailed by the conjunction of A with a set of conditions C* but is not entailed by A or C* alone. These conditions C* may be of different types, and these determine what type of level-generation is involved. What should be noted is that C* can be causal facts about the actual world. So, if my act-token of flipping a switch causes a light to go on, then it *level-generates* the act-token of turning on the light, though it does not *cause* my turning on the light, which is a different event from the light's going on.

Let us look at some examples of inheritance on the basis of *level-generation*, being a temporal part, and being a co-temporal part. If flipping a switch were the only way to turn on the light, then O("I turn on the light") should transmit its force to O("I flip the switch"). As for the temporal parts of necessary acts and the co-temporal parts of necessary compound acts, these do not seem to inherit any oughtness from each other (from one temporal or co-temporal part to another) but from the act that they are the temporal or co-temporal parts of; I ought to grab each end of my shoelaces if I ought to tie my shoelaces and I ought to jump and I ought to shoot if I ought to jump-shoot. What is also interesting is that C* can, in fact, contain these 'ought'-facts (cf. Goldman 1970, 25). For instance, if C* is O("I tie my shoelaces") then my tying my shoelaces *level-generates* my doing something I ought to do. Similarly, if I ought to turn on the lights, my turning on the light *level-generates* my doing something that I ought to do and since my flipping the switch *level-generates* my turning on the light, it also generates my doing something that I ought to do, for *level-generation* is transitive.

How does this work with our previous examples? Although it seems true to say "I drove at less than 100 m.p.h by driving at less than 65 m.p.h" it is not obvious how the act-tokens of my driving at less than 100 m.p.h and my driving at less than 65 m.p.h are related. It does not seem to be *level-generation*, because driving at less than 65 m.p.h entails on its own

driving at less than 100 m.p.h, thus violating 3(a) which says that A should not entail A' on its own. I propose to treat this as a different way of being on the *same level*. In general, where two property-instances are related such that they are either *conceptually identical* or by a more inclusive one containing a less inclusive one (one *conceptually entails* the other), then the property-instances are on the same level.

This seems to suggest the following principle of inheritance:

Inheritance*: If X objectively ought to do A, and either:
- a) A-ing and A'-ing are not distinct;
- b) A-ing and A'-ing are on the same level;
- c) A' level-generates A;
- d) A' is a temporal part of A;
- e) A' is a co-temporal part of A;

it follows that X objectively ought to do A'.

This applies recursively. If I objectively ought to turn on the light, then since my flipping the switch level-generates my turning on the light, then condition (c) is satisfied and I objectively ought to flip the switch. Since I ought to flip the switch, and (let us suppose) my moving my finger in a certain way is a temporal part of my flipping the switch, it follows from (d) that I objectively ought to move my finger in the required way.

However, this does not quite seem to work. One problem is my treating *conceptual entailment* as a way of being on the same level, for being on the same level is symmetric and implies [by satisfying condition (b)] not only correctly that if I objectively ought to drive at less than 65 m.p.h. then I objectively ought to drive at less than 100 m.p.h, but also incorrectly that if I objectively ought to drive at less than 100 m.p.h. then I objectively ought to drive at less than 65 m.p.h. This can be solved by modifying condition (b) to say that A' does not conceptually entail A. So, if A is driving at less than 65 m.p.h. and A' is driving at less than 100 m.p.h., A *conceptually entails* A' and A' inherits from A the force of O("I drive at less than 65 m.p.h") and make it true that O("I drive at less than 100 m.p.h"). But if A is driving at less than 100 m.p.h. and A' is driving at less than 65 m.p.h., A' *conceptually entails* A and even if O("I drive at less than 100 m.p.h") is true [which it will be if O("I drive at less than 65 m.p.h") is true] A' does not inherit any of its force. The inheritance can only go in the same direction as the *conceptual entailment*, which is both ways if A and A' *conceptually entail* each other, that is to say, they satisfy condition (a) in being not dis-

tinct. In this case it is reasonable for inheritance to be symmetric, as it was in deontic axiom (1*).

There are apparently less reasonable ways in which inheritance can be symmetric on inheritance*. We have already seen one: my reviewing the paper is a means for making it such that the paper is reviewed, so O("I review the paper") follows on this basis from O("Someone reviews the paper"). But my reviewing the paper *conceptually entails* that the paper is reviewed, so O("Someone reviews the paper") follows from O("I review the paper"). There are two possible ways my reviewing the paper and somebody's reviewing the paper seem to be related here, and inheritance goes one way according to one such relation and the other way according to the other.

What about the relation between my accepting the commission and my reviewing the paper? Here, reviewing the paper seems to be subsequent to accepting the commission, so these act-tokens do not seem to be in any of the relations described, for in all of these the act-tokens occur at the same time. The account above does not appear to deal with precisely the kind of means-end relationships that we introduced it for. Goldman (1970, 52-53) calls acts like accepting the commission "putting oneself in a position to do x".[4] Here, I do not think it is too much of a stretch to talk of A' being a cause of A, especially if we take this in the minimal sense of being an INUS condition of A. It is necessary for my reviewing the paper that all the conditions in C* that together level-generate my reviewing the paper from the basic actions I perform are true, and one of these, rather trivially, is that there is a paper to review. My accepting the commission is necessary for the truth of this condition. This is a case where there is a causal relation (weaker than causal necessitation) between the act-tokens.

---

[4] Goldman claims that by virtue of believing that act A' puts oneself in a position to do A and that we want to do A, we want also to do A' and our A'-ing is intentional. However, this claim does not help us. For one thing, nothing in inheritance* implies that it is only our intentional actions that we ought to do – if my writing the review causes a fly to move and thereby generates my act of moving the fly, and I ought to write the review, then I ought to move the fly. Remember that the content of the 'ought'-claim concerns an act-token, and this *particular* act of moving the fly is something I ought to do because without having done it I could not have reached the state of having written the review. Also, inheritance* transfers the force of the 'ought'-claim from the end to an actual means, not to what might only be believed to be a means.

However, this leads us back to something very like the paradox of the Good Samaritan, because it is necessary for my helping someone that there is someone to help, and I can put myself into a position to help someone by acting in a way that makes them need help. Not every act that is a cause of A is something I ought to do because A is something I ought to do, and it may well be something that I ought not to do. What is worse, we have not solved the original paradox of the Good Samaritan, since it is *conceptually entailed* by my helping someone that there is someone to help, and thus from (b) I objectively ought to be such that there is someone to help.

On the plus side, we have solved the problem of the first principle of explosion and Ross's paradox. Since the entailment in condition (3(a)) is a logical one and thus if $(C^* \wedge A) \vdash A'$ then equally $(C^* \wedge A) \vdash (A' \vee B)$ this might seem not to be the case, but here the counterfactual condition (3(b)) blocks this result; it is not true that if the agent had not done A, then he would not have done $A' \vee B$, because he might have done B and thereby done $(A' \vee B)$ without doing A. A does not level-generate $(A' \vee B)$; I do not post the letter "by" posting it or burning it.

What about the second principle of explosion? If $(C^* \wedge A) \vdash A'$ then equally $(C^* \wedge A) \supset (A' \wedge B)$ for any true B. But we can only say that $(C^* \wedge A) \vdash (A' \wedge B)$ when $(C^* \wedge A) \vdash B$, in which case, since $(C^* \wedge A) \vdash A'$ and $(C^* \wedge A) \vdash B$ it will be the case that O(A') and O(B) when O(A), and these can be conjoined in the normal way to give $O(A' \wedge B)$. In the general case when it is not true that $(C^* \wedge A) \vdash B$, A will not level-generate $(A' \wedge B)$ just because it level-generates A', and it will be only A' itself and not $(A' \wedge B)$ that inherits from A. Nor does $(A' \wedge B)$ inherit from A', for although A' and $(A' \wedge B)$ are on the same level it is $(A' \wedge B)$ that *conceptually entails* A'; A' would inherit by virtue of $O(A' \wedge B)$ were this 'ought'-claim true but $(A' \wedge B)$ does not inherit by virtue of O(A').

This suggests the following modified principle:

Inheritance**: If X objectively ought to do A, and either:

a) A-ing and A'-ing are not distinct;
b) A-ing and A'-ing are on the same level, and A' does not *conceptually entail* A;
c) A level-generates A';
d) A' is a temporal part of A;
e) A' is a co-temporal part of A;
f) A' is a cause of A;

it follows that X objectively ought to do A'.

This is a principle for necessary means, when there is one and only one way to A.

What if A-ing is merely sufficient for A'-ing, as it was when I could bring about the paper being reviewed either by my reviewing the paper or by your reviewing the paper? There I said that it did not follow that I ought to review the paper. Or a more trivial case when I ought to have lunch but I have a choice between several options. In these cases I would say that although it is not the case that X ought to A' it is the case that X ought to *intend* to A'.

This might seem odd, and I will not offer a complete defence here. Should X really intend to take each and every option? Doesn't this imply that he ought to have two lunches? To avoid this we have to suppose intentions to be interacting in a particular way. Suppose that I have to choose between a ham sandwich and a yoghurt for lunch, and decide on the ham sandwich. My intention to have the ham sandwich for lunch is "live," so to speak – it is transmitting information to my motor centres and feeding back into my cognitive function so that, typically at least, I believe that I will eat the ham sandwich and not eat the yoghurt. However, suppose that my intended action is thwarted by the discovery that I have no bread. Do I then have to go through the decision-making process again in order to form the intention to eat the yoghurt? It seems more economical to suppose that I already had this intention and that it simply became "live" when I was no longer able to do as I originally intended; at this point I believe that I will eat the yoghurt and that I will not eat the sandwich.

If it is felt that this is too much of a distortion of the ordinary concept of an intention, we can say instead that I am disposed to form this intention. These intentions or dispositions to intend are normatively guided by obligations to have intentions, which obligations are local in nature. So, if asked why I did not eat the ham sandwich it is correct to respond that I ate the yoghurt instead, and that by doing so I did something that I ought to do, i.e., have lunch. My response is not here an explanation of why my obligation to have the intention to eat the ham sandwich does not count – it does not cite an exception – but on the contrary accepts that there is this obligation and that it was violated but that this is a local violation necessitated by my complying with the competing obligation to intend to eat the yoghurt. If it were genuinely an exception there would be no normative question to answer. So, it is not the case that I ought to eat the yoghurt, even when this carries out an intention that I ought to have, and even

though I do something that I ought to do by doing so. It is not inconsistent, on this view, to have competing intentions, even intentions that cannot be jointly satisfied.

I think that this result is quite general, and not for only those that are not necessary. For everywhere that a preparatory act is necessary, it is possible for there to be an obligation not to intend that act. So, one could have an obligation not to intend to accept the commission, and also have, because the paper ought to be reviewed, an obligation to intend to accept the commission. One has, or is disposed to have, both intentions, despite their being contradictory, depending on which aspect of the situation one is attending to. Here, since accepting the commission is not more likely to be deontically superior to not accepting the commission, and given that one is already in the state of not accepting the commission, the obligation to have the intention to not accept the commission is *escaped* by virtue of being already in its end-state. When an obligation to have an intention is escaped in this way any obligations to have the inconsistent intention is not one that one ought to have all things considered, for to satisfy this intention would involve consciously acting in such a way as to lead from a deontically better world to a worse. If the world one is already in is deontically better than the ones one could reach through intentional action, it cannot be the case that one ought to act, even if by doing so one satisfies an intention that one ought, in some dispositional sense, to have.

I think that, finally, we can use this in response to the Paradox of the Good Samaritan. If I ought to help one in need, then I ought to intend to be such that there is one in need, and if I ought to intend this then I ought also to intend to make it so that there is one in need. However, quite independently of these 'ought'-facts it is also true that I ought to intend not to make it so that there is one in need. The first of these intentions is not "live" – it is not one for which "I will make it so that there is one in need" is believed to be true or can be made true outside of the science-fiction possibilities of backwards causation or time-travel. Hence, it is the second of these intentions that one ought to have all things considered. Note, however, that this relies on actions having no deontic values of their own, so to speak. If deontically better worlds are defined in terms of the number of good acts rather than the goodness of the states of affairs resulting from those acts, then it might be true after all that it ought to be the case that there be people in need and who need us to perform good acts (even if it ought not to be the case that we should act purposefully to create this op-

portunity for ourselves), for good acts presuppose that the world is not deontically best as it is. Deontically inferior states of affairs are the necessary evil of deontically superior acts. Some responses to the problem of evil paint a picture rather like this.

So, having the intention that there be people in need does inherit from the fact that I ought to help the one in need that it is an intention that I ought to have, but it is not one that I ought to have all things considered in the sense of its being live, on account of the fact that independently I ought to have the intention for there not to be people in need. The principle underlying this is:

> Inheritance of Local Obligation to Intend: If X objectively ought$_L$ to do A, and either:
>
> a) A' level-generates A;
> b) A' is a temporal part of A;
> c) A' is a co-temporal part of A;
> d) A' is a cause of A;
>
> it follows that X objectively ought$_L$ to intend do A'

where ought$_L$ is what I have called a local ought and one that sometimes ought all things considered to be violated. This principle is meant to supplement rather than replace

> Inheritance of Obligation to Act: If X objectively ought to do A, and either:
>
> a) A-ing and A'-ing are not distinct;
> b) A-ing and A'-ing are on the same level, and A' does not *conceptually entail* A;
> c) A' level-generates A;
> d) A' is a temporal part of A;
> e) A' is a co-temporal part of A;
>
> it follows that X objectively ought to do A'.

Inheritance of Obligation to Act is Inheritance** minus the inheritance from effect to cause and is still meant to express what objectively ought to be done all things considered. This principle preserves the intuition that one ought to drive at less than 65 m.p.h. because one ought to drive at less than 100 m.p.h. but it does not follow from Inheritance of Local Obligation to Intend that one ought$_L$ to intend to drive at less than 100 m.p.h. I think this is correct, since not everything that ought to be done ought to

be done intentionally. However, if there is an intention to A, an intention to A' will follow automatically anyway by the definition of *conceptual entailment*, because it will be impossible to have A as the content of an intention without also having A' as part of the content of an intention. At the other end of the spectrum, it follows from Inheritance of Local Obligation to Intend that I ought$_L$ to intend to make somebody needy but it does not follow from the fact that it is a cause that I ought to make somebody needy, there being no longer inheritance from effect to cause in Inheritance of Obligation to Act. Unfortunately, it does still seem to follow because of the conceptual entailment conditions; I still have not managed to rule this out and must leave it as an outstanding problem.

For the cases in between that are most naturally in a by-relation one both ought to do them and ought$_L$ to intend to do them. This is because conditions (a) to (c) of Inheritance of Local Obligation to Intend are the same as conditions (c) to (e) Inheritance of Obligation to Act. This even applies, I would say, to side-effects. If it is the case that I ought all things considered to write the review it must be that any negative consequences of writing the review, including reasons I may have against bring about the side-effects, must be outweighed. The picture is less clear, I think, with necessary means. Must it be in this case also that the negative consequences of preparatory (or, indeed, subsequent) acts are equally outweighed? The problem is that at the time of the preparatory act it seems possible that there may be more reasons against it than for it, as there was for purposefully making somebody needy. Thus, only an obligation to have an intention follows for causes.

8. Conclusion

In this paper I have discussed the view that inheritance principles license paradoxical results, including the impermissible result that everything is something that one ought to do provided that something is something that one ought to do. I proposed two principles for inheritance from ends to means, where I defined the relation of ends to means following ideas from Goldman (1970). The first of these was a principle for inheritance from ends to necessary means and was meant in part to account for a plausible semantics of 'ought'-sentences where the truth of "I ought to A" implied that there was an 'ought'-fact "A" rather than, as Broome seemed to

suppose, an 'ought'-fact "B" where "A" was an implication of "B" but not synonymous with it. The second of these was a principle for inheritance from ends to sufficient means and I argued that it was not the case that one ought to take the sufficient means but that one ought to intend to take the sufficient means. These 'oughts' are local but strict, in that any time they are not complied with counts as a violation and must be normatively justified by virtue of a competing 'ought.' Exactly how these local obligations interact has not been fully worked out, but one of its principal features is that it allows an agent to have intentions – and says that the agent even should have those intentions – even in cases where those intentions are not ones that can be carried out or are inconsistent with other intentions the agent may have (see Botting 2013).

References

BOTTING, D. (2013): The Distinctive Rationality of Intentions. *Organon F* 20, No. 4, 507-526.

BROOME, J. (2007): Requirements. In: Rønnow-Rasmussen, T. – Petersson, B. – Josefsson, J. – Egonsson, D. (eds.): *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz.* Available at: www.fil.lu.se/hommageawlodek

CARIANI, F. (2009): 'Ought' is not a Box. Available at: www.philosophy.northwestern.edu/people/faculty/documents/OughtBasicFinalCut.pdf

CHISHOLM, R. (1981): *The First Person: An Essay on Reference and Intentionality*. Minneapolis: University of Minnesota Press.

SCHROEDER, M. (2009): Means-End Coherence, Stringency, and Subjective Reasons. *Philosophical Studies* 143, 223-248.

SETIYA, K. – WAY, J. (n.d.): Instrumental Transmission. Available at: http://sophos.berkeley.edu/kolodny/InstrumentalTranmission5.pdf

WAY, J. (2010): Defending the Wide-Scope Approach to Instrumental Reason. *Philosophical Studies* 147, 213-233.

WEDGWOOD, R. (2006): The Meaning of 'Ought'. In: Russ, S.-L. (ed.): *Oxford Studies in Metaethics*. Oxford – New York: Oxford University Press.

Organon F 21 (3) 2014: 363-380

Záměr Wittgensteinova *Traktátu* (1): Předmluva a motto

PETR GLOMBÍČEK

Filozofická fakulta. Ostravská univerzita v Ostravě
Reální 5. 701 03 Ostrava. Česká republika
Filozofický ústav Akademie věd České republiky, v.v.i.
Jilská 1. 110 00 Praha 1. Česká republika
glombicek@osu.cz

ZASLÁN: 08-03-2014 ▪ AKCEPTOVÁN: 14-05-2014

ABSTRACT: The article presents a first part of an interpretation of the intention of Wittgenstein's *Tractatus*. The intention itself used to be considered a rather marginal topic until so called new-Wittgensteinian interpretations. The present article considers main sources to show what kind of content we can ascribe to the book. Its aim is to prove that *Tractatus* is not purely practical exercise, however without stripping the book of its therapeutic side. The first part considers the preface and the motto of the book.

KEYWORDS: Ferdinand Kürnberger – new wittgenstein – Paul Ernst – *Tractatus* – Wittgenstein.

Oč šlo Wittgensteinovi při psaní *Traktátu*? Taková otázka může znít u tak známé knihy banálně. Zároveň se interpretaci *Traktátu* za posledních padesát let věnovalo nebývalé množství monografií. V následujícím textu by měla interpretace předmluvy knihy a jejího motta společně s podporou ve Wittgensteinových *Zápisnících 1914-16* a jeho korespondencí z doby mezi dopsáním a vydáním knihy, ve které se pokouší čtenářům, na nichž mu zvlášť záleží, pointu knihy přiblížit, tedy uceleně říci, jaký cíl autor *Traktátu* se svou knihou v době jejího dopsání spojoval.

Tento úkol je kupodivu v sekundární literatuře zanedbáván. Zároveň ovšem *Traktát* patří mezi nejznámější příklady obtížně srozumitelných filo-

sofických knih. Pýthický styl Wittgensteinova psaní je zlopověstný a jeho kniha se přes svůj nevelký rozsah zabývá podle všeho ohromnou šíří témat od ontologie, přes epistemologii, sémantiku, logiku, etiku, dotýká se mystiky, estetiky, teorie subjektivity, filosofie vědy. Určitě by bylo namístě vykládat jednotlivá témata v kontextu celkového záměru knihy. A ten se nezdá být zcela jasný.

Názorně to ukázaly diskuse posledních let, ve kterých se střetli tzv. noví wittgensteiniáni, podle nichž je *Traktát* víceméně výhradně terapeutické dílo, jež je vědomě obsahově sebezničující, se zastánci tradičnějšího pohledu, podle něhož lze v *Traktátu* nalézt nějaké konkrétní nauky (hlavně asi tzv. zobrazovací teorie významu, případně nějaká ontologie atp.).[1] Přínos nových wittgensteiniánů je možná především právě v tematizaci celkového záměru *Traktátu*, který tradiční interpretace spíše přehlížely. Novým wittgensteiniánům ale byla kromě různých vnitřních nejasností vyčítána snaha oddělit v knize (podle nich záměrně nesmyslný) obsah od rámujících plnohodnotných formulací s tím, že kritéria, kde přesně toto rozdělení vést, se zdají diskutabilní. Ve světle této nedávné debaty je opět jasnější, že stojí za to zabývat se samostatně záměrem, s nímž byl *Traktát* napsán.

Podíváme se tedy nejprve, jak je smysl knihy deklarován v její předmluvě. Zkusíme také zjistit, co se o cíli knihy můžeme dozvědět z jejího motta. Více nám ale pomůže Wittgensteinova korespondence. Především máme k dispozici známý dopis nakladateli Ludwigu Fickerovi, kterému Wittgenstein nabízel svou knihu k vydání a přitom se mu snažil vysvětlit její relevanci. Pak máme korespondenci s Bertrandem Russellem, kterému Wittgenstein svůj text poslal při první příležitosti k přečtení, ochotně odpovídal na jeho otázky, a hlavně ho upozorňoval na celkový záměr knihy, který podle Wittgensteina Russellovi unikal. Kromě Russellova názoru Wittgensteina podle všeho nejvíce zajímalo, co na jeho rukopis řekne Gottlob Frege. Bohužel se nám nedochovaly Wittgensteinovy dopisy a nezbývá než se spokojit s Fregovou stranou výměny, ale i ta nám poskytuje několik vodítek. V pozadí celé interpretace samozřejmě stojí také Wittgensteinovy přípravné texty, především pracovní strojopis knihy, známý jako *Prototractatus* a podklad pro něj, Wittgensteinovy (nedopatřením dochované) *Zápisníky 1914-16*.

[1] Nejznámější jsou texty Cory Diamond a Jamese Conanta. Klasickým svazkem k tématu pak byl sborník Read – Crary (2000). Dostupné české shrnutí s odkazy na další literaturu viz Došek (2013, 222-250).

Z těchto materiálů vysvítá, že Wittgenstein bral obsah své knihy vážněji, než by se mohlo zdát z novo-wittgensteiniánských interpretací, zároveň však potvrzují to hlavní, s čím noví wittgensteiniáni přišli: Wittgenstein nemínil svou knihu jako teoretický příspěvek, resp. jako pouhé sdělení nové informace. To, co v knize vystupuje jako jednotlivé filosofické teze, bychom měli chápat čistě jako nástroje pro dosažení celkového cíle, kterým je léčit potřebu řešit filosofické problémy tak, že bychom hledali odpovědi na filosofické otázky. To ovšem podle všeho nemá znamenat, že by se knize úplně upíral obsah. *Traktát* obsahuje podle předmluvy nějaké „pravdivé myšlenky" (takže není čistě praktickou terapií bez doktrinálního obsahu). Nicméně, tyto myšlenky mají podle všeho charakter poněkud meta-filosofický: odpovídají na otázku po možnostech filosofie, po možnosti vynášet soudy na filosofická témata. Děje se tak ovšem zkoumáním teorie souzení, aby se předvedla specifická povaha souzení o filosofických tématech. *Traktát* si klade za cíl předvést, že starosti, které dostávají výraz ve filosofických otázkách, jsou legitimní, nicméně snaha hledat na specificky filosofické otázky odpovědi je patologická (i když nemusí být bezcenná), takže nestačí k tomu, aby nás starostí za těmito otázkami spolehlivě zbavila. Kniha se omezuje na tento negativní cíl zbavit čtenáře nadějí, jež by mohl spojovat s doktrinální filosofií. Nenabízí vlastní všelék na sklony k filosofování. Wittgenstein totiž podle všeho tyto sklony samotné nechápe jako cosi chorobného.

Ke svému cíli chce kniha dospět vypracováním rozlišení mezi tzv. říkáním a ukazováním jako sémantickými kategoriemi. Pochopíme-li jednou toto rozlišení, měli bychom se zbavit nutkání konstatovat (a tedy *říkat*) to, co spadá do kategorie ukazování, resp. měli bychom přestat obě kategorie směšovat, což by nás mělo také osvobodit od falešných filosofických nauk, otázek a snad aspoň zčásti také od samotných starostí, jež tyto otázky mají vyjadřovat. Tento článek se už nemůže zabývat povahou obou kategorií a jejich odlišením, natož uplatněním, ale zkusmo bychom mohli pro ilustraci tento rozdíl přirovnat k rozdílu, jaký Kant vede mezi pojmy a idejemi čistého rozumu. Ty první se uplatňují v soudech, jež nám poskytují poznatky – a ty se mohou týkat jen smyslové zkušenosti. Ty druhé by bylo tradiční filosofickou chybou uplatňovat v soudech, jež by si činily nárok na roli poznatků, avšak jsou nepostradatelné jako principy jednání včetně organizace a zaměření celku našeho poznání.

Článek se, jak řečeno, musí vyvarovat snahy suplovat úplný výklad celé knihy. Nemůže tedy vypracovat rozlišení říkat/ukazovat, nemůže zkoumat, jak dalece se podařilo deklarovaný cíl knihy naplnit (a zda je to vůbec mož-

né), nemůže doložit, které pasáže *Traktátu* je namístě brát jako vážná plnohodnotná tvrzení (výraz pravdivých myšlenek, jež slibuje předmluva) a které případně nikoli. Jinak se ale pokusí dosavadní tvrzení doložit. V této první části začnu od předmluvy knihy a jejího motta. K ostatním pramenům se dostanu v dalším díle.

Na rozdíl od většiny interpretací tedy nebudu vycházet ze závěru knihy a ukazovat, k čemu se v ní dospělo. Nepůjde mi o to, rozhodnout spor tradičních interpretů s novými wittgensteiniány v celé šíři. Hodlám jen dokázat, že Wittgensteinovy deklarace celkového záměru knihy počítají podstatně jak s netriviálním obsahem knihy, tak s jejím praktickým přesahem (a říci více o tom, jak tento přesah vypadá). V tomto ohledu půjdu střední cestou mezi oběma krajnostmi. Je-li ale tento výklad správný, pak by bylo (čistě z předmluvy a dalších deklarací záměru) jasné, že novowittgensteiniáni se mýlí, pokud se domnívají, že Wittgensteinův záměr byl čistě praktický a kniha neměla mít žádný doktrinální obsah. Zůstala by jim ovšem zásluha za důrazné upozornění na význam praktického cíle knihy, který se tradičně zanedbával a který bychom (také díky nim) formulovali hlavně ve druhé části článku trochu jinak a snad aspoň o něco jasněji, než se to obvykle děje, nakolik to ovšem lze provést bez explicitní interpretace závěru knihy. Po zde předloženém výkladu také zůstává novowittgensteiniánům otevřená cesta dokazovat, že obsah knihy neodpovídá autorovým deklaracím z předmluvy a vysvětlením jejího záměru v korespondenci. To by znamenalo tvrdit, že text knihy sám o sobě poukazuje k jinému autorskému záměru, než byl záměr historického autora.

1. Předmluva Traktátu

Podívejme se nejprve na formulaci cílů knihy v její předmluvě. Jednak lépe porozumíme problému, s nímž se potýkáme, jednak se snad dobereme prvních odpovědí.

> Snad bude této knize rozumět jen ten, kdo už na myšlenky, jež jsou v ní vyjádřeny – nebo na myšlenky podobné – připadl. – Není tedy učebnicí. – Jejího cíle by bylo dosaženo, kdyby tomu, kdo ji čte s porozuměním, přinesla potěšení.
> Kniha pojednává filosofické problémy a ukazuje – jak věřím – že způsob, jak se tyto problémy kladou, spočívá na neporozumění logice našeho jazyka. Celý smysl knihy by se asi dal shrnout slovy: Co se vůbec dá říci,

dá se říci jasně; a o čem nelze mluvit, k tomu se musí mlčet. (Wittgenstein 2001b, 2)

V předmluvě autor deklaruje, že kniha není učebnicí. Má udělat radost čtenáři, jemuž je určena, totiž tomu, kdo o tématech v ní probíraných už sám přemýšlel.[2] Těmito tématy jsou „filosofické problémy". Přičemž autor doufá, že kniha ukázala, že formulace filosofických problémů stojí na neporozumění logice našeho jazyka. Formulovat filosofický problém tedy podle něj znamená jít proti jazykové logice. Filosofické problémy by tedy vůbec neměly být formulovány.

Zde je namístě připomenout, že výraz „jazyková logika", resp. „neporozumění logice našeho jazyka" bychom neměli v tomto textu chápat jako nějaký technický termín rané analytické filosofie. Podle vlastního vyjádření Wittgenstein tento termín, společně s relevantním kontextem, převzal z doslovu Paula Ernsta k pohádkám bratří Grimmů.[3] Později, stále s odkazem na Paula Ernsta, dokonce používá výraz „mytologie ve formách našeho jazyka", který lze nalézt už u Nietzscheho (byť asi v trošku posunutém významu).[4] Logikou se zde nemíní jednoduše formální logika, nýbrž spíš

---

[2] Wittgenstein v korespondenci s překladatelem *Traktátu* do angličtiny kupodivu zdůrazňuje jednotné číslo, jakoby zamýšlený čtenář byl jenom jeden. Kdybychom hledali takového empirického čtenáře, byl by jím nejspíše Russell, který ale sotva byl jediným čtenářem, v něhož autor doufal. Srov. Wittgensteinův dopis Ogdenovi z 10. května 1922 – Wittgenstein (1973, 49).

[3] „Wenn mein Buch je veröffentlicht wird so muss sich in seiner Vorrede die Vorrede Paul Ernsts zu den Grimmschen Märchen gedacht werden die ich schon in der Logisch-philosophischer Abhandlung als Quelle des Ausdrucks ‚Missverstehen der Sprachlogik' hätte erwähnen müssen" (MS 110, 184, 20. 6. 1931). Srov. Baker – Hacker (2005, 314n.).

[4] Viz Wittgenstein (2001a, 187). Srov.: „Es liegt eine philosophische Mythologie in der Sprache versteckt, welche alle Augenblicke wieder herausbricht, so vorsichtig man sonst auch sein mag" Nietzsche, Friedrich, *Menschliches Allzumenschliches* II: § 11. Srov. Baker – Hacker (2005, 314n.); Srov. také Künne (1996). Ernst pro svůj doslov podle všeho zkompiloval několik svých starších textů, takže konkrétně odstavec, který Wittgensteina zaujal, lze nejsnáze najít v textu „Die Stoffe und Der Dichter" v knize Ernst (1940): „Der weitaus überwiegende Teil der noch heute verwendbaren Motive und Sujets stammt ganz bestimmt nicht aus der Wirklichkeit. Er ist oft uraltes Gut der Völker, auf rätselhafte und immer noch nicht genügend erklärte Weise bei den entferntesten und verschiedensten Völkern auftretend: entstanden durch Wandlungen der Sprache, indem eine spätere Zeit die Sprachlogik der Vergangenheit nicht mehr verstand und durch Erfindungen deutete; durch Wandlungen der Anschauungen über den Weltzu-

(opět wittgensteinovsky řečeno) *gramatika*: obecně zákonitosti fungování daného jevu, konkrétně zacházení s jazykem, resp. života v něm. Prvním použitím výrazu „logika" v předmluvě *Traktátu* se ocitáme blíže Wittgensteinovým pozdějším komentářům k Frazerově *Zlaté ratolesti* než Fregovu *Pojmopisu*. Je-li námětem *Traktátu* filosofie, pak hlavně jako soubor praktik, které chceme pochopit, respektive jako soubor praktik, jež se neadekvátně pokoušejí urovnat zmatek daný odcizením běžnému jazyku. *Traktát* má tedy podle svého autora obsah i praktický účel jako metafilosofická kniha. Ernst totiž v doslovu vysvětluje, že pohádky představují pokusy „racionalizujícím příběhem vyřešit problém, který nelze řešit zkušeností se skutečností" a to tam, kde se naši předkové vyrovnávali se světonázorem ještě starších dob v situaci, kdy „pozdější doba už nerozuměla jazykové logice minulých dob." Ernst uvažuje o pohádkách jako dobových racionalizacích starších světonázorů, jež vývojem jazyka přestaly být srozumitelné. Pro Wittgensteina je podle všeho analogicky filosofická teorie pohádkou ve smyslu racionalizujícího vyprávění, které reaguje na neklid způsobený neporozuměním zákonitostem běžného jazyka. V tomto případě ovšem nemusí za problémem stát vývoj jazyka, ale jen vzdálenost běžné jazykové praxi. V té ovšem mohou přetrvávat staré obraty.[5]

Především se ovšem nabízí otázka, zda a jak je možné být podle Wittgensteinem nastíněné představy filosofickým problémem vůbec dotčen. Úvodní slova předmluvy nám připomenou předmluvu biskupa Berkeleyho k jeho nejslavnější knize *Pojednání o principech lidského poznání*, kde filosofy přirovnává k těm, kteří zvířili prach, načež si stěžují, že nevidí, zatímco lidé

sammenhang, über den Tod, die Seele, das Jenseits, Gott usw., indem man unverstandene Reste des früheren Glaubens rationalistisch deutete; durch Wandern der Stoffe zu anderen Völkern, durch Weitererzählen bei veränderten Zuständen des Volkes und mit Anpassen an das Neue. Der Vorgang ist im wesentlichen immer der: ein durch die Wirklichkeitserfahrung unlösbares Problem wird durch eine erfundene rationalisierende Geschichte gelöst. Im Fortgang der Zeiten stellen sich in dieser Geschichte wieder unlösbare Probleme heraus, und eine neue Erdichtung kommt wieder wirklichkeitsnäher; in der folgenden Zeit wird die Wirklichkeitskritik wieder schärfer, und eine neue Rationalisierung kommt: bis man zuletzt das Ganze als belanglos oder töricht überhaupt fallen läßt."

[5] Grimmovské pohádky se vrátí ještě v Wittgenstein (2001b, 4.014), kde se základní sémantický vztah zobrazování ilustruje přirovnáním vztahu gramofonové desky, hudební myšlenky a notového zápisu ke vztahu dvou mládenců, dvou lilií a dvou koní v pohádce *Zlaté děti* (Grimm – Grimm 1988, 137-139).

kráčející širokou cestou zdravého rozumu takové starosti nemají (viz Berkeley 2007, 89). Filosofie se v obou případech jeví jako choroba či zvůle, které se slušný člověk pokud možno obloukem vyhne. U Wittgensteina stejně jako u Berkeleyho ovšem není úplně jasné, že skutečně filosofií takto jednoznačně pohrdají. Přesněji, ani jeden by zřejmě nechtěl hájit představu, že propadnout filosofii je samo o sobě špatné a být filosofem je méněcenné. Kdyby pokládali každý pokus vyrovnat se s filosofickými problémy za chorobný nebo zlotřilý, nemohli by vystupovat jako autoři filosofické knihy, která na tuto patologii poukazuje.

Navíc, i kdyby filosofické problémy neměly být formulovány, neznamená to, že se nemají řešit, když už jednou formulovány byly. Nicméně, stojí-li formulace těchto problémů na nepochopení fungování jazyka, pak přímé řešení těchto problémů zřejmě nemůže vést ke kýženému výsledku. Účinně řešit filosofický problém pak nemůže znamenat odpovídat na filosofickou otázku, nýbrž odhalit zmatek, na němž stojí sama formulace problému. To je také podle všeho smyslem *Traktátu*, který tento úkol neplní jen pro jednotlivé filosofické problémy, nýbrž poukazuje na kořen filosofických problémů jako takových. Nacházíme-li tedy v knize teze, jež vypadají jako pokusy přímo odpovědět na otázky z jednotlivých filosofických disciplín, pak stojíme před úkolem vtělit je do naznačeného celkového záměru knihy.[6]

Filosofie tady vystupuje ve dvou formách. Lze ji provozovat jako přímé odpovídání na filosofické otázky, nebo jako eliminaci potřeby tyto otázky klást a to tak, že budeme poukazovat na neporozumění jazykové logice, na němž formulace těchto otázek stojí. To první se na první pohled zdá být z pozice předmluvy zavrženíhodné. Víme ovšem, že Wittgenstein sám byl jinak v některých případech pozoruhodně otevřený takříkajíc estetickému ocenění i těch filosofických děl, s jejichž obsahem včetně celkového záměru třeba zásadně nesouhlasil. A v kontextu zpětně uznaného vlivu Paula Ernsta se nabízí otázka, proč by tradiční filosofické nauky měly mít menší právo na existenci než pohádky, které prý také vznikají z nepochopení jazykové logiky. Přinejmenším dokud nepochopíme zmíněný původ filosofických otázek, jsou nauky filosofů stejnými racionalizacemi a jako takové jsou oprávněné.

[6] Filosofické teze, jež v knize nalezneme, samozřejmě nemusejí být nevyhnutelně nesmyslné. Přinejmenším objasnění povahy filosofických soudů by mělo patřit k myšlenkám, jež *Traktát* podle své předmluvy obsahuje a jež jsou podle ní pravdivé. Samozřejmě zůstává sporné, jak tuto tezi předmluvy sladit s teorií významu, jež se v knize dále předkládá a jež ústí v návrh, říkat pouze věty přírodovědy (Srov. Wittgenstein 2001b, 6.53).

Když jednou nahlédneme, odkud se filosofické otázky berou, začneme se ohlížet po přiměřenějším způsobu, jak své odcizení jazyku řešit.

Druhá zmíněná forma filosofování je pro Wittgensteina očividně chvályhodná, a odpovídá také představě, že smysl knihy je etický.[7] Smysl knihy by totiž byl v takovém případě navýsost praktický: *Traktát* by nás vedl k nahlédnutí, že a jak filosofické problémy, jež náš trápí, pramení v tom, že nejsme doma ve svém jazyce. Napravit takové neporozumění by pak znamenalo osvobodit se od kletby filosofické otázky. Stále ovšem zbývá otázka, jak je to s hodnotou samotné posedlosti filosofickými problémy.

Zdálo by se, že na tuto otázku nemůžeme po přečtení prvních dvou odstavců předmluvy odpovědět. Podívejme se ovšem ještě na dva přehlížené detaily.

Podle začátku předmluvy se kniha a) zabývá filosofickými problémy, b) ukazuje, že *formulace* těchto problémů stojí na nedorozumění. Mít problém a formulovat jej jsou jistě různé věci. Už víme, že snaha vyrovnat se s filosofickým problémem, ba ani s formulovaným filosofickým problémem není pro autora *Traktátu* sama nevyhnutelně špatná. Chybu bychom ovšem udělali, kdybychom vzali vážně formulaci filosofického problému jako otázku, na kterou je třeba najít přímou odpověď. V tomto ohledu jsou filosofické otázky opravdu jako hravé hádanky, jejichž řešením také typicky není přímá odpověď na doslovně chápanou otázku (srov. Diamond – White 1997c, 50).

Zkusme vést obdobné rozlišení, jaké jsme vedli na straně řešení, také straně filosofického problému samotného. Jako nemusí být špatné zabývat se filosofickými problémy či jejich formulacemi, tak třeba nemusí být špatné trpět filosofickými problémy, mít filosofické problémy, dělat si jistý druh starostí. Zmatek přece podle předmluvy stojí toliko za formulacemi těchto starostí do podoby otázek, jež čekají na odpověď.[8] Pak by se otevírala možnost, že filosofické starosti zkrátka patří k životu a patologické formy nabývají ve svém formulování, jež přirozeně vede k hledání odpovědí na tyto formulace, jež mají podobu otázek. V prvním plánu se nabízí možnost, že filosofické starosti i jejich řešení mají původně praktickou povahu. Pak by bylo možné, že filosofické starosti jsou cosi, s čím je zkrátka třeba žít, resp. jejich řešení nejde přes jejich formulování, a rozhodně jejich řešení nespočívá v odpovědi na otázky, do kterých tyto starosti formulujeme, ba snad tyto

[7] Wittgenstein Fickerovi 20. října 1919, viz Wittgenstein (1997b, 130).

[8] Srov. Wittgenstein (1997c, 50): „Die Entshethung der Probleme: die drückende Spannung, die sich einmal in eine Frage zusammenballt und sich objektiviert."

odpovědi k řešení starostí samy přímo podstatně nepřispívají, nebo přinejmenším samy nejsou hledaným řešením. Přinejmenším je jasné, že Wittgensteinův text můžeme legitimně zkusit číst také jinak než jako výraz extrémního filosofického kvietismu. Také by pak dávala dobrý smysl Wittgensteinova slova z dopisu nakladateli Fickerovi o etickém smyslu knihy.

A druhou indicii můžeme v prvních dvou odstavcích najít hned na prvním řádku. Je zde řeč o myšlenkách, jež jsou v knize vyjádřeny. *Traktát* má tedy mít nějaký obsah, což kupodivu vůbec není triviální zjištění. Často se má totiž za to, že celá kniha vrcholí paradoxem, kdy autor poukazuje na skutečnost, že věty knihy nesplňují kritéria smysluplnosti těmito větami formulovaná, z čehož se vyvozuje silný závěr ohledně smysluplnosti jejího obsahu (srov. poznámku 1). Cíl knihy se v takových interpretacích redukuje na terapeutickou proměnu čtenářova postoje k filosofii, zatímco obsah se knize vlastně úplně upře. Pak by ale v knize přísně vzato nemohly být vyjádřeny žádné myšlenky. Přečtení dalšího odstavce by nás mělo této starosti zbavit.

> Kniha tedy chce vytyčit hranice myšlení – nebo spíše nikoli myšlení, nýbrž výrazu myšlenek: Vždyť aby se mohly vytyčit hranice myšlení, museli bychom být schopni myslet obě strany této hranice (museli bychom tedy být schopni myslet, co se myslet nedá).
> Hranice se tedy bude dát vést jedině v jazyce a co leží za touto hranicí, bude zkrátka nesmysl. (Wittgenstein 2001b, 2)

Kniha má vytyčit myšlení hranice. A hodlá to udělat vymezením legitimních jazykových výrazů myšlenek, protože v jazyce lze formulovat také výrazy, které nejsou výrazy žádných myšlenek (a zároveň tam podle všeho lze formulovat výrazy všech legitimních myšlenek), zatímco myšlenky nelze rozdělit na legitimní a nelegitimní, nýbrž lze nanejvýš oddělit myšlenky od toho, co myšlenky nejsou. To, co leží za hranicemi myšlení, je přece zkrátka nesmysl a nedá se myslet, resp. není to myšlenkou. Samotné vedení hranic, resp. specifikace podmínek smysluplnosti pro výrazy a případné třídění výrazů podle stanovených kritérií ale musí samo také být myšlením. A formulace tohoto myšlení v jazyce musí být legitimní. Takovou formulaci musí mít Wittgenstein na mysli, když hned na začátku předmluvy zmíní „myšlenky, jež jsou v ní vyjádřeny“ (totiž v předkládané knize), a na konci předmluvy se k nim vrátí s tím, že jsou pravdivé s definitivní platností a jejich formulování označí za jeden z hlavních přínosů knihy.

Pomoci může jednoduché rozlišení mezi výrazem (*der Ausdruck der Gedanken*), myšlenkou (*die Gedanke*), která je tímto výrazem vyjádřena, a sa-

motným aktem myšlení (*das Denken*) této myšlenky (potažmo myslícím subjektem).[9] Asi každý čtenář *Traktátu* dřív nebo později narazí na otázku možnosti něčeho jako myšlenky, které nelze vyjádřit. Předmluva naproti tomu podle všeho počítá s představou, že pro každou myšlenku existuje jazykový výraz, ne však naopak. Na této představě stojí možnost vytyčit hranice myslitelného prostřednictvím vytyčení hranic vyjádřitelného a to uvnitř jazyka. Taková představa ovšem stále nevylučuje faktickou možnost, aby subjekt myšlení zaujímal postoje k nemyslitelnému, resp. k tomu, co bychom byli v pokušení nazvat nevyjádřitelnými myšlenkami, resp. nesmyslnými myšlenkami, jež se také dají formulovat v jazyce. Podle citátu z předmluvy totiž sice prý nejsme schopni myslet něco, co se myslet nedá (nelze myslet nemyšlenku), ale jak uvidíme níže na příkladech z korespondence s Russellem, tuto formulaci bychom rozhodně neměli obráceně chápat jako tezi, že každá věta určitého běžného jazyka (natož věta, kterou někdo hlásá jako výraz nějaké myšlenky a vážně se k ní hlásí) je podle Wittgensteina opravdu větou (která vyjadřuje myšlenku) – a to navzdory tvrzení, že věty našeho běžného jazyka jsou logicky v pořádku tak, jak jsou (viz Wittgenstein 2001b, 5.5563). Tato věta nám jen připomíná původ filosofických problémů v odcizení běžnému jazyku. Ostatně, vykolíkovat myšlení hranice právě v jazyce se Wittgenstein podle předmluvy rozhoduje právě s odůvodněním, že v jazyce lze formulovat výrazy pro nesmysly, jež leží za jím vytyčenou hranicí, takže v jazyce můžeme vidět obě strany hranice, zatímco v myšlení samém nic takového provést nejde. Ostatně, Wittgenstein v citovaném úryvku upřesňuje, že vlastně nehodlá omezovat myšlení, nýbrž jen vyjadřování. Nesmysly se dají leda říkat, myslet nikoli.

Čím se tedy *Traktát* zabývá, když se podle předmluvy zabývá filosofickými problémy, aby ukázal, že formulace těchto problémů stojí na neporozumění jazykové logice? Co je obsahem oněch myšlenek v knize vyjádřených? Mohli bychom zkusit říci, že se *Traktát* má zabývat jakýmsi kvazimyšlením, totiž zaujímáním nějakého typu postojů k něčemu, co za myšlenky pouze zaměňujeme, když se snažíme zachytit onen druh neklidu, který chápeme jako filosofické znepokojení, a vyrovnat se s ním (typicky pomocí filosofické nauky). Vyšla by nám interpretace „pravdivých myšlenek, které kniha obsahuje" jako myšlenek, které se týkají zaujímání postojů k něčemu, co nejsou myšlenky (a co se podle předmluvy nedá myslet). Ty

[9] Srov. Wittgenstein (2001b, 3.11), kde je rozlišena věta jako znak od větného smyslu. Znak je projekcí situace a myšlení větného smyslu je metodou projekce.

by se ve zmíněných pravdivých myšlenkách *Traktátu* vyskytovaly v tzv. nepřímých kontextech. Takové zaumnosti se ovšem lze snadno vyhnout, když zůstaneme u výkladu „pravdivých myšlenek knihy" v duchu výše uvedeného cíle: formulovat teorii významu, která stanoví kritéria smysluplnosti, jež umožní rozeznat smysluplné výrazy od nesmyslů. To, že a jak formulace filosofických problémů stojí na nepochopení toho, jak funguje jazyk, se kniha podle předmluvy nechystá říkat – kniha to „ukazuje". Rozlišení říkat versus ukazovat přitom podle všeho patří k jádru *traktátovské* sémantiky. Předmluva nám prezentuje *Traktát* jako knihu, kterou bychom měli v souladu s tzv. zobrazovací teorií významu, jež by podle běžného čtení měla být jejím obsahem, chápat jako větu, resp. jako funkci jednotlivých vět, které stejně jako kniha vcelku jsou pravdivé a mají smysl, který ukazují. Napětí se závěrem knihy, kde se říká, „[m]oje věty objasňují tím, že ten, kdo mi rozumí, nakonec rozpozná, že jsou nesmyslné, když jimi – po nich – vystoupí nad ně (Musí takříkajíc odhodit žebřík, poté, co po něm vylezl)" (Wittgenstein 2001b, 6.54), nelze přehlédnout. Explicitně by se tady v knize mělo provozovat teoretizování o jazykovém významu ve službě metafilosofickému projektu. To by mělo tvořit obsah knihy a tady (nikoli v jednotlivých filosofických disciplínách jako ontologie či epistemologie) bychom měli hledat ony pravdivé myšlenky, jež jsou v ní vyjádřeny.

> Jak dalece se mé úsilí kryje s úsilím jiných filosofů, nechci posuzovat. To, co jsem zde napsal, si v jednotlivostech ani nečiní žádný nárok na novost; a proto neuvádím žádné zdroje, protože je mi lhostejné, zda to, co jsem myslel, myslel už přede mnou někdo jiný.
> Zmíním jen to, že za podnět ke svým myšlenkám vděčím z velké části velkolepým dílům Fregovým a pracím svého přítele pana Bertranda Russella. (Wittgenstein 2001b, 2)

Především je tady znovu řeč o myšlenkách, které kniha obsahuje. Jinak Wittgensteinovo vyjádření k originalitě *Traktátu* působí velice pyšně. Jako inspiraci ovšem zmiňuje Fregova „velkolepá díla" a „práce svého přítele" Russella. Jak jsme už viděli, později bude litovat, že zapomněl na autora doslovu ke grimmovským pohádkám Paula Ernsta.

Vedle již řečných se později v knize explicitně zmiňuje ještě Heinrich Hertz stejně, jako tam lze nalézt celkem otevřené aluze na Schopenhauera. Russell s Fregem (a podle Wittgensteina zřejmě také Ernst) se vážně zabývali specifikací podmínek smysluplnosti jazykových výrazů jako prostředkem k tomu, aby řekli něco o racionalitě. Pochopit fungování jazyka podle nich

mělo pomoci pochopit povahu myšlení. V tomto ohledu se k nim Wittgenstein hlásí. Jeho rezignace na originalitu vlastně jen opakuje první odstavec předmluvy, kde říkal, že kniha není učebnicí (nejde jí o to sdělit nové informace) a že jí bude možná rozumět jen ten, kdo na myšlenky v ní vyjádřené už sám někdy připadl.

> Má-li tato práce hodnotu, pak stojí na dvojím. Zaprvé na tom, že jsou v ní vyjádřeny myšlenky, a tato hodnota je tím větší, čím lépe jsou tyto myšlenky vyjádřeny. Čím lépe se trefil hřebík na hlavičku. – Zde si jsem vědom, že mé síly ke zvládnutí takového úkolu jsou skrovné. – Snad přijdou jiní a provedou to lépe.
> Naproti tomu *pravdivost* zde sdělených myšlenek se mi zdá nenapadnutelná a definitivní. Jsem tedy toho mínění, že v tom podstatném jsem problémy vyřešil s konečnou platností. A pokud se v tom nemýlím, pak stojí hodnota této kniha zadruhé na tom, že ukazuje, jak málo se dosáhlo tím, že se vyřešily. (Wittgenstein 2001b, 2-3)

Závěr předmluvy shrnuje, v čem spočívá hodnota práce (má-li nějakou). 1) V knize jsou vyjádřeny myšlenky. 2) Kniha ukazuje, jak málo se dosáhne vyřešením problémů, jež jsou v ní vyřešeny. První bod je rozveden ve dvou ohledech: a) zmíněné myšlenky se nepovedlo vyjádřit nejlépe, takže hodnota knihy by mohla být mnohem vyšší; b) pravdivost sdělených myšlenek je naproti tomu nepochybná a Wittgenstein deklaruje přesvědčení, že probírané problémy vyřešil s definitivní platností.

Potvrzuje se zde znovu, že kniha skutečně obsahuje podle svého autora myšlenky a to myšlenky pravdivé. Vidíme, že v předmluvě byla postupně řeč nejen o hodnotě knihy, ale též o jejím *smyslu*, opakovaně o *myšlenkách*, jež sděluje, a také o jejich *pravdivosti*. Tyto výrazy jsou klíčovou součástí obsahu *Traktátu* bez ohledu na rozdíly mezi jeho interpretacemi. Při samozřejmosti jejich používání v předmluvě je těžko můžeme odbýt jako netechnické případy, na něž nelze vážně vztahovat měřítka *traktátovské* terminologie (pokud něco takového existuje).[10]

Text předmluvy také není ke knize připojen narychlo před vydáním. Předmluvu najdeme skoro slova od slova, včetně všech zmíněných příkladů už v tzv. *ProtoTraktátu*, pracovní verzi knihy datované do 1916.[11] V ní také

[10] Korespondence s anglickým překladatelem opět ukazuje, že terminologie není to, na čem Wittgenstein lpěl. Nejlépe je to vidět na jeho opomíjení rozlišení mezi tím, co je *sinnlos* a *unsinnig*, jež je zásadní pro většinu interpretací. Srov. Kienzler (2009, 223-247).

[11] Detailní rozbor vzniku knihy z dochovaných materiálů viz Potter (2013, 13-39).

najdeme několik drobných textových variant, které ukazují, že Wittgenstein skutečně textu předmluvy věnoval pozornost. Víme také, že od dokončení knihy v roce 1918 až do jejího vydání o tři roky později Wittgenstein na knize nic neměnil. Podle všeho s ní byl spokojen a rozhodně neměl nutkavou potřebu neustále text přepracovávat, jako to dělával se svými rukopisy později. Je tedy snad namístě číst předmluvu jako plnohodnotnou součást knihy včetně interpretace klíčových termínů. Navíc předmluva ukázkově plní typickou roli: stručně a jasně říká, oč v knize jde.[12] A tento cíl spočívá v demarkaci myslitelného prostřednictvím toho, co lze vyjádřit. Je-li tedy předtím i potom v předmluvě řeč o myšlenkách, jež kniha vyjadřuje, nemůžeme výraz „myšlenky" chápat v nějakém odlišném smyslu nebo jako metaforu či odbýt jeho použití jako pouhou nedbalost. Předmluva pregnantně formuluje cíl spisu a říká, že spis obsahuje myšlenky, jež jsou pravdivé (i když se je prý nepovedlo vyjádřit zrovna nejlépe). Navíc kniha jako celek má podle předmluvy smysl, a nejspíš dokonce hodnotu.[13]

Když Wittgenstein za jeden z hlavních kladů knihy označuje to, že vyjadřuje myšlenky, říká tím zajisté, že kniha má nějaký smysluplný obsah (pravdivé myšlenky, jež řeší filosofické problémy). Toto konstatování ovšem stále působí neuvěřitelně banálně – dokud je nečteme jako tiché zdůraznění, že předkládaná kniha se v tomto ohledu liší od jiných filosofických knih. Těžko by Wittgenstein mohl ostřeji odmítnout představu, že jeho kniha je výhradně praktickou terapií.

Wittgenstein také v předmluvě postupně řekl, že a) knize bude rozumět možná jen ten, kdo už na myšlenky, které jsou v ní vyjádřeny, sám připadl; b) že si v jednotlivostech nečiní nárok na původnost; c) hodnota knihy spočívá ve vyjádření nepochybně pravdivých myšlenek (jakkoli nedokonalém). Podle předmluvy je Wittgenstein přesvědčen, že s definitivní platností vyřešil filosofické problémy jednoduše tím, že formuloval kritéria smysluplnosti jazykových výrazů a že z těchto kritérií vysvítá, že možnost formulovat filosofické problémy stojí na narušení těchto kritérií. Jak ale dodává, „hřebík na

---

[12] Srov. Fregovu výtku, že kniha postrádá skutečnou předmluvu, jež by řekla, oč v knize jde. Fregův dopis Wittgensteinovi z 30. září 1919 – viz Frege (2011, 44-47).

[13] Srov. ovšem oproti tomu Wittgensteinův dopis Ogdenovi z 8. listopadu 1921, kde k navrhovanému názvu knihy *Filosofická logika* poznamenává, že je to nesmysl a že neví, co filosofická logika je, takže takový název by byl vhodný leda, kdybychom chtěli říct, že když je kniha nesmysl, mohl by být nesmyslný už její název. Srov. Wittgenstein (1973, 4). Což lze ovšem odbýt jako sarkasmus, který si zahrává s interpretačními potížemi, jež zde probíráme.

hlavičku“ trefil „v tom podstatném“, ale jeho síly jsou skrovné, možná se nevyjádřil nejlépe (a třeba to někdo napraví). Jádrem knihy by mělo být předvést, jak formulace filosofických problémů stojí na vymknutí z jazykové logiky (což by měla být hlavní myšlenka knihy). Na tuto myšlenku ve slabší nebo silnější formě a s různými závěry skutečně před Wittgensteinem připadla řada jiných, kteří se ji také pokusili odpovídajícím způsobem demonstrovat. Wittgenstein jako by se zde hlásil k delší tradici s tím, že doufá starou myšlenku lépe vyjádřit a předvést její relevanci. Jednak Wittgenstein říká, že zásadně pochybené a zmatené nejsou jen teorie některých škol nebo filosofických disciplín, nýbrž formulace filosofických problémů vůbec a toto zmatení stojí na odcizení běžnému jazyku s jeho fungováním. A za druhé se podle všeho domnívá, že lépe formuloval kritéria smysluplnosti jazykových výrazů, a přišel tak na způsob, jak zřetelněji než jeho předchůdci rozeznat případy nesmyslů (srov. Wittgenstein 2001b, 3.325). Zde bychom mohli hledat zdroj radosti, kterou by kniha měla přinést tomu, kdo ji čte s porozuměním. Půjde o někoho, kdo sám v nějaké formě připadl na myšlenku, že formulace filosofických problémů se vzpírají jazyku, případně se s touto myšlenkou pokusil o nějakou teorii významu. U Wittgensteina by ke své radosti měl tuto myšlenku nalézt ostřeji formulovanou, případně vyargumentovanou a demonstrovanou pevnějšími kritérii smysluplnosti.

Poslední věta předmluvy nás vrací k otázce po hodnotě posedlosti filosofickým problémem, nakolik je nezávislá na explicitní otázce, na niž by se hledala odpověď. Wittgenstein totiž uzavírá předmluvu poukazem na druhý zdroj hodnoty své knihy: „ukazuje, jak málo se dosáhne tím, když se vyřeší“ (totiž filosofické problémy). Kniha nás má totiž vlastně jen zbavit nadějí, jež bychom spojovali s odpovědí na otázky. Takže to, co nás nutí si je klást, nezmizí. Pochopíme, že nauky filosofů jsou jen racionalizace, jež mají čelit zmatku z odcizení jazyku. Odcizení samo ale tento objev přetrvá.[14] Říci, že jeho řešení má praktickou, nikoli teoretickou povahu, ovšem možná znamená jít nad rámec rané Wittgensteinovy filosofie, která jakoby se potýkala právě s tím, že v ní schází pojem praktické racionality.

[14] Mimochodem, je jistě zajímavé postavit vedle tohoto pesimisticky znějícího závěru motto druhé Wittgensteinovy knihy, jež se obvykle chápe jako zásadní obrat a odhalení chyb *Traktátu*: „Pokrok má tu vlastnost, že vždycky vypadá větší, než ve skutečnosti je.“

2. Motto

Ale zpátky k autorským vysvětlením pointy knihy. Z předmluvy je jasnější též motto knihy: „... a vše, co člověk ví, co není jen zaslechnutý ruch a šum, dá se říci třemi slovy." Odpovídá tomu, jak Wittgenstein shrnuje to, co sám nazývá smyslem knihy: „Co se vůbec dá říct, dá se říct jasně; a o čem nelze mluvit, k tomu se musí mlčet." Při této příležitosti je namístě zmínit, že v kontextu, z něhož motto pochází, neznamená ono „jasně" totéž jako „precizně", „zřetelně" či „rozlišeně":

> Zeptám-li se polovzdělance, jaký je rozdíl mezi antickým a moderním, mezi klasickým a romantickým uměním, velmi rozčileně mi odpoví: Pane, ta otázka vyvolává ohromnou spoustu představ. Jde o námět na celé knihy a zimní semestr. Zeptám-li se naopak někoho řádně a dokonale vzdělaného, dostane se mi odpovědi: Pane, to se dá říct třemi slovy. Staré umění se týkalo těla, v tom novém jde o duši. Proto bylo staré umění plastické, a to nové je lyrické, muzikální, zkrátka romantické. Bravo! Máme zde jako na dlani ohromnou spoustu představ, když jsou člověku opravdu vlastní – a vše, co člověk ví a co není jen zaslechnutý ruch a šum, dá se říci třemi slovy.[15] (Kürnberger 1877, 338-351)

Polovzdělanec by se odvolal na potřebu důkladného rozboru, zatímco znalec je s to stručně odpovědět. Mluvit jasně zde znamená neutopit se v detailech a mluvit bez odkazu na autority. Obojí lze v případě potřeby doplnit.

Určitě bychom neměli motto interpretačně přetěžovat tím, že bychom na něm chtěli vystavět originální interpretaci Wittgensteinovy filosofie. Neměli bychom je nicméně přehlížet. Wittgenstein míval u svých větších textů obvykle nějaké motto a jeho výběru dokázal věnovat překvapivou péči.

---

[15] „Wenn ich einen Halbgebildeten frage: Was ist der Unterschied zwischen der antiken und der modernen, zwischen der klassischen und der romantischen Kunst? so wird er in großer Verwirrung antworten: Herr, diese Frage regt ganze Welten von Vorstellungen auf. Das ist ein Stoff für ganze Bücher und Wintersemester. Wenn ich dagegen einen Durchgebildeten und Ganzgebildeten frage, so werde ich die Antwort erhalten: Herr, das ist mit drei Worten zu sagen. Die Kunst der Alten ging vom Körper aus, die Kunst der Neuern geht von der Seele aus. Die Kunst der Alten war deshalb plastisch, die Kunst der neuern ist lyrisch, musikalisch, kurz romantisch. Bravo! So haben ganze Welten von Vorstellungen, wenn man sie wirklich beherrscht, in einer Nuß Platz, und alles, was man weiß, nicht bloß rauschen und brausen gehört hat, läßt sich in drei Worten sagen." (Kürnberger 1877, 338-351).

Především, v kontextu odstavce je lépe vidět, že opozice mezi tím, co víme, a tím, co je jen zaslechnutý ruch a šum, proti sobě nestaví bezprostřední vnímání jednotliviny a logický nebo gramatický nesmysl. Na jedné straně je jasná obecná formulace založená na přímé znalosti tématu. Na druhé pak vyhýbavé mlžení, jež vychází ze zprostředkované znalosti, resp. odkazu na cizí znalost. Zaslechnutý ruch a šum tedy není diskvalifikován jako bezcenný obsah, nýbrž jako obsah, který není osvojený.

Zadruhé, z kontextu je zřejmé, že v Kürnbergerově textu nejde o skepsi či meze lidského poznání. Pointa nespočívá v poukazu na to, že ve skutečnosti víme tak málo, že nám k vyjádření našeho poznání stačí pár slov. Pointou je spíše rozlišení mezi skutečným věděním a intelektuálním snobismem. O rozsah vědomostí zde primárně nejde, může být u znalce i snoba stejný, rozhodně není u znalce menší. Primárně jde o postoj k vědomostem. Skutečný znalec mluví zpříma a jasně.

Zatřetí, rozdíl mezi klasickým a moderním uměním je skutečně téma na knihy a univerzitní kursy. Napadne nás, zda Kürnberger nemluví ironicky, což by jistě mohlo mít vliv také na Wittgensteinovo použití jeho slov. Citát pochází z textu *Opoziční stavba pomníků* a slouží v něm především postesku nad obecnou zahleděností do minulosti, a následné výzvě k samostatnému jasnému myšlení, jakému by se mohly stavět nové pomníky, resp. myšlení, jež se samo takovým pomníkem stává. Tento kontext dobře souzní s Wittgensteinovým prohlášením nezávislosti v předmluvě, kde deklaruje nezájem o případné shody se staršími filosofy a omezuje se na stručný dík za inspiraci Fregovi s Russellem. Přičítat Kürnbergerovi ironii zde tedy není na místě. Text ostatně nevybízí k nějakým unáhleným soudům ani nepopírá potřebu studia pro osvojení znalosti, jež umožní vynášet s přehledem obecné soudy. Pochybnost o vážnosti Kürnbergerova příkladu se zakládá na obracení implikace. Ten, kdo je znalý, zajisté dokáže vynášet obecné soudy. Tuto celkem intuitivní představu bychom neměli zaměňovat s tvrzením, že každý, kdo vynáší obecné soudy, je znalec. Vynášení obecných soudů tak není kritériem znalosti, ale neschopnost vynášet obecné soudy, může posloužit jako kritérium neznalosti.

Motto bychom mohli parafrázovat: „To, co opravdu víme a nemáme to jen z druhé ruky, to můžeme říct stručně a jasně." A z kontextu, v jakém se citát původně vyskytuje, se dá chápat především jako deklarace modernismu: výzva stát se vzorem samostatného myšlení poučeného na vlastní přímé obeznámenosti, jež neulpívá v minulosti a nedrží se autorit. Takové chápání jde velmi dobře dohromady s tím, co Wittgenstein říká v předmluvě, ale

není jisté, že si motto vybral právě pro takové sdělení. Jistě v mottu ovšem nacházíme kontrast přímé a nepřímé znalosti, z nichž druhá se hodnotí záporně. Samozřejmě se nabízí vztáhnout „zaslechnutý ruch a šum" k odcizení běžnému jazyku, které plodí filosofické teorie, jež jenom zkoušejí racionalizovat bezradnost, zatímco to, co se „dá říct třemi slovy" by bylo tím, co dokážeme říct z přímé konfrontace s realitou. S dalekosáhlejšími závěry ale musíme počkat, dokud neprozkoumáme další prameny. Těmi hlavními bude korespondence s inspirátory zmíněnými v předmluvě, totiž s Russellem a Fregem, a také dopis nakladateli Fickerovi. K těm se dostaneme příště.

Literatura

BAKER, G. – HACKER, P. (2005): *Understanding and Meaning. Analytical Commentary to Philosophical Investigations. Part 1: Essays*. Oxford: Blackwell.

BERKELEY, G. (2007): *Pojednání o základech lidského poznání*. Praha: Oikoymenh.

DIAMOND, C. – WHITE, R. (1977): Riddles and Anselm's Riddle. *Proceedings of the Aristotelian Society*, Suppl. Vol. 51, 143-186.

DOŠEK, T. (2013): Co je dosažitelné jen ze žebříku, mě nezajímá. *Organon F* 20, č. 2, 222-250.

ENGELMANN, P. (1967): *Letters from Ludwig Wittgenstein*. Oxford: Blackwell.

ERNST, P. (1940): *Völker und Zeiten im Spiegel ihrer Dichtung*. München: Albert Langen und Georg Müller.

FREGE, G. (2011): Frege-Wittgenstein Correspondence. In: De Pellegrin, E. (ed.): *Interactive Wittgenstein: Essays in Memory of Georg Henrik von Wright*. Wien: Springer.

GRIMM, J. – GRIMM, W. (1988): *Pohádky*. Praha: Svoboda.

KIENZLER, W. (2009): Die Sprache des Tractatus: klar oder deutlich? Karl Kraus, Wittgenstein und die Frage der Terminologie. In: Gebauer, G. – Goppelsröder, F. – Volbers, J. (eds.) (2009): *Wittgenstein, Philosophie als „Arbeit an Einem selbst"*. Paderborn: Fink, 223-247.

KÜNNE, W. (1996): Paul Ernst und Ludwig Wittgenstein. *Wittgenstein Studies* 3, No. 1, available at: http://sammelpunkt.philo.at:8080/482/1/18-1-96.txt.

KÜRNBERGER, F. (1877): Denkmalsetzen in der Opposition. In: Kürnberger, F.: *Literarische Herzensachen*. Wien: Rosner, 338-351.

POTTER, M. (2013): Wittgenstein's Pre-Tractatus Manuscripts: A New Appraisal. In: Sullivan, P. – Potter, M. (eds.) (2013): *Tractatus. History and Interpretation*. Oxford: Oxford University Press, 13-39.

RAMSEY, F. (1931): *Foundations of Mathematics and Other Logical Essays*. London: Routledge.

READ, R. – CRARY, A. (eds.) (2000): *The New Wittgenstein*. London: Routledge.

MCGUINESS, B. (ed.) (1979): *L. Wittgenstein und Wiener Kreis*. Oxford: Blackwell.

WITTGENSTEIN, L. (1973): *Letters at C. K. Ogden*. Oxford: Blackwell.

WITTGENSTEIN, L. (1997a): *Cambridge Letters*. Oxford: Blackwell.
WITTGENSTEIN, L. (1997b): *Tagebücher und Briefe*. Innsbruck: Haymon-Verlag.
WITTGENSTEIN, L. (1997c): Tagebücher 1914-16. In: Wittgenstein, L.: *Werkausgabe*. Bd. 1. Frankfurt: Suhrkamp.
WITTGENSTEIN, L. (2001a): Filosofie. *Organon F* 8, č. 2, 174-189.
WITTGENSTEIN, L. (2001b): *Logisch-philosophische Abhandlung*. Kritische Edition. Frankfurt: Suhrkamp.

Organon F 21 (3) 2014: 381-398

Implicitní pravidla

JAROSLAV PEREGRIN

Filosofický ústav Akademie věd České republiky, v.v.i.
Jilská 1. 110 00 Praha 1. Česká republika
Filosofická fakulta. Univerzita Hradec Králové
Náměstí Svobody 331. 500 03 Hradec Králové. Česká republika
jarda@peregrin.cz

ZASLÁN: 21-02-2014 • AKCEPTOVÁN: 12-04-2014

ABSTRACT: In his criticism of my book *Člověk a pravidla* [*Man and rules*], Michal Ivan scrutinized my notion of implicit rule, concluding that it is flawed. In this contribution, I defend my approach, explaining the notion in greater detail. I state that my talk about the existence of an *implicit rule* refers to the social setting in which some kinds of social (especially linguistic) actions are governed by *normative attitudes* of the members of the society. These normative attitudes institute the propriety which make instances of actions of the kinds either *correct* or *incorrect*; hence people can follow or violate the rule, the rule can come into being, develop, and fade away – without it being explicitly articulated.

KEYWORDS: Implicit rules – language – rules – rule following.

Jsem rád, že moje kniha *Člověk a pravidla* (viz Peregrin 2011a) vzbuzuje intenzivní kritické reakce – naposledy obsáhlý a promyšlený text Michala Ivana (viz Ivan 2014). Tu knihu jsem napsal mimo jiné také jako jakousi stručnou „inventuru" toho, co už o pravidlech vím (nebo se domnívám, že vím); s tím, že některým z těch věcí se chci v budoucnu dále věnovat a rozebírat je do větší hloubky,[1] a kritické připomínky mi v tomhle samozřejmě

[1] Nějaké plody onoho „většího rozebírání" už tu jsou – viz zejména Peregrin (2011b; 2012a; 2012b; 2012c; 2012d; 2014).

velmi pomáhají. Musím nicméně konstatovat, že i přes různá zjednodušení, na které mí kritici oprávněně poukazují, mi koncepce pravidel a normativity, kterou jsem v té knize předložil, stále připadá nosná.

Ivanova kritika navazuje na výhrady, které se objevily již v úvaze, ke které moje kniha inspirovala Ladislava Koreně. Koreň (2012) píše o tom, že moje úvahy se nebezpečně blíží bludnému kruhu, neboť podle nich „propozičné poznatky predpokladajú propozičný kód, ten však predpokladá jazyk, ktorý zas predpokladá pravidlá, ktoré sú sami propozičnými poznatkami – a tak ďalej dokola!" Koreň ovšem vidí i přímočarou cestu ven z tohoto kruhu: podle něj bych měl „pripustiť, že to, čo v niektorých úvahách nazýva *pravidlami*, nie sú v skutočnosti plnokrvné pravidlá (*qua* deontické obmedzenia), ale akési proto-pravidlá (...)" (Koreň 2012, 615).

Když jsem Koreňovu kritiku četl, neměl jsem pocit, že by mě nějak příliš zasahovala, zejména proto, že se mi zdálo, že ono řešení, po kterém Koreň volá, fakticky v knize najdeme. (Koreň to také neprezentoval jako nějaký fatální nedostatek, ale jako drobnější vadu na kráse jinak pozoruhodné teorie.) Hovořím tam přece o rudimentárních pravidlech (což není nic jiného než Koreňova „proto-pravidla"), i o tom, jak se od těchto rudimentárních pravidel propracováváme k pravidlům plnohodnotným. Poté, co mi podobnou věc mnohem vehementněji vyčetl i Ivan (který ji vidí jako fatální), jsem si ale uvědomil, že něco z toho, co v této knize říkám, skutečně může být problematické a zavádějící. Konkrétně se jedná o následující pasáž:

> A je to právě schopnost takového poznávání v podobě *vědění-že*, na čem staví pravidla. Pravidlo je vlastně také určitým modem tohoto druhu vědění – jde, můžeme říci, o *know-that-something-ought-to-be*, zkráceně *ought-to-be* čili *mělo-by-být*. Tím se dostáváme k terminologii Wilfrida Sellarse. Podle něj existují dva druhy „normativního vědění": vědění typu *mělo-by-se-dělat* (což je korelát příkazu, nebo alespoň nějakého bezprostřední vodítka pro jednání) a vědění typu *mělo-by-být* (přesvědčení o žádoucnosti určitého stavu). Každé takové *mělo-by-být* vede k různým *mělo-by-se-dělat* – tato *mělo-by-se-dělat* jsou tím, co je třeba činit, aby nastalo či přetrvalo ono *mělo-by-být*. (Peregrin 2011a, 65)

Moji kritici tuto pasáž zřejmě interpretují tak, že normativní postoje, které jsou podle mého názoru pravidlům konstitutivní, nutně musejí být *propozičními* postoji; a tudíž, že pravidly se může řídit jedině ten, kdo je schopen propozičního myšlení a vědění-že. To jsem rozhodně říci nechtěl (jakkoli musím uznat, že uvedená pasáž je v tomto ohledu velmi matoucí).

Určitě tomu tak není u rudimentárních pravidel, to jest u pravidel, skrze které se k propozičnímu myšlení (a k vědění-že) teprve propracováváme. A není tomu tak obecně ani u plnohodnotných pravidel ve společnosti, jako je ta naše – v tomto případě je sice *možné* pravidla vyjádřit jako (normativní) propozice, avšak není nutné, aby každý, kdo zaujímá normativní postoje, tak musel činit explicitně a propozičně. Uvedená pasáž se měla vztahovat na situaci, kdy se již k existenci propracovala plnohodnotná (a tedy již nikoli jenom rudimentární) pravidla a kdy jsou ti, kdo se pravidly řídí, již plnohodnotnými protagonisty propozičního myšlení.

Je ovšem pravda, že pokud to, co jsem v knize takto napsal, interpretujeme tak, že pravidla musejí být nutně podložena *propozičními* postoji – to jest tak, jak to interpretují moji kritici, rozumím-li jim dobře – jeví se jistě i některé další části mé knihy jako problematické. A já teď vidím, že na tom rozhodně nejsem bez viny. Takže připouštím, že Ivanova kritika může být v tomto smyslu opodstatněnější, než se bude jevit z mého dalšího textu. Já nicméně budu vycházet z vysvětlení, které jsem nyní podal, tj. že pravidla – alespoň ve své rudimentární podobě – propozičnímu myšlení předcházejí, nepředpokládají ho.

A ještě jednu věc, na kterou Ivan upozorňuje, je třeba připsat rozhodně k mé tíži; a to je chyba v překladu Wittgensteina v Peregrin (2011a, 26). Ta je hloupá a omlouvám se za ni.

Pokusím se nyní shrnout, jak to podle mě je s pravidly a v jakém smyslu hovořím o explicitních a implicitních pravidlech.

Zásadní podle mne je, že existuje rozdíl mezi chováním, které je pouze pravidelné (je v souladu s nějakým pravidlem, je tímto pravidlem charakterizovatelné[2]) a chováním, které se tímto pravidlem skutečně řídí (které by tu nebylo, kdyby tu nebylo to pravidlo). Je tu zřejmý rozdíl mezi tím, co dělá zajíc, který vběhne na silnici a náhodou se tam pohybuje tak, že to je v souladu s dopravními předpisy, a tím, co dělá řidič, který se těmito předpisy skutečně řídí. V čem tento rozdíl spočívá?

Na první pohled se to zdá být jasné: ten rozdíl je v *mysli* toho, o kterého jde, zatímco ten zajíc ve své mysli nemá nic jako dopravní předpisy, onen řidič je tam má a řídí se jimi *vědomě*. Já mám pocit, že jakkoli samozřejmě

[2] Hovořit o tom, že je chování někoho nebo něčeho „charakterizovatelné pravidlem", může být ovšem matoucí. Rozumí se tím to, že toto chování je na pohled s oním pravidlem v souladu; a ono pravidlo se tak díky tomu může zdát být užitečné i pro předpovídání dalšího chování příslušné entity.

toto vysvětlení vypadá, není ve svých důsledcích udržitelné. Dovedu si představit třeba negramotného řidiče, který nikdy dopravní předpisy neviděl, ale cestou pokusu a omylu (či napodobování) se naučil se na silnici správně chovat, a i o takovém je podle mne třeba říkat, že se řídí dopravními předpisy – na rozdíl od výše uvedeného zajíce není jeho chování v souladu s předpisy jenom náhodou, ale v důsledku toho, že rozlišuje (jakkoli jaksi pouze „implicitně") mezi tím, co je na silnici správné, a tím, co ne. Je samozřejmě možné namítat, že ten negramotný řidič sice nemá v mysli doslova předpisy, ale má tam *něco*, čím se odlišuje od zajíce či jiného člověka, který se chová jenom pravidelně. Já však nevěřím, že by se tohle *něco* dalo nějak uchopit (a je, myslím zásluhou Kripkovy knihy o Wittgensteinovi – Kripke 1982 – že některé slepé uličky, do kterých pokusy o takové uchopení vedou, zmapoval); a, jak praví klasik, „nic je stejně tak dobré jako nějaké něco, o němž nelze nic říct".

To navíc ani nemluvím o obecném problému, který se pojí s každým vysvětlením opírajícím se o to, co je v něčí mysli. Do myslí lidem samozřejmě nevidím, takže nevím, jestli to tam *opravdu* je; a taková vysvětlení proto zákonitě stojí více či méně na vodě. To mě vede k závěru, že k tomu, abych odlišil chování, které je charakterizovatelné jako řízení se pravidlem, od chování, které je pouze pravidelné, musím svůj zrak nikoli zanořit do mysli toho, o jehož chování jde, ale naopak ho pozvednout k širšímu společenskému kontextu, ve kterém se tato osoba pohybuje. Pravidla jsou tím pádem pro mě především sociální instituce. (Tím samozřejmě nechci popírat, že řízení se pravidlem má všelijaké psychologické aspekty a že může být nadmíru zajímavé se těmito psychologickými aspekty zabývat. Tvrdím jenom, že o tyto *psychologické* aspekty nelze opřít *filosofický* výklad toho, v čem řízení se pravidlem spočívá.)

Abychom si osvětlili, jaká je (podle mne) povaha společenských institucí, představme si takovou věc, jako je zkouška na vysoké škole. (Kolega Ivan u mě právě nedávno jednu takovou úspěšně složil.) Když někoho zkouším, spočívá to obvykle fakticky v tom, že za mnou přijde, já se ho na něco ptám a on mi odpovídá. Občas za mnou ale někdo přijde a odpovídá na mé otázky, aniž by to byla zkouška. V čem je ten rozdíl? V tom, že v případě zkoušky zkoušenému sdělím známku a někam mu ji zapíšu? Asi nikoli, představíme-li si pomateného profesora (stav, ke kterému se já sám doufám zatím jenom blížím), který má pocit, že má někoho zkoušet a zapíše mu do indexu známku, přestože už je dávno v důchodu a na vysoké škole už neučí, tak vidíme, že tyhle věci z rozhovoru skutečnou zkoušku neudělají. Je tedy rozdíl

dán stavem mysli zkoušejícího a zkoušeného, v tom, že to oba „berou" jako zkoušku? S odkazem na různé pomatence opět vidíme, že tohle nemůže být zásadní. (Představuji si, že právě když jsem kolegu Ivana zkoušel, někde v psychiatrické léčebně třeba nějaký „Peregrin" „zkoušel" nějakého „Ivana", a oba byli přesvědčeni, že jde o zkoušku na Karlově univerzitě.)

Co podle mne dělá z rozhovoru dvou lidí skutečnou zkoušku, je tedy především širší společenský kontext, existence příslušné školy, jejích regulí atd., a to všechno je zase průsečíkem fantasticky komplexních postojů velké spousty lidí. Někdo je studentem, rektorem univerzity či zkoušejícím především proto, že spousta lidí uznává pravidla, která konstituují tyto role a jejich současné obsazení právě těmito osobami. (Vzpomínám si na rok 1989 a na osvobozující šok, kterým pro mě bylo plné uvědomění si faktu, že komunisté jsou tak mocní už skoro jenom proto, že my všichni je za tak mocné máme, a že když my všichni s tím, koordinovaně, přestaneme, jejich moc se zhroutí jako domeček z karet ...).

A to, co jsem právě naznačil na příkladu institucí, jako je univerzita a zkouška na takové univerzitě, podle mne platí obecněji o jakémkoli pravidlu: pravidlo vzniká v průsečíku určitých postojů mnoha lidí, konkrétněji ono *je* tímto průsečíkem. (Tím odpovídám na Ivanovu otázku, jakého druhu věc to je – je to věc stejného druhu, jako zkouška; můžeme tomu říkat *institucionální realita*.) Postoje, o které tady jde, nazývám *normativní* a mám za to, že to jsou dále již ne dost dobře analyzovatelné lidské postoje, identifikovatelné v podstatě behaviorálně; když jsou verbalizovány, nabývají podoby vět o tom, že je něco nějak správné nebo že něco nějak má být.

Pro pochopení toho, co to pravidla jsou, mi připadá zásadní přemýšlet o tom, jak se mohla vůbec nějaká pravidla objevit. Je totiž zcela zřejmé, že do současného stavu, ve kterém je naše společnost propletena velkou spoustou pravidel nejrůznějšího typu, jsme se museli nějak dostat ze stavu, kdy jsme, podobně jako spousta našich zvířecích bratranců – žádná pravidla neměli. A protože nepředpokládám, že schopnost řídit se pravidly mohla vzniknout jako nějaká jednorázová genetická mutace, domnívám se, že pravidla musela nejprve existovat jenom v nějaké jednoduché, rudimentární (zárodečné) podobě a z té se postupně vyvíjet k oněm komplexním podobám, jaké mají pravidla, která známe z našeho současného světa. (Ivan píše, že netematizuji „charakter implicitných pravidiel a ich vzťah s rudimentárnymi pravidlami". Měl jsem pocit, že tento vztah je z mé knihy jasný: rudimentární pravidla nemohou než být implicitní. Píše-li ale Ivan (2014, 86), „[p]okiaľ ide o prepojenie medzi rudimentárnymi a implicitnými pravidla-

mi, zdá sa, že je totožné s prepojením medzi akýmikoľvek explicitnými pravidlami a tými implicitnými“, vzbuzuje to ve mně vážnou obavu, že slovu „rudimentární“ rozumí nějak zcela jinak, než je normální, totiž jako *zárodečný* či *primitivní*.)

Představuji si, že se rudimentární pravidla mohla objevit například nějak takto (netřeba snad dodávat, že se jedná o spekulaci, nic jiného tady k dispozici není). Představme si tlupu nějakých našich prehistorických předků, kteří se ještě příliš neliší od jiných zvířat žijících v tlupách, to jest třeba vlků nebo lidoopů. V tlupě existuje nějaká hierarchie (jak to známe z tlup zvířat) a také všelijaké zaběhnuté způsoby chování. Řekněme, že jeden z takových způsobů spočívá v tom, že žádný jedinec se k jinému nepřibližuje na příliš těsnou vzdálenost, řekněme menší než půl metru. Tento způsob chování existuje na úrovni dispozic či reflexů: když se ke mně někdo přiblíží, dám víceméně mimovolně najevo svou nevoli a tím ho odpudím; naopak když se k někomu takto přiblížím já, jsem od toho já odrazen jeho nepřátelskou reakcí.

Z čistě reflexivních reakcí na to, že se ke mně někdo příliš přiblíží, se podle mne v průběhu vývoje lidstva vyvinuly komplexnější (jakkoli ne nutně nějak plně uvědomělé) postoje toho druhu, že nejenom mimovolně vrčím na někoho, kdo je u mě příliš blízko, ale projevuji nevoli i nad tím, když se někdo přibližuje příliš blízko k někomu jinému – nejde tedy už jenom o čistě obranou reakci, ale o reakci, která je vztažena obecně k nějakému druhu chování, ať už se týká bezprostředně mě, nebo ne. To jsou podle mne zárodky normativních postojů a zárodky pravidel.

Zdůrazňuji, že aby člověk (nebo možná i zvíře) mohl zaujímat takové rudimentární normativní postoje, nemusí mít nic jako plnohodnotná přesvědčení, tj. nemusí myslet propozičně či pohybovat se v prostoru důvodů. (Naopak, prostor důvodů *předpokládá* pravidla, takže takováto rudimentární pravidla stojí v jeho základě, namísto toho, aby byla jeho produktem.) Myslím, že z hlediska psychologie odpovídají něčemu na té úrovni, pro jakou Gendlerová (2008a; 2008b) razila termín *alief* – něčemu jako zárodkům přesvědčení, která ale ještě nejsou plnohodnotnými přesvědčeními.

Tyto zárodky se pak mohou dále vyvíjet tím způsobem, že se odmítavé (případně souhlasné) reakce na některé druhy chování, které společností rezonují, institucionalizují tak, že nabývají podoby různých odměn a trestů. Potom, když je k dispozici dostatečně vyvinutý jazyk, mohou být vyjádřeny a nějaký praGuth-Jarkovský může na stěnu jeskyně napsat „Člověk se nemá k jinému přespříliš přibližovat!“, čímž se toto pravidlo stane explicitním.

Z fylogenetického hlediska tedy pravidla nejprve zcela jistě existují jako implicitní, a teprve mnohem později, když se ustanoví jazyk, mohou nabýt explicitní podoby. (Tohle se samozřejmě odvíjí od mého přesvědčení, že i jazyk se zakládá na pravidlech – pokud bychom toto přesvědčení odmítli, otevřel by se snad prostor pro představu, že se *nejprve* kompletně vyvinul jazyk, a teprve potom se objevila pravidla, která tím pádem už mohla být všechna explicitní. Já se však domnívám, že jediný způsob, jak skutečně vysvětlit, jak funguje jazyková komunikace, počítá s tím, že to je pravidly řízená činnost – viz Peregrin 2012b; 2012c; 2012d.) Implicitními tedy rozhodně byla ona rudimentární pravidla, která stála na počátku vývoje pravidel; a implicitními jsou některá pravidla i v době, kdy už máme i explicitní pravidla. (Ptá-li se tedy Ivan (2014, 84), „Môžu implicitné pravidlá existovať bez toho, aby existovali pravidlá explicitné? Alebo je nutné, aby explicitné pravidlá vytvorili systém pravidiel, v ktorom sa ďalej dokážeme pohybovať prostredníctvom implicitných pravidiel?“, pak odpovědí je, že implicitní pravidla podle mne nejenom mohou, ale musejí předcházet pravidlům explicitním.)

Jak se mi ve světle těchto mých názorů jeví Ivanova kritika? Zdá se mi, že se s ním snad shodnu na tom, že je třeba rozlišovat mezi tím, co já nazývám řízení se explicitními pravidly, a chováním, které takovým řízením se explicitními pravidly není. Ivan (2014, 87) si ovšem vypůjčuje Wittgensteinovu terminologii, a hovoří o procesech „obsahujících“[3] pravidlo: „musíme rozlišovať medzi tým, čo by sa dalo nazvať ‚procesom, ktorý je v *súlade s* pravidlom‘ a ‚procesom obsahujúcim pravidlo‘“ (Wittgenstein 2002, 41), pričom *obsiahnuté* ‚znamená, že vyjadrenie tohto pravidla tvorí súčasť týchto procesov‘ (Wittgenstein 2002, 41)“. Ivan navíc konstatuje, že „Wittgenstein nepripúšťa inú možnosť“. Jakkoli ztotožnění toho, čemu já říkám řízení se explicitními pravidly, s Wittgensteinovými „procesy obsahujícími pravidlo“ je jistým zjednodušením, myslím, že pro účely tohoto textu jej můžeme přijmout.

Podstatné mi připadá, co Ivan myslí tím, že „Wittgenstein nepripúšťa inú možnosť“. Chce tím snad říci, že vše, co už je za hranicemi řízení se explicitními pravidly, je stejné povahy – jsou to všechno „procesy“, které jsou *pouze* „v souladu s pravidlem“? To by bylo v přímém rozporu s tím, co si myslím já. Já jsem skálopevně přesvědčen, že i když pomineme explicitní pravidla, stále musíme rozlišovat mezi tím, když se někdo řídí pravidlem (byť není explicitně přítomno) a tím, když jeho chování pouze vykazuje

[3] V originále “involving”; Glombíček to ve svém českém překladu *Modré a hnědé knihy* (Wittgenstein, 1958) překládá jako „proces, na němž je účastno pravidlo“.

pravidelnost, která je s pravidlem v souladu. Já se domnívám, že bez takového rozlišení se neobejdeme, a proto hovořím v tom prvním případě o řízení se implicitním pravidlem a v onom druhém o pouhé pravidelnosti.

Tohle, myslím, velmi jasně vyložil Sellars na příkladu jazyka:

> ... můžeme říci, že učit se užívat jazyk znamená se dospívat k tomu, že člověk činí A v situaci C, A' v situaci C', atd. kvůli systému „tahů", ke kterému tyto činy patří, a přitom můžeme odmítnout, že učit se užívat jazyk znamená dospívat k tomu, že člověk činí A v situaci C, A' v situaci C' *s úmyslem realizovat* systém tahů. Musíme zkrátka rozlišovat mezi chováním, které je ‚vedené vzorcem', a tím které se ‚řídí pravidly', kde to druhé je složitější jev, který ten první zahrnuje, ale není s ním totožné. Chování, které se řídí pravidly, v jistém smyslu obsahuje jak hru, tak metahru, kde ta druhá je hrou, ve které tvoří pravidla následovaná v rámci té první součást chování řízeného pravidly. (Sellars 1963, 208-209)

Sellars tedy v důsledku rozlišuje mezi (1) chováním, které je pouze pravidelné, (2) chováním, které je vedené vzorcem a (3) chováním, které je řízené pravidly. Mně tato terminologie připadá poněkud matoucí, takže to podávám tak, že ve druhém i ve třetím případě jde o chování řízené pravidly, přičemž v tom druhém hovořím o implicitních pravidlech, zatímco v tom třetím o pravidlech explicitních.

Není mi jasné, do jaké míry Ivan s rozlišením, které já popisuji jako rozdíl mezi řízením se implicitním pravidlem a pouze pravidelným chováním, nesouhlasí. (Pokud ho odmítá, pak je náš pohled na celou tu věc skutečně těžko smiřitelný, protože mně toto rozlišení připadá zcela zřejmé a nepominutelné.) Na některých místech jeho textu se zdá, že právě poukaz na jeho iluzornost je jeho vůbec hlavním argumentem proti mně. Například na závěr své obsáhlé úvahy o tancování, ve které konstatuje, že když někdo zvládne nějaký tanec, tancuje, aniž by při tom měl na mysli nějaká pravidla (s čímž jistě nelze než souhlasit), konstatuje:

> No tancovať môžeme rôzne, pohybovať sa náhodnými pohybmi a mykať telom. Niektoré náhodné tance, obzvlášť pokiaľ má niekto „tanec v krvi", môžu byť v súlade s pravidlom konkrétneho tanca. Čo podľa mňa robí Wittgenstein, je upozornenie, že v tomto prípade ľudia nedodržiavajú pravidlo. Keďže však konajú pravidelne, z ich správania môžeme pravidlá odvodiť. (Ivan 2014, 91)

Myslím, že takto jednoduché to není. Asi se s Ivanem shodneme, že málokdo tancuje podle explicitních pravidel; tancování je podle mě obvykle záležitost

jisté praktické „zručnosti“, která se získává tak, že napodobuji ostatní, případně se řídím radami a opravami nějakého učitele. Velká většina lidí by podle mne nedokázala slovy popsat, jak se správně tancuje ten či onen tanec. Takže v tomto případě se v podstatě téměř vůbec nejedná o řízení se explicitními pravidly (o „procesy obsahující pravidla“). Chce tedy Ivan říci, že pravidla jsou v tomto případě ve hře jenom v tom smyslu, že by jejich pomocí mohl vnější pozorovatel popsat, co lidé na parketu dělají, stejně tak jako v případě, kdyby se na parket vyrojili králíci a úžasnou shodou náhod se začali pohybovat tak jako tančící lidé? Že tedy mezi tančícími lidmi a náhodně se pohybujícími králíky není vůbec žádný rozdíl? To mi připadá naprosto absurdní.

Můj názor je ten, že v případě tancujících lidí, na rozdíl od poskakujících králíků, hrají určitá pravidla zásadní roli, i když nejsou během toho tance explicitně přítomna (v hlavách tanečníků ani nikde jinde na parketu); a že lze říci, že ti lidé dělají to, co dělají, *kvůli těmto pravidlům*. To, co dělají, totiž dělají v důsledku toho, že byli nějakou výchovou, výcvikem či usměrňováním vpraveni do jistého druhu sellarsovského „chování vedeného vzorcem“, že normativní postoje, které k nim uplatňovali ostatní, nějak internalizovali natolik, že jsou jimi nyní implicitně vedeni. Takže říká-li Ivan (2014, 90-91), „[k]eď uvažujeme o tancovaní ako o akomsi zautomatizovanom pohybovaní, preformulovanie našich praktík pomocou prítomnosti pravidla nič nepridáva – muselo by mať nejakú silu, ktorá by vysvetlila toto správanie“, nezbývá mě než s tím zásadně nesouhlasit. I některé formy „zautomatizovaného pohybování“ jsou vykonávány „kvůli pravidlu“ – a to tehdy, když jsou součástí komplexnějších vzorců chování, které zahrnují („normativní“) postoje a reakce k těmto formám, v jejichž důsledku se tyto formy stávají takovými, jakými jsou.

Ivan (2014, 91) konstatuje: „Wittgenstein podľa mňa upozorňuje na to, že v tomto prípade ľudia nedodržiavajú pravidlo.“ Jakkoli z hlediska naší diskuse není tak docela podstatné, co si o té věci myslí Wittgenstein, tady musím s Ivanem nesouhlasit a nechápu, čím je podle něj tento jeho závěr podložen. Zdá se mi, že připisovat Wittgensteinovi, že nevidí rozdíl mezi tím, co dělá tancující člověk, a tím, co dělá králík, který se shodou okolností pohybuje jako tancující člověk, je prostě pošetilé.

Mate mě ovšem, že na některých místech svého textu se i sám Ivan vyjadřuje tak, jako by považoval za samozřejmé, že člověk se *může* řídit pravidlem, aniž by toto pravidlo bylo nějak explicitně „přítomno“. Tak např. Ivan (2014, 92) píše: „Koncom zdôvodňovania, prečo dodržiavame pravidlo nejakým spôsobom, je nezdôvodnené konanie (dodržiavanie pravidla) a nie

pravidlo, ani interpretácia pravidla, ani fakt o našom mentálnom stave." Tohle je přesně to, co si myslím já; ale pokud jde o *dodržování pravidla* a pokud při tom není to pravidlo explicitně přítomno a interpretováno, jak to nazvat jinak než řízení se *implicitním* pravidlem? A dále: „Okrem takého konania podľa pravidla, kedy sa rozhodujem, ako si ho vyložiť, dodržiavame pravidlo aj bez akéhokoľvek rozhodovania. ‚Jestliže se držím pravidla, nevolím. Řídím se pravidlem *slepě*' (Wittgenstein 2006, §219). Jednoducho, neovládame jazyk, lebo chápeme pravidlá, ale len vďaka tomu, že konáme istým spôsobom, možno povedať, že aplikujeme pravidlá (Baker 1984, 280)." To je opět něco, s čím naprosto souhlasím, jen mi není jasné, jak to zapadá to toho, co říká Ivan výše. Z jeho předchozího textu jsem nabyl dojem, že není-li v nějakém „procesu" pravidlo „přítomné" (jako „symbol"), pak jde o pouhou pravidelnost, a nikoli dodržování pravidla. Protože pokud je, pak tady máme onu trichotomii, o kterou jde mně: 1. řízení se explicitním pravidlem (symbolem); 2. řízení se pravidlem, které není explicitní; 3. pouhá pravidelnost.

Je ovšem pravda, že způsob, kterým v knize hovořím o explicitních a implicitních pravidlech, je velmi schematický a celý tento pojmový aparát by bylo dobré poněkud zjemnit a diferencovat. Představme si pravidlo, které je vyjadřováno nápisem nad jezdícími schody v metru: „Na schodech stůjte vpravo!" Lidé, přicházející ke schodům, si tento nápis přečtou a řídí se pravidlem, které vyjadřuje (někteří samozřejmě ne). To je případ explicitního pravidla. Představme si na druhé straně, že pravidlo stát na schodech na pravé straně se vyvine spontánně – lidé začnou naznačovat těm, kdo vpravo nestojí, svoji nevoli a nováčci v metru brzy vycítí, že když stojí na schodech vlevo a blokují ty, kteří chtějí po schodech nahoru stoupat, stávají se terčem nevraživosti a řídí se tím (nebo opět neřídí). To je případ implicitního pravidla.

Představme si ale, že pravidlo je, tak jako v prvním uvedeném případě, zavedeno explicitně, ale příslušný nápis je poté odstraněn. Někteří lidé, kteří přicházejí ke schodům, si na ten nápis vzpomenou a to jim připomene, že by měli stát vpravo. Jiní už na něj zapomněli nebo ho nikdy neviděli a to, že by měli stát vpravo, vycítí z reakcí ostatních. V tomto případě se zřejmě nedá jednoznačně hovořit ani o explicitním, ani o implicitním pravidle. Přiznávám, že mě v knize šlo především o ty krajní body a tím, co je mezi nimi, jsem se příliš nezaobíral.

Nyní se mi zdá, že rozlišení mezi *implicitním* a *explicitním* by bylo třeba zjemnit minimálně tak, že bychom na jedné straně rozlišovali mezi explicitními a implicitními *pravidly*, a na druhé straně rozlišovali mezi případy explicitního řízení se pravidlem a případy implicitního řízení. (Kritériem pro

to první rozlišení by mohlo být například to, jakým způsobem vešlo dané pravidlo v platnost: zda bylo „předepsáno“, či zda se vyvinulo spontánně. Kritérium pro to druhé rozlišení by se pak mohlo více blížit tomu, o čem uvažuje Wittgenstein v *Modré a hnědé knize*, jak to výše cituje Ivan.[4])

V každém případě by pak bylo lze říci, že se někdo *implicitně* řídí *explicitním* pravidlem. To by mohl být případ lidí, kteří chodí pod cedulí „Na schodech stůjte vpravo!“, aniž by si ji všimli a stojí vpravo proto, že jsou do tohoto modu chování vmanipulováni ostatními; případně těch, kdo si ten nápis nejprve četli a řídili se podle něj, ale pak už to začali dělat automaticky. V jistém smyslu by pak snad bylo možné hovořit i o tom, že se někdo naopak *explicitně* řídí *implicitním* pravidlem. To by mohl být případ toho, kdo je do toho, aby stál vpravo, vmanipulováván ostatními, ale tento fakt reflektuje a uvědomuje si, že se řídí určitým pravidlem, i když toto pravidlo není nikde explicitně formulováno.

Dalším pojmovým rozlišením, které jsem v knize nezaváděl a které by mohlo být užitečné, je rozlišení mezi *řízením* se pravidlem a *dodržování* pravidla. O člověku, který nikdy nejel v metru asi můžeme říci, že *dodržuje* (ve smyslu *neporušuje*) pravidlo, že na schodech se má stát vpravo, ale říci o něm, že se tímto pravidlem *řídí*, může znít poněkud podivně.

Tohle rozlišení ovšem souvisí s rozlišením, kterým se v knize zabývám: totiž s (vágním) rozlišením mezi *preskriptivními* a *restriktivními* pravidly. Preskriptivní pravidlo mi, zjednodušeně řečeno, říká, že něco mám udělat, že mám vyvinout nějakou aktivitu, a dodržet ho tedy nelze jinak, než tak, že udělám, co mi přikazuje – tedy že se jím budu řídit. Avšak většina těch pravidel, o které mi v knize jde a které se týkají jazyka a společenských „virtuálních prostorů“, o jakých tam hovořím, jsou pravidly restriktivními. Neříkají nám, co udělat máme, ale co udělat *nesmíme*. Dodržet taková pravidla pak lze někdy i bez toho, abychom jim věnovali jakoukoli pozornost – takže v jejich případě se dodržování pravidla stává něčím, co mohu dělat, aniž bych se tím pravidlem v nějakém podstatném slova smyslu řídil.

V knize jsem upozorňoval na to, že Wittgensteinovy úvahy o pokračování číselné řady ze *Zkoumání* mohou být matoucí proto, že se týkají povýtce preskriptivního typu pravidla (takové pravidlo stanoví, co musím udělat – totiž jaké následující číslo musím vyslovit či zapsat), zatímco většina pravidel, která jsou podstatná z hlediska fungování jazyka, jsou povýtce restriktivní (tato pravidla nám nikdy neříkají, co konkrétně máme někdy říci či napsat).

---

[4] Zdá se mi být také ve shodě s tím, co navrhuje Svoboda (2012).

Řekl jsem, že se mi zdá, že Ivan popírá hranici mezi řízením se implicitním pravidlem a pouhou pravidelností. Na některých místech jeho textu se mi ale jeho argument proti uvedené hranici zdá nabývat poněkud odlišné podoby; jako by už neříkal, že vůbec neexistuje, ale jenom to, že je příliš vágní a nejasná. Ptá se například: „To, aký je rozdiel medzi dodržaním a nedodržaním pravidla, medzi robením danej činnosti akosi náhodou a jej realizáciou na základe naučenia pravidiel, nemôžeme určiť paušálne vopred, ale musíme sledovať kritériá konkrétnych praktík, v ktorých rôzne pravidlá majú svoj život“ (Ivan 2014, 91). Ano, tady není žádné jasné kritérium. Je to ale fatální? Nemáme také žádné jasné kritérium pro to, kdy něco nazveme lesem a kdy ne. (Možná ministerstvo zemědělství nějaké má; ale to nemusí mít mnoho společného s tím, jak se slovo „les“ skutečně používá.) Slovo „les“ nám přesto velice úspěšně slouží. Já se v žádném případě nesnažím tvrdit, že mezi chováním, které je pouze pravidelné, a chováním, které se řídí implicitním pravidlem, je ostrá hranice nebo že bych takovou hranici dokázal narýsovat; o tom, že takový rozdíl existuje, se mi ale nezdá být možné pochybovat.

Všimněme si také, že ani hranice okolo dodržování explicitních pravidel (Ivanova případu, kdy je pravidlo „přítomné“, kterou, nakolik mu rozumím, on sám uznává) není v žádném případě ostrá. Když pojedu autem a budu si u toho pobrukovat dopravní předpisy, je tu pravidlo rozhodně přítomné; činí to ale z mého počínání skutečné řízení se explicitním pravidlem? Co když si ty předpisy přeříkávám bez toho, abych jim rozuměl, co když jsou dokonce v jazyce, kterým nehovořím? Takže říká-li, že „jazykové hry, ktorými opisujeme dodržiavanie pravidla, keď je pravidlo prítomné v uvedenom zmysle a keď pravidlá na základe pozorovania odvodíme, sú rôzne“ (Ivan 2014, 92), pak je tu podle mne úplně stejný druh různosti, jako mezi pouhými pravidelnostmi a implicitními pravidly – není tu žádná ostrá hranice, ale můžeme udat spoustu případů, z nichž je různost zřejmá.

Zásadním problémem se mi zdá být to, že Ivan kolísá mezi „sociologickým“ pojetím pravidel, které je blízké mě, a pojetím „psychologickým“, které podle mne vede do slepé uličky. Např. Ivan (2014, 91) říká: „Dodržiavaním pravidiel (a nie konaním zo zvyku) je nejaká činnosť vďaka tomu, že je zapojená do celého radu praktík, ktoré vytvárajú život (konkrétneho) pravidla.“ To se mi zdá být přesně to, co říkám já.

Na jiných místech se však naopak vyjadřuje tak, jako bychom o tom, zda nějaký jedinec dodržuje pravidlo, mohli rozhodnout čistě na základě obsahů jeho mysli/mozku; tak např. nás vyzývá, abychom si implicitní pravidla představili jako tabulku v mozku (srov. Ivan 2014, 96). To mi dává asi stejně tak

smyslu jako „Představme si univerzitu jako tabulku v mozku“. Pravidlo, jak já tento pojem chápu, nemůže být nikdy v ničím mozku, existuje bytostně v průsečíku postojů různých jedinců, a tak nutně musí být mimo každý jednotlivý mozek. Tak jako každá jiná institucionální realita. (Je to *realita* v tom smyslu, že to, co si o ní myslí kterýkoli jednotlivec, na ní samo o sobě nemůže nic změnit. Ostatně rozdíl mezi *řízením se pravidlem* a *domněnkou, že se řídím pravidlem* byl základní bod Wittgensteinova pojetí pravidel.)

Uvažme z tohoto hlediska „ilustraci“, kterou Ivan uvádí v závěru svého textu a o níž se asi domnívá, že vynáší na světlo nějaké fatální nedostatky mého postoje:

> Na ilustráciu si predstavme niekoľko ľudí idúcich po pohyblivých schodoch: 1) A stojí na eskalátore vždy vpravo, lebo má kompulzie; 2) B sa vpravo jednoducho cíti lepšie, bezpečnejšie; 3) C robí to, čo ostatní; 4) D si podľa správania ostatných odvodil pravidlo „Nestoj v ceste rýchlejším!“; 5) E si prečítal pokyn „Drž sa vpravo“. Pri pohľade na schody povieme, že všetci dodržiavajú pravidlo držania sa vpravo. Je u všetkých prítomné? Dodržiavajú ho rovnakým spôsobom? (Ivan 2014, 98)

To, o čem zde hovoří, jsou psychologické motivace lidí, kteří jedou po schodech. Ale vyvozovat z psychologických motivací jednotlivce něco o existenci pravidla je ošidné: znovu musím připomenout Wittgensteinův základní vhled týkající se toho, že když se někdo domnívá, že se řídí pravidlem, rozhodně to samo o sobě neznamená, že se pravidlem skutečně řídí.

Abychom mohli o někom, kdo jede po schodech, říci, že se řídí pravidlem, musíme především vědět, že tu to pravidlo je; a aby tu bylo pravidlo, musí tu být nějaká komunita lidí, díky které tu je. To pravidlo by mohlo být například formulováno v přepravním řádu metra, a protože každý, kdo využívá služeb metra, se zavazuje řídit přepravním řádem, každý, kdo jede metrem, se v tomto smyslu tímto pravidlem řídí. Samozřejmě někdo se jím řídí i v tom smyslu, že se velmi pečlivě vyhýbá jeho porušování; zatímco jiný na něj kašle, někdo ho dokonce naschvál porušuje.

Jiná možnost, bližší realitě, by byla ta, že by se to pravidlo etablovalo jako nepsané, prostřednictvím normativních postojů velké většiny cestujících – každý, kdo se nebude chovat v souladu s ním, bude terčem nevraživosti, což ho asi postupně usměrní. A zase, bude tu většina, která se bude snažit nevraživosti cestujících předejít; pak tu budou někteří, kteří ji vůbec nezaregistrují; a pak tu budou třeba i takoví, kteří o ní budou vědět, ale naschvál se jí budou vzpírat. Může tomu samozřejmě být i tak, že ty normativní po-

stoje budou záležitostí jenom nějaké vybrané podskupiny cestujících, řekněme nějakých „pravých pražáků“, kteří se budou dodržováním takových pravidel chtít distancovat od všelijakých těch „náplav“ a „vidláků“.

Které z postav, uvedených na scénu Ivanem, se tedy budou pravidlem řídit? Na základě toho, co o nich píše, se to podle mne vůbec nedá říci. Museli bychom vědět nejenom to, jak se na schodech chovají (a to ne jenom v jednom okamžiku), ale jak reagují na chování jiných na schodech, jaké jsou jejich institucionální vazby na případné autority, které takové pravidlo vydávají atd. A tady se také nelze nevrátit k tomu, co už jsem napsal výše: bylo by bláhové očekávat, že nám nějaká teorie pravidel dodá jasná kritéria pro rozhodnutí, kdy se kdo řídí pravidlem, a kdy ne. Současně si ale nemyslím, že by to znamenalo, že takový pojem řízení se pravidlem patří na filosofické smetiště.

Vezměme pro srovnání pojem náboženství, o kterém snad nikdo nebude pochybovat, že je při vysvětlování konání některých lidí a některých lidských společenství užitečný. Kde začíná a kde končí náboženství? Je buddhismus náboženství? Je jím konfucianismus? Je jím nějaké uctívání totemů? A kdy řekneme, že je někdo účasten na daném náboženství? Když to říká? Když se účastní podstatných rituálních praktik? Ale které jsou ty podstatné a jak moc se jich musí účastnit? Bylo by bláhové se domnívat, že na takové otázky existují jednoznačné odpovědi; a současně by bylo bláhové se domnívat, že absence takových jednoznačných odpovědí znehodnocuje celý ten pojem. Podstatné je, že existuje spousta jednoznačných náboženství a spousta lidí, kteří se jich jednoznačně účastní; a že právě jejich příslušnost k oněm náboženstvím vysvětluje spoustu z toho, co dělají. A podobně je to podle mne s pravidly: jak jsem argumentoval na jiných místech, jsem skálopevně přesvědčen, že jedině skrze pravidla můžeme vysvětlit jazykový význam (viz poznámka 4); a mám tendenci si myslet, že tohle je rozšiřitelné na význam v podstatě v jakémkoli širším smyslu, takže pravidla jsou klíčem ke specificky lidskému světu.

Projdu nyní podrobněji Ivanovy kritické připomínky, jak je shrnuje v závěru svého textu:

> Prvý problém s implicitnými pravidlami spočíva v otázke, či možno implicitné pravidlo vôbec dodržiavať. Mali by sme povedať, že dodržiavame pravidlo, no nevieme aké. Nielen to, napokon, pravidlo môžeme zabudnúť a iba robiť to, čo treba. Mali by sme dodržiavať pravidlo, ktoré ani nie je možné vyjadriť. (Ivan 2014, 95)

Dodržovat implicitní pravidlo znamená zohledňovat normativní postoje členů příslušné komunity, případně je nějak „internalizovat“. Tancuji-li nějaký

konkrétní tanec, dodržuji také spoustu pravidel, které vůbec nemusím být schopen formulovat (znamená to, že nevím, jaká ta pravidla jsou?). Máme-li k dispozici jazyk, můžeme implicitní pravidlo vyjádřit. Test, který odlišuje situaci, ve které jde o implicitní pravidla, a tu, kde se jedná nanejvýše o pravidelnosti, spočívá především v detekci přítomnosti normativních postojů – to jest reakcí lidí na to, co v relevantním ohledu dělají jiní lidé.

> Mohli by sme povedať, že svojou hrou pravidlá *vytvárame*. V akom zmysle? Používame tu metaforu bez toho, aby sme vedeli, čo ňou skutočne myslíme. (Ivan 2014, 96)

Obávám se, že tady chce Ivan říci, že já nevím, co skutečně myslím tím, když říkám, že *vytváříme pravidla*. Tohle já ale opravdu vím (ať už je to, co tím myslím, pro někoho přijatelné nebo ne) a snad jsem to jasně popsal výše. Pravidlo pro mě, jak už jsem říkal několikrát, vzniká v průsečíku normativních postojů různých lidí, tyto rezonující postoje vedou k jistému usměrňování lidského chování.

> Pokiaľ implicitné pravidlá nemožno spoznať *priamo*, ale iba v niektorých prípadoch ich možno preložiť do slov, ako vieme, či formulované pravidlo zodpovedá tomu implicitnému? Pokiaľ sú diskurzívne formulované pravidlá interpretáciou implicitných pravidiel, ktoré ich ontologicky predchádzajú, a zároveň ich nedokážeme nikdy spoznať, otázkou zostáva, ako ich dokážeme interpretovať. (Ivan 2014, 96)

Pokud jde o tu první otázku, příliš ji nerozumím. O „překládání" implicitních pravidel do slov bych nehovořil, spíše o jejich slovním vyjádření.[5] A otázka, jak víme, že vyjádření nějakého pravidla vyjadřuje právě toto pravidlo, mi nezní o mnoho smysluplněji než otázka, jak víme, že věta *Venku prší* vyjadřuje, že venku prší. (Jak by řekl Wittgenstein, víme to proto, že umíme česky.) Co Ivan myslí „interpretací" implicitních pravidel, stejně tak jako to, proč se domnívá, že tato pravidla nikdy nedokážeme poznat, mi zůstává záhadou.

> S pojmom pravidla sa spájajú praktiky ako dodržiavanie pravidla, výnimka z pravidla, vytváranie pravidla a iné. Žiadne z nich nie je možné spájať s implicitnými pravidlami. Podstatná je aj otázka, či je možné tieto pravidlá takpovediac pochopiť – tvrdiť, že dochádza k nejakému implicit-

[5] Tady bych odkázal na Svobodovo rozlišení mezi S-pravidly a J-pravidly – viz Svoboda (2012).

> nému chápaniu, vytvára klamný obraz o konaní po „absorbovaní" pravidiel, ktoré je vhodnejšie predstaviť si bez procesu uchopovania, bez propozičného obsahu, skrátka ako v istom zmysle zautomatizované a nediferencované konanie. (Ivan 2014, 97)

Musím říci, že vůbec nerozumím, proč by s implicitními pravidly nebylo možné spojovat „praktiky ako dodržiavanie pravidla, výnimka z pravidla, vytváranie pravidla a iné". Tak, jak já pojmu implicitní pravidlo rozumím (a jak jsem se to snažil vysvětlit výše), mohou implicitní pravidla spontánně vznikat (není je ovšem možné vytvářet nějakými dekrety), mohou být dodržována (či porušována) a jistě z nich mohou existovat i výjimky. (Představme si znovu implicitní pravidlo týkající se toho, že se na jezdících schodech v metru jezdí napravo. Dovedu si velice dobře představit, že bude-li Karel Gott stát nalevo, budou jinak v takové situaci aktivované normativní postoje většiny ostatních cestujících potlačeny obdivem k Mistrovi.) Nepřipadá mi podstatné, zda v případě, že se někdo stane vnímavým k nějakým normativním postojům svého okolí, budeme hovořit o „implicitním pochopení", o nabytí nějaké společenské dovednosti či třeba o zvládnutí techniky.

> To, že idea robenia niečoho správneho a nesprávneho sa nijako automaticky nespája s pravidlami, som sa pokúsil ukázať v časti Normatívne praktiky. Iste, nič nám nebráni rozšíriť používanie pojmu pravidiel aj na tento prípad. Pohybovali by sme sa však v kruhu. Najskôr by sme chceli pojem implicitného pravidla spojiť s druhým spôsobom správania sa (vykonať niečo správne/nesprávne), no na to, aby sme to mohli urobiť, museli by sme najskôr identifikovať zodpovedajúce prípady použitia výrazu konať správne. Ako ich však vyberieme? (Ivan 2014, 98)

Přiznám se, že nerozumím, proč by neměla být pravidla „automaticky" spojená se správností a nesprávností. Pravidlo, jak toto slovo běžně užíváme, je podle mne nástrojem určení nějaké správnosti a nesprávnosti. Takže v tomto směru se mi zdá být pravidlo se správností/nesprávností propojeno zcela intimně. Připouštím, že v opačném směru to není tak zřejmé; ale připustíme-li, že pravidlo nemusí být explicitní (a bez toho podle mého názoru jednak nedokážeme vysvětlit, jak pravidla vznikla a jednak se nedokážeme vypořádat s Wittgensteinovým regresním argumentem – viz Wittgenstein 1953, §85), pak, zdá se mi, je i ten opačný směr přímočarý.

> Ak by sme pripustili, že vždy, keď dokážeme odvodiť pravidlá, môžeme povedať, že dodržiavame implicitné pravidlá, ich pomyselná všadeprítomnosť by zahalila ostatné spôsoby správania. (Ivan 2014, 98)

Předpokládám, že případ „keď dokážeme odvodiť pravidlá“, je případem, který vykazuje pravidelnosti, což je v podstatě každý případ, neboť nějakou pravidelnost lze najít všude. Rozhodně ale ne v každém takovém případě jde o implicitní pravidla v mém slova smyslu – ta předpokládají normativní postoje.

> Dodržiavať pravidlá je jedna súčasť konania medzi inými. Povedať, že všetko zmysluplné konanie je založené na pravidlách, je metafyzická predstava. Môžeme to brať ako definíciu, no otázna je užitočnosť pojmu zmysluplného konania, ktorý takto získame. Podobne úsilie určiť vopred kritérium rozlíšenia medzi pravidlom a konaním zo zvyku, kritérium dodržiavania pravidla pre akékoľvek možné pravidlo, je odsúdená na neúspech, keďže pravidlá fungujú rôzne. Navyše, čo som sa pokúsil ukázať, dodržiavanie pravidiel má rôzne podoby. (Ivan 2014, 98-99)

Tady jsem, musím říci, trochu na rozpacích z té „metafyziky“. Možná se v lecčems mýlím, ale ke své koncepci jsem se propracoval skrze svůj intelektuální zápas s tím, jak vysvětlit, jak funguje jazyk, a co vlastně říkáme tím, když říkáme, že nějaké slovo má takový a takový význam. Dopracoval jsem se k přesvědčení, že typy zvuků, které používáme ke komunikaci, se stávají v rámci našich komunikačních praktik smysluplnými podobným způsobem, jakým se v rámci šachů stávají z kousků dřeva pěšci, střelci či věže. Nahlédl jsem, že pravidla mají schopnost otevírat nám nové prostory pro nové druhy akcí; a domnívám se, že tohle by mohlo být klíčem k charakterizaci toho, čím se lidský („smysluplný“) svět liší od světů, ve kterém žijí jiní živočichové. Možná nemám pravdu, ale proč metafyzika?

> Môžeme teda súhlasiť s Koreňom, že Peregrin očakáva od pravidiel príliš veľkú teoretickú úlohu. Preto sa pokúša spojiť aspekty rôznych pravidiel do jedného robustného konceptu a použiť ho na vysvetlenie rôznorodých situácií. (Ivan 2014, 99)

Ano, mě se pravidla začala jevit jako klíč k vysvětlení významů, k charakterizaci specificky lidského světa i k lidské mysli; možná od nich tedy skutečně očekávám „príliš veľkú teoretickú úlohu“. To, že se snažím spojit „aspekty rôznych pravidiel do jedného robustného konceptu“, ale asi vnímám jako mnohem méně problematickou věc než Ivan. Jsem si vědom toho, že spoustu věcí zjednodušuji; nicméně jsem přesvědčen, že i to je úloha filosofie. Mám pocit, řečeno s Wittgensteinem, že „filosofický problém má podobu: ‚Nevyznám se v tom‘“; a napomoci k tomu, aby se „v tom“ člověk vyznal, mohu jedině tak, že předvedu nějaké *přehledné znázornění* (*übersicht-*

liche Darstellung – viz Peregrin 1992) celé té problematické věci, na základě kterého bude v chaosu možné spatřit nějaký náznak řádu. Filosofie tedy podle mne nemá popisovat a předpovídat (protože pak se marně pokouší konkurovat empirickým vědám), ale především předvádět zjednodušené modely, jejichž pomocí se začínáme v tom, v čem se nevyznáme, orientovat.

Literatura

GENDLER, T. S. (2008a): Alief in Action (and Reaction). *Mind & Language* 23, 552-585.

GENDLER, T. S. (2008b): Alief and Belief. *Journal of Philosophy* 105, 634-663.

IVAN, M. (2014): O pravidlách. *Organon F* 21, č. 1, 82-100.

KOREŇ, L. (2012): Pravidlá: spoločnosť, jazyk a racionalita. *Teorie vědy/Theory of Science* 33, 591-615.

KRIPKE, S. (1982): *Wittgenstein on Rules and Private Language*. Cambridge (Mass.): Harvard University Press.

PEREGRIN, J. (1992): Sprache und ihre Formalisierung. *Deutsche Zeitschrift für Philosophie* 40, 237-244.

PEREGRIN, J. (2011a): *Člověk a pravidla. Kde se berou rozum, jazyk a svoboda*. Praha: Dokořán.

PEREGRIN, J. (2011b): Creatures of Norms as Uncanny Niche Constructors. In: Hříbek, T. – Hvorecký, J. (eds.): *Knowledge, Value, Evolution*. London: College Publications, 189-198.

PEREGRIN, J. (2012a): Člověk jako normativní tvor. *Teorie vědy/Theory of Science* 34, 3-23.

PEREGRIN, J. (2012b): The Normative Dimension of Discourse. In: Allan, K. – Jasczolt, K. (eds.): *Cambridge Handbook of Pragmatics*. Cambridge: Cambridge University Press, 209-225.

PEREGRIN, J. (2012c): Semantics without Meaning? In: Schantz, R. (ed): *Prospects of Meaning*. Berlin: de Gruyter, 479-502.

PEREGRIN, J. (2012d): Inferentialism and the Normativity of Meaning. *Philosophia* 40, 75-97.

PEREGRIN, J. (2014): Rules as the Impetus of Cultural Evolution. *Topoi*, vyjde.

SELLARS, W. (1963): Empiricism and Abstract Entities. In: Schilpp, P. A. (ed): *The Philosophy of Rudolf Carnap*. LaSalle: Open Court, 431-468.

SVOBODA, V. (2012): Pravidla, normy a analytický filozofický diskurz. *Organon F* 19, 143-179.

WITTGENSTEIN, L. (1953): *Philosophische Untersuchungen*. Oxford: Blackwell. Český překlad: *Filosofická zkoumání*. Praha: FLÚ AV ČR 1993.

WITTGENSTEIN, L. (1958): *The Blue and Brown Books*. Oxford: Blackwell. Český překlad: *Modrá a hnědá kniha*. Praha: Filosofia 2006.

Organon F 21 (3) 2014: 399-408

Externalizmus, skepticizmus a zdôvodnenie: Odpovede M. Pichovi a M. Taligovi

MARTIN NUHLÍČEK

Katedra filozofie a dejín filozofie. Filozofická fakulta. Univerzita Komenského v Bratislave
Šafárikovo nám. 6. 814 99 Bratislava 1. Slovenská republika
nuhlicek@fphil.uniba.sk

V tomto texte sa pokúsim odpovedať M. Pichovi a M. Taligovi, ktorých články Picha (2013) a Taliga (2013b) boli reakciami – vo viacerých ohľadoch kritickými – na moju stať *Prečo externalistické kritériá zdôvodnenia principiálne zlyhávajú?* (pozri Nuhlíček 2013b). Obidvom autorom ďakujem za ich podnetné odpovede, ktorými ma primäli znovu premyslieť niektoré otázky týkajúce sa epistemického zdôvodnenia, ako aj skepticizmu. Vďaka nim sa odhalilo niekoľko miest v mojej argumentácii, ktoré sa môžu javiť z vecného hľadiska ako problematické, prípadne môžu vyvolávať riziko nesprávneho pochopenia kvôli formulačnej ťažkopádnosti. Preto by som rád využil túto príležitosť a vrátil by som sa k niektorým kľúčovým bodom mojej predošlej state, bližšie by som ich vysvetlil, poopravil či lepšie sformuloval. Verím, že viaceré Pichove a Taligove kritické námietky sa potom ukážu ako prosté nedorozumenia. Obávam sa však, že medzi tvrdeniami jedného i druhého autora sa nájdu aj také, s ktorými nemôžem súhlasiť. Teda okrem spresnenia mojich vlastných stanovísk chcem priestor tohto textu využiť aj na stručnú kritickú diskusiu o vybraných otázkach, ktoré vo svojich statiach tematizujú spomenutí autori. Zrejme ťažko možno pochybovať o tom, že vyjasnenie názorových pozícií je základnou podmienkou každej konštruktívnej diskusie. Môj text predkladám s ambíciou prispieť k tomuto cieľu.

V prvej časti tohto textu budem reagovať na článok M. Pichu, v druhej časti na článok M. Taligu. V prípade M. Taligu budem na niektoré veci reagovať v nadväznosti na diskusiu o niektorých epistemologických otázkach,

ktorá medzi ním, mnou a ďalšími autormi už istý čas prebieha v odborných periodikách (pozri najmä Taliga 2009; 2010; Démuth 2012; Nuhlíček – Szapuová 2012; Taliga 2012; Gahér 2013; Nuhlíček 2013a; Szapuová 2013; a Taliga 2013a).

1. Odpoveď M. Pichovi

M. Picha si všíma, že v texte Nuhlíček (2013b, 39-40) poukazujem na neschopnosť internalistického prístupu k zdôvodneniu vyrovnať sa s hrozbou filozofického skepticizmu a následne začínam hovoriť o externalizme ako alternatíve k internalizmu. Pripúšťam, že takáto postupnosť v mojom texte mohla vyvolať dojem, že externalizmus vznikol primárne s ambíciou riešiť problém skepticizmu, na ktorý bol prikrátky internalizmus. Takýto dojem je samozrejme nesprávny a rád by som uviedol veci na pravú mieru. Ako je známe, externalistické chápanie zdôvodnenia sa historicky objavilo v rámci intenzívnej diskusie o poznaní a zdôvodnení, ktorá na filozofickej scéne prebiehala približne od polovice 60. rokov 20. storočia, predovšetkým v reakcii na tzv. Gettierov problém, spochybňujúci tradičnú trojzložkovú analýzu poznania. Problém skepticizmu patrí v epistemológii nepochybne medzi najdiskutovanejšie otázky, pričom ani internalizmus, ani externalizmus sa nevyhli konfrontovaniu s touto otázkou. Avšak opäť prízvukujem, že snahu odpovedať na skepticizmus (v jeho klasickej, karteziánskej podobe) nemožno považovať za hlavnú motiváciu vzniku externalistickej alternatívy v chápaní poznania či zdôvodnenia. V tomto zmysle môžem súhlasiť s tvrdením, že externalizmus je pokusom o odpoveď na to, čo Picha nazýva konformný skepticizmus (por. Picha 2013, 380-381), pokiaľ ho chápeme ako tézu, že (internalistické) zdôvodnenie nie je postačujúcou podmienkou pre poznanie. Rovnaký záver stojí tiež v jadre spomínaného Gettierovho problému (pozri Gettier 1963).

Pristavím sa ešte pri probléme skepticizmu. Je namieste spresnenie, že skepticizmus, s ktorým pracujem, celkom nezodpovedá pôvodnému karteziánskemu pochybovaniu. Descartes sa pokúsil zaviesť globálnu pochybnosť, no tá je podľa môjho názoru v konzekventnej podobe racionálne nemožná (a ani Descartes ju podľa mňa nedosiahol). Globálny skepticizmus, teda pochybovanie o absolútne všetkých presvedčeniach, vedie k triviálnym paradoxom. Nedá sa sformulovať v podobe tvrdenia, ktoré by si mohlo nárokovať platnosť poznatku (aj samo toto tvrdenie by muselo byť spochyb-

nené); nedá sa podporiť argumentom či dôvodom (vedomosti o vzťahoch podpory medzi tvrdeniami by boli tiež spochybnené) a pod. Preto pokiaľ hovorím o skepticizme, mám na mysli skôr silný lokálny skepticizmus, čiže spochybňovanie zdôvodnenia a poznania v určitej vymedzenej oblasti (alebo triede) presvedčení. V dejinách filozofie boli najčastejšími cieľmi skeptickej pochybnosti triedy presvedčení o vonkajšom svete, ale aj o minulosti, alebo o iných mysliach, zatiaľ čo tradične sa takáto pochybnosť väčšinou nevzťahovala na presvedčenia týkajúce sa rozumového usudzovania či sebauvedomenia.

Prečo uvádzam túto odbočku? Zdá sa mi, že Picha formuluje pozíciu, ktorú nazýva radikálnym skepticizmom, trochu neopatrne. Na jednej strane ju vymedzuje pomocou modality – *nie je možné* určiť, ktoré presvedčenia sú pravdivé (pozri Picha 2013, 379) – v tom spočíva vlastná radikálnosť tejto pozície; ale na druhej strane ju tiež charakterizuje z hľadiska rozsahu pochybnosti *de facto* ako pochybovanie o všetkom: „Táto radikálna pozícia tvrdí, že ktorékoľvek naše presvedčenie – o svete navôkol, o našom vnútornom živote, o minulosti, o matematike – môže byť nepravdivé. *Všetky naše presvedčenia* môžu byť produkované postupmi, na ktorých začiatku nie je zodpovedajúci fakt, ale zákerný protihráč“ (Picha 2013, 379; kurzíva M. N.). Takéto vymedzenie príliš pripomína globálny skepticizmus, ktorý nepovažujem za akceptovateľný z dôvodov uvedených v predošlom odseku. Súhlasím s M. Pichom, že v rámci diskusie o internalizme a externalizme je vhodnejšie predpokladať tzv. konformný skepticizmus, ktorý spočíva v spochybňovaní zdôvodnenia presvedčení. No navyše si myslím, že takýto skepticizmus je tiež jediný zmysluplný, pokiaľ jeho alternatívu, radikálny skepticizmus, definujeme v termínoch globálnej pochybnosti.[1]

Nakoniec prejdem k jadru Pichovej kritiky môjho článku. Mrzí ma, ak som vyvolal dojem, že externalistom pripisujem ambíciu primárne vyvrátiť radikálny skepticizmus. Takýto výklad pochopiteľne nie je presný; a aj M. Picha sa správne (a ústretovo) nazdáva, že chcem povedať niečo iné. Pokiaľ však dobre rozumiem následnej autorovej rekonštrukcii mojej argumentácie, tak nemôžem povedať, že adekvátne vystihol moju pointu. Picha mi vyčíta, že pomocou tzv. argumentu tretej osoby a tzv. argumentu prvej osoby

[1] Pritom je zrejmé, že sám autor vidí problémy plynúce z globálneho spochybňovania všetkých presvedčení: „Bez niekoľkých pravdivých presvedčení skrátka nevieme, kam sa pri teoretizovaní chceme dostať. Nevieme, s čím návrhy porovnávať; nevieme ani to, ako ich vlastne zdôvodňovať“ (Picha 2013, 380).

(pozri Nuhlíček 2013b, 43-46), ktoré sú namierené proti prominentnej externalistickej teórii, reliabilizmu, kritizujem túto koncepciu z hľadiska jej neúčinnosti proti radikálnemu skepticizmu (pozri Picha 2013, 383). Domnievam sa, že autor si tieto argumenty vykladá tak, že pomocou nich poukazujem na nemožnosť oprieť sa s absolútnou istotou o pravdivosť akéhokoľvek presvedčenia (napríklad kvôli karteziánskej hypotéze zlého démona). Ak je to pravda, tak ma však nechápe správne. V mojom článku sa pokúšam poukázať na niečo iné: na *praktickú nemožnosť* dospieť k zdôvodneniu ľubovoľného presvedčenia, ktorá podľa mňa hrozí, pokiaľ prijmeme externalistickú pozíciu. V skratke, spomenutý reliabilizmus opisuje zdôvodnenie presvedčení v termínoch pôvodu presvedčení v spoľahlivých kognitívnych procesoch. No zároveň neposkytuje žiadny plauzibilný praktický „návod", ako identifikovať takto chápané zdôvodnenie nejakého presvedčenia. Argumenty tretej osoby a prvej osoby v kontexte môjho článku nie sú vyjadrením radikálneho skepticizmu, ale majú predstavovať rozhodujúcu ilustráciu práve spomenutej praktickej nemožnosti identifikovať zdôvodnenie (a konzekventne aj poznanie) z externalistickej perspektívy – a to *nezávisle* od skeptickej argumentácie. Externalizmus nás podľa môjho názoru ponecháva v stave kompletnej epistemickej nevedomosti; a v tomto zmysle treba chápať aj moje tvrdenie, že „neschopnosť externalizmu z hľadiska subjektu diferencovať správne a nesprávne presvedčenia považujem za hlavný dôvod jeho teoretického zlyhania" (Nuhlíček 2013b, 47).[2]

Chcem zdôrazniť, že neschopnosť *identifikovať zdôvodnenie* presvedčení z externalistickej perspektívy treba striktne odlišovať od otázky adekvátnosti externalistického *opisu zdôvodnenia*. Inými slovami, moje námietky nemajú za cieľ ukázať, že externalizmus sa mýli v opise zdôvodnenia, ale sú namierené proti praktickej použiteľnosti tejto teórie. Nevylučujem, že externalistický opis zdôvodnenia môže byť pravdivý; no ak aj je, tak je nám nanič, pokiaľ nám, subjektom, neumožňuje funkčne hodnotiť epistemický status našich presvedčení – to je podstata mojej kritiky externalizmu.[3]

---

[2] Zdá sa, že nedostatok, ktorý externalizmu pripisujem, bol prítomný už v zárodku tejto pozície. Alvin Goldman, na ktorého plecia padá nepochybne najväčšia zásluha pri artikulovaní externalizmu, poznamenal už vo svojej priekopníckej práci: „Pravdivostné podmienky by sa nemali zamieňať s verifikačnými podmienkami. Moja analýza ‚S vie p' nemá za cieľ poskytnúť procedúry na zistenie, či niekto (vrátane seba samého) vie danú propozíciu" (Goldman 1967, 372).

[3] V rovine opisu zdôvodnenia si externalizmus dokonca dokáže poradiť aj s problémom skepticizmu, a to veľmi jednoducho. Napríklad reliabilizmus môže vyhlásiť, že po-

Takže súhlasím s M. Pichom, že bez akéhokoľvek vzťažného bodu, ktorý predpokladáme ako pravdivý, je veľmi náročné, ak nie nemožné, o čomkoľvek teoretizovať (pozri Picha 2013, 380). Takisto s ním súhlasím v tom, že reliabilizmus nemá problém s hrozbou epistemického regresu (pozri Picha 2013, 383). Reliabilizmus totiž pracuje s konceptom pôvodu presvedčení (v tom zmysle, že pochádzajú z nejakého kognitívneho procesu), ktorého súčasťou nie sú vzťahy k iným presvedčeniam, a teda možnosti regresu presvedčení sa už definične vyhýba. Avšak s týmito tvrdeniami som sa v článku Nuhlíček (2013b) ani nepokúšal polemizovať. Môj kritický „útok" na externalizmus som viedol iným smerom – ktorým, to som sa pokúsil lepšie objasniť v tejto odpovedi. Ak smiem využiť Pichovu metaforu, ktorú mi adresoval v závere svojej reakcie, tak je to tak, akoby sme sa pokúšali vybrať si auto, ktoré čo najmenej znečisťuje vzduch, no narazili by sme na to, že istá skupina výrobcov údaje tohto typu zákazníkom vôbec nesprístupňuje.

2. Odpoveď M. Taligovi

M. Taliga vo svojom texte na väčšom priestore rozoberá, údajne v protiklade k môjmu pohľadu, že zdôvodnenie je istým spôsobom previazané s otázkou pravdivosti presvedčení – zdôvodňujeme, že určité presvedčenie je pravdivé; zdôvodňujeme s cieľom dosiahnuť pravdu (pozri Taliga 2013b, 386 a n.). V tomto môžem s Taligom len súhlasiť a mrzí ma, ak som sa predtým vyjadril tak nejasne, že to viedlo ku konfúzii. Nebudem sa tu vracať k celej argumentácii, ktorú uvádzam na kritizovanom mieste (pozri Nuhlíček 2013a, 146 a n.), ale pokúsim sa ešte raz – verím, že zrozumiteľnejšie – sformulovať, aké sú v mojom chápaní základné črty vzťahu medzi zdôvodnením a pravdivosťou presvedčení.

Podľa môjho názoru si netreba pliesť dve otázky: 1. Na čo nám slúži alebo má slúžiť zdôvodnenie? 2. Čo je to zdôvodnenie a v čom spočíva? Ako je zrejmé, prvá otázka sa týka motivácií pre zavedenie podmienky zdôvodne-

kiaľ niektoré presvedčenie fakticky pochádza od karteziánskeho zlého démona, tak je nezdôvodnené (lebo presvedčenia produkované týmto zdrojom nevykazujú tendenciu k pravdivosti), ale pokiaľ presvedčenie fakticky pochádza z dobre fungujúcej zmyslovej percepcie, tak je zdôvodnené. Iné je to v rovine identifikovania zdôvodnenia, v ktorej nám reliabilizmus neposkytuje žiadny dobrý „návod", ako by sa z hľadiska subjektu dalo diferencovať medzi uvedenými dvomi možnosťami, čo vedie k stavu, ktorý Picha výstižne nazýva *reliabilistický skepticizmus* (pozri Picha 2013, 382).

nia a jej úlohy pri opisovaní poznania, zatiaľ čo predmetom druhej otázky sú nevyhnutné a postačujúce podmienky, za akých nejaké presvedčenie nadobúda status zdôvodnenosti. Z hľadiska prvej otázky nemám výhrady proti Taligovmu upozorneniu, že zdôvodňovanie s iným ako epistemickým cieľom, t. j. s cieľom maximalizovať pravdivé presvedčenia, nie je pre epistemológiu zaujímavé – naopak, plne sa s ním stotožňujem. Z hľadiska druhej otázky sa však dostávajú do popredia iné črty. Začnem (pomerne banálnym) pripomenutím, že v rámci tradičnej tripartitnej definície poznania sú podmienka zdôvodnenia a podmienka pravdivosti nominálne dve rôzne, samostatne stojace podmienky. Okrem iného je to vidno aj v tom, že táto definícia pripúšťa, aby subjekt mal zdôvodnené, no nepravdivé presvedčenie, alebo opačne, pravdivé, a pritom nezdôvodnené presvedčenie.[4] Toto pripomenutie ďalej upriamuje našu pozornosť na inú dôležitú charakteristiku: to, *čomu* pripisujeme zdôvodnenie (alebo aj pravdivosť), je *presvedčenie*. Inak povedané, pokiaľ sa pýtame, *čo* je nositeľom vlastnosti „byť zdôvodnený" (alebo „byť pravdivý"), tak prirodzenou odpoveďou je, že je to prvok *presvedčenia*. Nezdôvodňujeme pravdivosť, ako by sa niektorí mohli nazdávať; vlastnosť „byť zdôvodnený" nepripisujeme pravdivosti niečoho. Aj z hľadiska klasického delenia medzi *episteme* a *doxa* stojí na počiatku skúmania mienka, presvedčenie, o ktorom následne zisťujeme, či je pravdivé, resp. zdôvodnené, aby mohlo byť poznaním (teda či má alebo nemá takéto vlastnosti). Oddelenosť podmienok zdôvodnenia a pravdivosti nakoniec dokumentuje i fakt, že pri vysvetľovaní, za akých okolností je nejaké presvedčenie zdôvodnené, sa v rámci teórií zdôvodnenia nepoužíva priame odvolávanie sa na faktor pravdivosti.[5] Toto tvrdenie sa síce môže zdať podozrivé vo svetle vymedzenia externalistického zdôvodnenia ako *sklonu k pravdivosti* presvedčení v závislosti od ich pôvodu, ktoré uvádzam v Nuhlíček (2013a,

[4] Dodávam, že v obidvoch prípadoch by nešlo o príklady poznania, nakoľko by neboli splnené *všetky* podmienky, ktoré tripartitná definícia kladie ako nevyhnutné pre poznanie.

[5] Predstavme si „rozvinutú" tripartitnú definíciu poznania ako konjunkciu výrokov, ktorá by vznikla, keby sme v tejto definícii za výraz „je pravdivé" dosadili súbor podmienok špecifikujúcich, kedy je nejaké presvedčenie pravdivé, a za výraz „je zdôvodnené" analogicky dosadili súbor podmienok zdôvodnenia nejakého presvedčenia (v obidvoch prípadoch v závislosti od zvolenej teórie pravdivosti resp. teórie zdôvodnenia). Pokiaľ by sme zdôvodnenie charakterizovali v termínoch pravdivosti, alebo by sme nejakou okľukou redukovali zdôvodnenie na pravdivosť, tak príslušná časť našej definície by sa v súvislosti so zdôvodnením mohla stať kruhovou alebo redundantnou.

147), a s ktorým Taliga polemizuje v Taliga (2013b, 387), no pri bližšom pohľade sa neukazuje žiadny spor ani nekonzistentnosť: vlastnosť „mať sklon k pravdivosti" predsa nie je totožná s vlastnosťou „byť pravdivý" (nie sú extenzívne ekvivalentné). Uvedené externalistické vymedzenie zdôvodnenia teda neopisuje zdôvodnenie priamo v termínoch pravdivosti, čoho dôkazom je opäť možnosť mať zdôvodnené (t. j. majúce sklon k pravdivosti), no v skutočnosti nepravdivé presvedčenie, ktorú takéto chápanie pripúšťa.[6] Cieľom predošlých poznámok je ilustrovanie tézy, že z hľadiska druhej otázky položenej v úvode tohto odseku požiadavka pravdivosti presvedčenia priamo nevstupuje do explikácie jeho zdôvodnenia. To nijako nepodkopáva názor, že zdôvodňujeme s cieľom dosahovať maximum pravdivých presvedčení – iba pritom nesmieme strácať zo zreteľa jednoduchú diferenciu, že oná pravdivosť nie je súčasťou teoretického vymedzenia zdôvodnenia. V tomto zmysle treba čítať moje vyjadrenia, že to, čo zdôvodňujeme, je presvedčenie samé a nie jeho pravdivosť, ako aj to, že podmienka zdôvodnenia (a jej explikácia) je nezávislá od podmienky pravdivosti.

V súvislosti so zdôvodnením však M. Taliga opakovane obhajoval názor, že podľa neho nadišiel čas *úplne sa vzdať* požiadavky zdôvodňovania poznatkov (pozri najmä Taliga 2009). Tejto témy sa dotkol aj v Taliga (2013b, 392 a n.) a ja mu ďakujem za vysvetlenie jeho postoja. Pri tejto téme sa ešte na chvíľu zastavím aj ja, pretože sa mi zdá, že sa mi nepodarilo uspokojivo objasniť jednu okolnosť. Taliga píše, že podľa jeho mienky napokon jeho a moje „názory nie sú až také odlišné" (Taliga 2013b, 393), nakoľko on odmieta zdôvodnenie ako také a ja sa (najmä v Nuhlíček 2013b) dostávam na pohľad k podobnému záveru, že existujúce teórie zdôvodnenia, či už internalistické alebo externalistické, sú neprínosné. Napriek tomu, že naše stanoviská môžu vyzerať príbuzne, vidím medzi nimi jeden podstatný rozdiel: ja *netvrdím*, tak ako M. Taliga, *že zdôvodňovanie treba paušálne odmietnuť*. Taliga na viacerých miestach (systematicky v spomenutej stati Taliga 2009) rozoberá možné problémy a prekážky, z toho mnohé opodstatnené, do ktorých sa môže zaplietať teoretická požiadavka zdôvodňovania. Proti tomu nenamietam; zrejme sotva dnes nájdeme teóriu zdôvodnenia, ktorú by

[6] Napríklad v prípade reliabilizmu by subjekt mohol mať presvedčenie, ktoré pochádza zo spoľahlivého kognitívneho zdroja, t. j. zo zdroja, ktorý vo všeobecnosti produkuje prevahu pravdivých presvedčení, vďaka čomu by jeho presvedčenie vykazovalo sklon k pravdivosti. To však ešte nie je zárukou, že dané presvedčenie by bolo skutočne pravdivé – sami reliabilisti pripúšťajú, že aj spoľahlivé kognitívne zdroje nás môžu občas zavádzať (napr. percepcia vykonávaná v šere, hmle, či iných neprajných okolnostiach).

sme mohli bez váhania označiť ako jednoznačne úspešnú a všeobecne prijímanú. Myslím si však, že z toho ešte *nevyplýva*, že sama požiadavka zdôvodnenia je pomýlená a treba ju eliminovať. Preto nesúhlasím s Taligovým záverom: „Odmietnutie zdôvodnenia vedie spolu s faktom, že existuje poznanie, k téze, že poznatky nemusia byť (a ani nie sú) zdôvodnené" (Taliga 2013b, 394) – toto tvrdenie považujem za gro nezhody medzi ním a mnou v otázke zdôvodnenia.

Zastávam názor, že zdôvodnenie nie je umelo vznesená požiadavka na poznanie; vychádza z vysoko plauzibilných epistemických intuícií, podľa ktorých vedenie zahŕňa (alebo by malo zahŕňať) schopnosť vhodne zdôvodniť dané vedenie. Pokiaľ by niekto tvrdil, že niečo vie, ale nedokázal by ďalej uviesť prečo a ako to vie (čiže uviesť dôvody pre svoje vedenie), tak by sme zrejme mali tendenciu upierať mu, že to skutočne vie. Takéto úvahy sú v epistemológii dlho a dobre známe (jedným z najstarších, a pritom stále najlepších textov na túto tému je Ayer 1956, 28-34). Je pravda, že mnohé pokusy teoreticky rozpracovať problém zdôvodnenia narazili na nečakané ťažkosti, ale paušálne odmietnutie zdôvodňovania kvôli tomu sa mi vidí príliš unáhlené. Vypustenie tejto požiadavky by podľa mňa spôsobilo vážny nesúlad medzi teoretickým opisom ľudského poznania a bežnými epistemickými praktikami a intuíciami. Netvrdím, že epistemologické teórie musia otrocky nasledovať všetky bežné intuície, ale nazdávam sa, že by mali byť v súlade aspoň s hlavnými intuitívnymi postojmi, medzi ktoré zaraďujem aj tézu, že poznanie zahŕňa zdôvodnenie. Trvanie na stanovisku „poznania bez zdôvodnenia" je z hľadiska Taligovej argumentácie síce pochopiteľné, ale jeho neintuitívnosť je, myslím, privysokou cenou za jeho zdanlivú teoretickú výhodnosť.[7] M. Taliga mi môže adresovať výzvu, nech mu teda ukážem nejakú funkčnú a neproblematickú teóriu zdôvodnenia – pokiaľ mi takúto výzvu adresuje, budem nútený priznať, že žiadnu takú teóriu zdôvodnenia nepoznám. No ešte raz prízvukujem: 1. z toho, že sa doposiaľ neobjavila prijateľná teória zdôvodnenia, logicky nevyplýva, že požiadavka zdôvodnenia je ako taká neopodstatnená;[8] 2. prípadné odmietnutie požiadavky zdôvod-

[7] V tejto súvislosti mi napadajú známe slová T. Reida, že v prípade konfliktu medzi filozofiou a zdravým rozumom je to zákonite filozofia, ktorá musí ustúpiť.

[8] Nech mi M. Taliga odpustí, ak sa mýlim, ale javí sa mi, že napriek jeho dobrej znalosti problému indukcie sám robí induktívny záver spôsobom, ktorý inak odmieta. Konkrétne z faktu, že doteraz nenarazil na použiteľnú teóriu zdôvodnenia, zdá sa, induktívne vyvodzuje záver, že použiteľná teória zdôvodnenia vo všeobecnosti neexistuje.

nenia by nás dostalo do prílišného sporu s bežnými epistemickými intuíciami. Vyzerá to ako patová situácia, ale obávam sa, že jej alternatívy sú ešte menej uspokojivé.

Zvyšný priestor textu využijem na to, aby som sa ešte raz dotkol názorovej nezhody medzi M. Taligom a mnou, ktorá sa týka chápania skepticizmu a pokúsil sa stručne objasniť aspoň dve okolnosti, ktoré sa mi v tejto diskusii zdajú významné. Po prvé, Taliga akoby stále nerozumel, čo je v neporiadku s jeho vymedzením skepticizmu ako pozície, podľa ktorej explicitne *máme poznanie*. Som presvedčený, že ktokoľvek, kto aspoň laicky rozumie výrazu „skeptik", ho spája s pochybovaním; a pokiaľ sa pohybujeme v rovine epistemológie, tak ho celkom prirodzene spája s pochybovaním o poznaní. A skutočne, takýto pohľad na skepticizmus drvivo prevláda jednak v dejinách filozofie, a jednak aj v súčasných diskusiách o tejto problematike. Taliga však prichádza s opačným názorom a uvádza, že „skeptik *nemá prečo* tvrdiť, že nič nevieme, a ani spochybňovať samu skutočnosť, že nejaké poznanie predsa len máme" (Taliga 2013b, 394; kurzíva M. T.). Myslím, že tu sa ukazuje jadro nedorozumenia: Podľa môjho názoru skeptik nie je niekto, u koho skúmame, či má alebo nemá prečo tvrdiť to či ono. Naopak – každú pozíciu alebo argumentáciu, ktorá smeruje k spochybneniu poznania, z *definície* označujeme ako skeptickú a pod „skeptikom" chápeme (často fiktívnu) postavu, o ktorej si predstavujeme, že zastáva skeptickú pozíciu alebo predkladá skeptickú argumentáciu. Z môjho pohľadu teda nemôžem súhlasiť s názorom, podľa ktorého skeptik pripúšťa, či dokonca obhajuje poznanie. Prirodzene, Taligovi nič nebráni rozvíjať pozíciu, ktorú *on* nazýva skepticizmom, ak mu je sympatická, pričom z vecného hľadiska nemám pripomienky k zrozumiteľnosti jej opisu, ktorý uvádza (najflagrantnejšie v Taliga 2010). Pokiaľ sa ju však rozhodne naďalej označovať, v rozpore s tradíciou, ako skepticizmus, tak zrejme musí počítať s tým, že opakovane môže narážať na začudovanie a nepochopenie, tak ako tomu bolo z mojej strany v priebehu predchádzajúcich diskusií.

Po druhé, chcem spresniť tvrdenie, že skepticizmus nie je a nemusí byť logicky konzistentný smer (prvýkrát uvedené v Nuhlíček – Szapuová 2012, 151), ktoré, uznávam, môže vyvolávať isté rozpaky. Týmto tvrdením nemám na mysli, že nejaká konkrétna skeptická argumentácia (smerujúca k spochybneniu poznania) by mohla byť vnútorne logicky protirečivá, ale že rozličné skeptické argumentácie nemusia pri ich vzájomnom porovnaní vykazovať dokonalú zlučiteľnosť. Skepticizmus nemusí byť logicky jednoliaty v tom zmysle, že pripúšťa existenciu rôznorodých, a navzájom nie nevy-

hnutne zlučiteľných argumentačných stratégií na spochybnenie poznania, pričom tieto stratégie spája spoločný cieľ (vďaka čomu sa nazývajú skeptické!), no môžu sa rozchádzať v cestách, ako k nemu dospieť. M. Taliga predsa sám hovorí, že skeptik spochybňuje poznanie vymedzené určitým konkrétnym spôsobom (pozri Taliga 2013b, 394). Neznamená to potom, že existujú rozličné „skepticizmy", ktoré útočia na rôzne vymedzenia poznania, t. j. postupujú k tomu istému cieľu, ale rôznymi cestami?[9]

Literatúra

AYER, A. (1956): *The Problem of Kowledge*. London: MacMillan.
DÉMUTH, A. (2012): Môže skeptik skutočne pochybovať o všetkom? (Odpoveď M. Taligovi). *Filozofia* 67, č. 2, 141-146.
GAHÉR, F. (2013): Je zdôvodňovanie naozaj zbytočné? *Filozofia* 68, č. 2, 132-138.
GETTIER, E. (1963): Is Justified True Belief Knowledge? *Analysis* 23, No. 6, 121-123.
GOLDMAN, A. (1967): A Causal Theory of Knowing. *The Journal of Philosophy* 64, No. 12, 357-372.
NUHLÍČEK, M. – SZAPUOVÁ, M. (2012): Poznámky ku skepticizmu alebo Čo spochybňuje, a čo nespochybňuje skeptik. *Filozofia* 67, č. 2, 147-159.
NUHLÍČEK, M. (2013a): Čo je zdôvodnenie? (Odpoveď M. Taligovi). *Filozofia* 68, č. 2, 139-150.
NUHLÍČEK, M. (2013b): Prečo externalistické kritériá zdôvodnenia principiálne zlyhávajú? *Organon F* 20, č. 1, 37-49.
PICHA, M. (2013): Externalismus a skeptická výzva. *Organon F* 20, č. 3, 379-383.
SZAPUOVÁ, M. (2013): K podmienkam zmysluplnej filozofickej polemiky. *Filozofia* 68, č. 2, 151-156.
TALIGA, M. (2009): Nekonečný príbeh zdôvodňovania. *Filosofický časopis* 57, č. 3, 353-374.
TALIGA, M. (2010): Paradox skepticizmu? *Filozofia* 65, č. 7, 695-705.
TALIGA, M. (2012): Na čo sú dobré argumenty? (Odpoveď A. Démuthovi, M. Nuhlíčkovi a M. Szapuovej). *Filozofia* 67, č. 5, 417-425.
TALIGA, M. (2013a): O nepotrebnosti podmienky zdôvodnenia pre zmysluplnú filozofickú kritiku. (Odpoveď F. Gahérovi a M. Szapuovej). *Filozofia* 68, č. 7, 606-614.
TALIGA, M. (2013b): Pravdivosť alebo zdôvodnenie? (Odpoveď M. Nuhlíčkovi). *Organon F* 20, č. 3, 384-398.

[9] Pomocou takýchto úvah by azda bolo možné tiež nanovo zachytiť rozdiel medzi skepticizmom a agnosticizmom, ktorého jasná formulácia leží na srdci Taligovi: Špecifikom skepticizmu by mohol byť práve jeho charakter rôznorodého súboru úvah, argumentov a dôvodov mieriacich na spoločný cieľ – na spochybnenie poznania; prípadne by mohol byť vymedzený ako pozícia, ktorú dosiahneme aplikovaním takýchto úvah, argumentov a dôvodov.

Organon F 21 (3) 2014: 409-420

Marek Picha, Dagmar Pichová: *100 myšlenkových experimentů ve filozofii*
Praha: dybbuk 2013, 191 strán

Jeden zo spôsobov, ktorým môžeme bežnému človeku priblížiť druh činnosti, ktorej sa filozofi bežne vo svojom čase venujú, môže spočívať v tom, že takejto osobe uvedieme príklad typického filozofického argumentu alebo ... určitého filozofického myšlienkového experimentu. Nedávny knižný produkt Mareka Pichu a Dagmar Pichovej prináša veľa zaujímavých myšlienkových experimentov z bohatého spektra filozofických disciplín – počnúc metafyzikou, epistemológiou, či etikou, pokračujúc teóriou konania, metodológiou vedy, filozofiou náboženstva, filozofiou mysle, a končiac filozofiou umenia, či sémantikou. Autori v tejto práci približujú súbor myšlienkových „dobrodružstiev", ktoré majú potenciál zaujať širšiu čitateľskú obec, nielen tú filozoficky zainteresovanú.

Nie je to tak dávno, čo jeden zo spoluatorov, Marek Picha, prišiel na knižný trh s vydarenou publikáciou *Kdyby chyby: epistemologie myšlenkových experimentů* (pozri Picha 2011). Išlo o prácu, v ktorej venoval priestor širšiemu spektru otázok súvisiacich s definičným vymedzením myšlienkových experimentov, ich klasifikáciou či s ich epistemickým posudzovaním, resp. hodnotením. Nová kniha *100 myšlenkových experimentů ve filozofii* z pera manželov Pichovcov predstavuje nielen vhodný a prakticky orientovaný doplnok predchádzajúcej práce, ale aj samostatnú zbierku príkladov, doplnených o stručné uvedenie do tejto filozofickej metódy. Ako autori už v Predhovore deklarujú, ich kniha je určená „širokému publiku".

Kniha má okrem Predhovoru a stovky príkladov filozofických myšlienkových experimentov aj rozsiahlejší úvod, v ktorom sa autori pokúšajú vymedziť aspoň niektoré základné charakteristiky ich prístupu k myšlienkovým experimentom a princípy, ktoré stoja v pozadí ich analýz. Knihu navyše doplňajú hneď tri druhy registrov – tematický, menný (autorský) i názvový – čo výrazne uľahčuje vyhľadávanie myšlienkových experimentov podľa preferovaného kľúča. Jednotlivé experimenty však v knihe nasledujú v abecednom poradí podľa mien ich autorov – počnúc Akvinského „Kanibalmi" až po Ziffov „Zreteľný príklad". Každý jeden myšlienkový experiment je vtesnaný na jednu až dve strany. Expozícia predstavených experimentov zahŕňa jednak stručný úryvok z diela daného filozofa spolu s úplným bibliografickým odkazom na prácu, v ktorej sa myšlienkový experiment vyskytol, jednak jeho rekonštrukciu podľa schémy, ktorú Dagmar a Marek Pichovci v úvode vysvetľujú.

Autori knihy *100 myšlenkových experimentů ve filozofii* myšlienkové experimenty vymedzujú ako „*soubory pokynů určujících, co si představit, chceme-li něco zjistit*“ (s. 12). Aj keď uvedené vymedzenie ďalej charakterizujú a špecifikujú, a teda vysvetľujú, akú úlohu v tomto druhu filozofickej činnosti (a jej produktov) zohráva predstavivosť a zameranosť na cieľ, predsa len sa zdá byť toto ich vymedzenie širšie, než zrejme sami zamýšľali. Keď totiž neskôr špecifikujú, ktorý druh myšlienkových činností a ich produktov *ne*budú považovať za myšlienkové experimenty (pozri s. 15-23), nájdeme na ich zozname tri kategórie entít: *všeobecné argumenty*, *reálne psychologické experimenty* a *úvahy o predstavivosti*. V prípade niektorých všeobecných argumentov, medzi ktoré zaraďujú napríklad Pascalovu slávnu stávku o existencii Boha alebo aj Zenónovu apóriu letiaceho šípu, ale aj v prípade niektorých úvah o predstavivosti (napríklad Nagelovej filozofickej úvahy o tom, čo by znamenalo mať skúsenosť ako netopier), by tieto príklady spĺňali ich (definičné) vymedzenie. Ich dodatočné spresnenia a ilustrácie príkladov však majú naznačiť, že status myšlienkových experimentov majú mať len tie úvahy, v ktorých sa (i) vyžaduje vybavenie si konkrétnych detailov či scén predstavenej situácie; a (ii) cieľom predstavenia si určitej situácie je zistenie týkajúce sa tejto situácie, nie našej predstavivosti.

V tejto súvislosti sa musím priznať, že ma autori príliš nepresvedčili. Nejde v žiadnom prípade o to, že by som nerešpektoval ich právo zvoliť si určité obmedzenia, umožňujúce vyselektovanie niektorých filozofických úvah a argumentov ako myšlienkových experimentov. Skôr ide o to, že ich obmedzujúce podmienky sú príliš vágne. Napríklad pokiaľ ide o prvé obmedzenie, môžeme povedať, že a) niektoré konkrétne detaily predstavovanej situácie sú sekundárne z hľadiska informácie, ktorú má myšlienkový experiment sprostredkovať (čoho sú si napokon aj oni sami vedomí – pozri napríklad Thomsonovej príklad Problému transplantácie, ktorý autori analyzujú v úvodnej časti: s. 27-28); ďalej, že b) rozdiel medzi konkrétnymi detailmi opisu predstavovanej situácie a nejakým všeobecnejším opisom (všeobecným argumentom) je skôr mierou stupňa, a nie kvalitatívneho rozdielu. Navyše, pokiaľ ide o druhé špecifické obmedzenie, môžeme si všimnúť, že c) zistenia vzťahujúce sa na našu predstavivosť, a s ňou súvisiace presvedčenia, môžu byť v určitých prípadoch aj zaujímavou informáciou o svete, ktorého súčasťou sú okrem iného aj naše predstavy a spôsoby ich generovania či ďalšieho spracovania.

Nejde teda o to, že by sme odmietali výber predstavených myšlienkových situácií na základe nejakých kritérií, alebo že by sme videli problém v nešpecifikovanej selektívnosti; skôr tu narážame na problém veľmi hrubých kritérií, ktoré umožňujú interpretovať ako myšlienkové experimenty aj také úvahy, ktoré autori za také považovať nechcú.

Pokiaľ však ide o návrh rekonštruovať myšlienkové experimenty podľa schémy, ktorú autori prezentujú (a ktorá bola predstavená už v práci Picha 2011), môžeme oceniť ich návrh z hľadiska vyváženosti kombinácie ideálov jednoduchosti, zrozumiteľnosti i transparentnosti. Autori v úvode práce približujú trojstupňovú analýzu jednotlivých príkladov.

V prvom kroku analýzy formulujú tzv. *destilát* myšlienkového experimentu. Ide o predstavenie hypotetickej (prípadne už realizovanej) situácie, v ktorej sa menia okolnosti *obvyklého priebehu* i *upraveného priebehu*. Obvyklý priebeh je charakterizovaný pomocou dvojice (pojmov) *vstupnej konštelácie* a *výstupnej konštelácie*, zatiaľ čo upravený priebeh pomocou dvojice (pojmov) *upravenej konštelácie* a *záveru*. Aby sme si uvedené rozlíšenia ilustrovali, pozrime sa na destilát myšlienkového experimentu Descartesovho Klamúceho démona, ktorý (na s. 51) autori približujú týmito vymedzeniami:

Obvyklý priebeh predstavenej situácie, ktorú Descartes opisuje vo svojich *Meditáciách o prvej filozofii*, reprezentuje v ich priblížení vstupná konštelácia „*menej mocný klamúci démon*“ a výstupná konštelácia charakterizovaná spojením „*nemôže klamať o niektorých mentálnych obsahoch*“. Upravená konštelácia upraveného priebehu je zase vyjadrená podmienkou „*veľmi mocný klamúci démon*“ a záver vyjadruje spojenie „*nemôže klamať o existencii klamaného subjektu*“. Obvyklý priebeh tak reprezentuje dvojica konštelácií *menej mocný klamúci démon – nemôže klamať o niektorých mentálnych obsahoch*, a upravený priebeh zase dvojica *veľmi mocný klamúci démon – nemôže klamať o existencii klamaného subjektu*.

Problémom tohto konkrétneho príkladu je však to, že úryvok z Descartesa, ktorý mu predchádza, neobsahuje časť textu, z ktorej by sme mohli „vydestilovať“ východiskovú situáciu s tzv. menej mocným klamúcim démonom. Jediná pomyselná situácia, na ktorú Descartesov úryvok odkazuje, je situácia s démonom, ktorý je „nesmírně mocný a lstivý“, teda situácia odkazujúca na tzv. upravený priebeh príbehu.

Ak by sme si zvolili slovník teórie množín (čo autori práce nerobia), tak vstupnú a výstupnú konšteláciu obvyklého priebehu (určitej situácie), ako aj upravenú konšteláciu a záver upraveného priebehu (situácie) by sme mohli označiť za určitú konceptuálnu (myšlienkovú) funkciu, kde argumentu funkcie zodpovedá jej vstupná konštelácia, resp. upravená konštelácia (v prípade upraveného priebehu) a hodnote funkcie zase výstupná konštelácia, resp. záver (v prípade upraveného priebehu). Ak by sme to trochu zjednodušili, mohli by sme povedať, že jedným z cieľov myšlienkového experimentu môže byť overenie toho, či ide o funkciu s konštantným priebehom hodnôt pre rôzne argumenty.

Druhý krok analýzy, ktorú pri rekonštrukcii myšlienkových experimentov autori ponúkajú, predstavuje formulácia *princípu*, ktorý má reprezentovať „zobecnění závěru získaného promyšlením scénaře“ (s. 29). Autori čitateľa upo-

zorňujú na úskalia tohto kroku, pretože si vyžaduje určité zovšeobecnenie, ktoré nemusí byť explicitne vyjadrené v pôvodnom zdroji experimentu a pripúšťa viacero „stupňov" všeobecnosti. V prípade Descartesovho Klamúceho démona má princíp túto podobu: „Nejmocnější klamavý démon by nemohl klamat o všem" (s. 51). Ide o princíp, ktorý zovšeobecňuje informáciu obsiahnutú v danom myšlienkovom experimente – že totiž ak by aj naše poznávacie schopnosti boli vystavené klamaniu zo strany Descartesovho démona, niektoré naše presvedčenia by boli voči tomu rezistentné.

Napokon, posledný, tretí krok predstavuje vyjadrenie určitej filozofickej *tézy*, ktorú má daná myšlienková úvaha *podporiť alebo vyvrátiť*. Ide o krok, ktorý zachytáva filozofický účel daného experimentu. Myšlienkový experiment tak má potenciál určitú filozofickú tézu buď pozitívne podporiť (zdôvodniť) alebo spochybniť (vyvrátiť). Napríklad, tézu Descartesovho myšlienkového experimentu s klamúcim démonom autori vyjadrujú tvrdením „Skepticizmus je chybný" (s. 51). Keď to zhrnieme, môžeme spolu s autormi povedať, že každý myšlienkový experiment predstavuje nejakú *prípadovú štúdiu* (s. 23), ktorá má podporiť určité filozofické tvrdenie.

Keďže autori ešte v úvode deklarujú, že ich „analýzy jsou míněny pouze jako vodítka, jako interpretační pomůcky", a tiež že „všechny rozbory jsou samozřejmě revidovatelné ..." (s. 32), ťažko im možno určité zjednodušenia v ich analýzach vyčítať. Ak by im totiž išlo o ambíciu detailne priblížiť jednotlivé myšlienkové experimenty spolu s ich širším filozofickým pozadím a diskusiou, táto kniha by takýto cieľ nenaplnila. Keďže však zámerom autorov bolo priblíženie základných zdrojov a schém množstva zaujímavých filozofických experimentov, potom môžeme konštatovať, že tento zámer sa v uvedenej publikácii podarilo naplniť. Je síce pravdou, že v prípade niektorých hypotetických experimentov by som privítal uvedenie širšieho textového (filozofického) kontextu (napríklad v prípade Frankfurtovho „Dobrovoľne závislého" či Parfitovho „Napraveného nobelistu"), no o väčšine vybraných myšlienkových experimentov možno povedať, že sa im ich podarilo zrozumiteľne a jednoducho priblížiť.

V práci tak možno nájsť exemplárne formulácie viacerých známych i menej známych filozofických experimentov, a to ako z antickej, stredovekej i novovekej filozofickej tradície, tak aj z prostredia analytickej filozofie 20. storočia. Nájdeme v nej rovnako Zenónovho Achilla a korytnačku, ako aj Gaunillov Stratený ostrov, Lockovho Dobrovoľného väzňa, Kantov Apriórny priestor, Russellovu Päťminútovú hypotézu, Quinovho Gavagaia a množstvo ďalších príkladov.

Aj keď autori zrejme nemali ambíciu, aby sa ich publikácia stala úvodom do filozofie myšlienkových experimentov, predsa len sa im podarilo predložiť veľmi cennú zásobáreň filozofických úvah a argumentov, ktoré pracujú s predstavovaním si hypotetických (myšlienkových) experimentov. Ich práca môže slúžiť

ako vhodná didaktická pomôcka na vyhľadávanie a rekonštrukciu mnohých myšlienkových experimentov, predovšetkým v úvodných kurzoch do filozofickej metodológie, teórie argumentácie, či filozofie myšlienkových experimentov. Na druhej strane, aj širšia čitateľská obec dostáva do rúk prácu, ktorá ju môže príjemne prekvapiť i pobaviť príkladmi úvah, ktorými filozofi vypĺňajú svoj čas.

Lukáš Bielik
bielikluc@yahoo.com

Literatúra

Picha, M. (2011): *Kdyby chyby: epistemologie myšlenkových experimentů*. Nakladatelství Olomouc.

David Shoemaker: *Personal Identity and Ethics: A Brief Introduction*
Ontario: Broadview Press 2009, 296 pages

In his book *Personal Identity and Ethics – A Brief Introduction,* David Shoemaker presents a uniquely comprehensive treatment of the relationship between personal identity and identity-related ethical concerns. The author proceeds in a clear and reader-friendly manner, starting with the definitions of the key concepts in the field, going on to particular theories of personal identity and their connections to several practical concerns, and concluding with methodological issues of a higher level of abstraction.

The general strategy of the book is to see whether certain concerns that have traditionally been taken to presuppose the concept of personal identity (such as responsibility, distributive justice, compensation, etc.) can be (and need to be) justified by a theory of personal identity, and which theory, if any, can do the job. If successful, such a theory would cohere with our intuitions about when the practical concerns are justified and propose a criterion of personal identity which would *explain why* the concerns are justified. Of course, some of our intuitions about the appropriateness of the practical concerns may have to be revised if they turn out to be inconsistent with what is in the course of the investigation identified as the theory of personal identity with the greatest explanatory power. What is sought, then, is a *reflective equilibrium* between our intuitions about identity-related practical concerns and theories of personal identity.

The book is divided into eight chapters in two parts. The first part focuses on the role of personal identity in *self-regarding* ethics, covering practical issues which include the possibility of immortality and the rationality of anticipation

and self-concern. These concepts serve Shoemaker as a background for the discussion of four basic theories of personal identity – the soul theory, the memory theory, the body theory, and the brain theory – and two more sophisticated theories – the biological theory and the psychological theory. Shoemaker concludes that the four basic theories are quite inadequate in their own terms and that the more sophisticated theories are each controversial – explaining some of our intuitions and failing in others. However, with respect to anticipation and self-concern, the psychological theory seems to be more adequate than the biological approach. The possibility of immortality is shown to be very difficult to justify on any theory.

Shoemaker then discusses two radical approaches – the narrative theory and the "identity-does-not-matter" (IDM) view, which have been designed to fix the problems of the established theories. He shows that the success of the narrative view in explaining our concerns is dubious, and, moreover, the theory is inherently vague, which makes its application especially difficult. The IDM view, in contrast, is quite plausible and Shoemaker regularly turns to it throughout the book to seek solutions to other problems discussed, because it offers a more fine-grained analysis. It is made clear, however, that by adopting the IDM view, one is forced to give up the general assumption of the whole project: if identity does not matter, the practical concerns are not justified by the logical relation of numerical identity, but by other *continuity* relations, which differ from identity.

The second part of the book focuses on the relevance of personal identity for *other-regarding* ethics. Chapters four and five present the moral issues at the beginning of life, such as abortion, stem cell research, cloning, genetic intervention and population ethics. Shoemaker reaches the following conclusions:

- In the abortion debate, the only theory that could justify the identity between a fetus and the adult is the biological theory, but that theory is morally irrelevant, so it cannot be used to justify the immorality of killing fetuses.
- The only identity-based objection to stem-cell research can come from the soul theory, which is deeply flawed.
- No plausible theory of personal identity can support objections to human cloning.
- In realistic cases of genetic intervention, such intervention does not have the potential to change an individual's identity.
- Radical cases of enhancement may threaten one's narrative identity, but they can also be interpreted in a way that retains the narrative identity of the enhanced individual (partly due to the vagueness of the narrative criterion).

- Serious implications can be shown to follow from personal identity theory for intergenerational justice. No theory, unfortunately, offers a plausible solution to the problems.

A note of caution: the fact that no objections to certain practices can be drawn from personal identity theory does by no means mean that there aren't any other legitimate objections to the practices. But these are out of the focus of Shoemaker's book.

Chapter six deals with the moral issues at the end of life: advance directives and the death of multiple personalities in dissociative identity disorder. Both issues are extremely mind-boggling and thought-provoking. Shoemaker argues that, on any theory of personal identity, it is very difficult to justify our intuition that advance directives should be respected. With respect to multiple personalities, Shoemaker defends the view that alter-egos are different persons in one body. But in that case it becomes difficult to resolve a clash in our intuitions: on the one hand, we believe that all persons deserve moral protection and should not be killed (eliminated by treatment), on the other, the doctors who treat DID do not act immorally. All in all, these two difficult issues give slightly more support for the biological theory of personal identity.

Then two extensive chapters follow, treating moral responsibility and the implications of personal identity for ethical theory. The main conclusion of the first chapter is that while responsibility entails ownership of the actions for which one is responsible, ownership does not entail identity with the agent. Rather, it is the continuity of a subset of the person's psychological make-up. The precise definition of what falls into the subset is difficult to provide, but the IDM view comes closer to truth than its opponents.

The other chapter focuses on the assessment of the implications of Parfitian reductionism in personal identity for normative theory. Shoemaker deals with Parfit's utilitarianism and Brink's rational egoism, as well as contractarian and Kantian objections to reductionism. One important claim that crystalizes in the discussions is that it is crucial to answer the normative question of what entities are the proper targets of the individual concerns. Shoemaker discusses momentary experiencers, selves and persons. Each alternative entails a different conception of normative ethics.

The conclusion assesses the correct methodology to adopt when dealing with issues on the border of ethics and personal identity. Shoemaker's book is based on the assumption that the practical concerns derive their justification from a metaphysical theory of personal identity. This assumption has, however, been challenged. Some authors support an "ethics first" approach, claiming that metaphysics is irrelevant, because the practical concerns derive their justifica-

tion from practice, not theory. Shoemaker acknowledges the seriousness of this challenge, but believes that more needs to be said to show that the challenge affects all of the discussed concerns. And this brings us to the outcome of Shoemaker's work.

Shoemaker reaches what might at first sight be a surprising conclusion. Our intuitions about the individual practical concerns are not uniform enough to be explained by a single theory of personal identity. Thus, some require the psychological theory (anticipation) some seem to track biological continuity (compensation) while others cannot be explained by a theory of identity at all (responsibility). One may soften the impact of the conclusion, however, by the observation that the set of identity-related practical concerns is, in fact, a relatively heterogeneous class of concerns, which have traditionally been unified merely by the relatively superficial belief that they presuppose personal identity in time. However, once we start looking more closely at what such identity may consist in, it becomes clear that there is a number of intertwined relations, which normally go together, but can be conceptually distinguished, and that the concerns actually only attach to these more elementary relations. This, I believe, is the most general outcome of the book.

But by saying that, I do not mean to imply that other important conclusions have not been reached. In fact, every chapter contains a number of conclusions that are well-supported by clearly presented arguments, and where no decisive conclusion has been reached, the reader is always shown why one is so hard to achieve. All in all, I believe the book is an extremely useful tool for anyone who would like to map the enormously rich field between ethics and personal identity, as well as advanced students in the field who will benefit from Shoemaker's insight.

I what follows, I would like to address briefly some issues that came to my mind while reading the book. I am fully aware that the scope of an introductory book did not allow for their thorough discussion.

The metaphysical and epistemological criteria of personal identity

The first one concerns the relationship between the metaphysical and epistemological criteria of personal identity. In the introduction Shoemaker makes the distinction between a metaphysical criterion specifying what personal identity consists in, and an epistemological criterion providing a way of identifying personal identity. For instance, the soul theory may provide an answer to the question of what the identity between a person identified at an earlier time and a person identified at a later time consists in. But it will hardly serve as a useful epistemological criterion, because souls are usually thought of as immaterial

and independent of the material aspects of human beings. Shoemaker says (p. 15) that his main objective is to find a metaphysical criterion, but he admits that an acceptable metaphysical criterion would lose some points if it didn't fare well on the epistemological side. But throughout the book it soon becomes clear that the two criteria are actually tied much more closely than it seems. If we are interested in personal identity because we want to justify practical concerns, we have to be able to determine whether the metaphysical criterion of personal identity holds, that is, we have to have epistemic access to the metaphysical facts in which personal identity consists.

So we must inevitably ask the question: Which metaphysical criterion is epistemologically most successful? We have already seen that epistemic access to souls is impossible, so the soul theory fails. In fact, this was pointed out by Locke, who, driven by the motivation to construct a theory of personal identity that could justify accountability, developed the memory theory. He claimed we do have knowledge of our own persistence, but if our identity resided in souls, it would be impossible. Instead, sameness of consciousness, which is usually interpreted as memory connectedness, is what enables our epistemic access to our identity. How do I know that I am the same person who got into bed last night? I don't need to look for a soul, I don't even need to look in the mirror to check the sameness of body, I simply remember from the inside the experience of lying into my bed and all other experiences that I had the previous evening. So memory connections are a plausible, albeit fallible, tool for first person identification.

However, this theory quickly runs into trouble as a metaphysical criterion and needs drastic revisions. The most important is the shift from memory connectedness to memory continuity and further to the richer relation of psychological continuity. The shift is caused by the desire to use memory as a criterion of numerical identity, which is transitive, and by the contingent fact that people forget, which makes memory intransitive. To fix this, philosophers have suggested that chains of memories be used instead of direct memories as a criterion of identity. So even though I may not remember any experiences from my teen age when I am old and forgetful, I will be identical to the teenager, because I will remember times in which I remembered the teen-age experiences. This fixes the logical and metaphysical problem, but seems to introduce epistemological difficulties. Suppose that I remember my life in 1980, but not my earlier life in 1960. But I did remember it in 1980. Since I remember 1980, I can be introspectively sure that I lived in 1980. But what good is the fact that in 1980 I remembered 1960 to my current knowledge of my life in 1960? The mental time travel that is enabled by direct memories stops in 1980. The memory of me living in 1980 does not carry any information about me

remembering at that time my life in 1960. So even though the memory continuity theory may be more suitable as a metaphysical criterion of personal identity, as an epistemological criterion it faces a problem that the memory connectedness criterion is immune to.

Many authors reject even the memory continuity theory, and instead of a single psychological relation they propose a number of such, including the intention-future experience relation, sameness of beliefs over time, and similarity of character. A sufficient number of such relations is termed *strong psychological connectedness*, and their chain is termed *psychological continuity*. The psychological continuity theory is generally regarded as one of the most successful ones, and Shoemaker concludes that one of its advantages is that it accounts for self-identification very well (p. 84). But I think the only relation that enables introspective self-identification is memory connectedness and all the relations that authors have added to it to fix its metaphysical problems presuppose memory in our introspective effort to identify and re-identify them. For instance, the only evidence that I have the same intentions and character as I had a week ago is that I remember them. Thus, memory connectedness is the key to our self-identification, but it cannot serve as a metaphysical criterion of personal identity. This creates a problem for the further debate of practical concerns, because if we agree that they are not justified solely by memory connectedness, we are abandoning the safest epistemological criterion.

Anticipation and self-concern

Another issue I would like to discuss is the advantage that the psychological theory is supposed to have over the biological theory in explaining the rationality of anticipation and self-concern.

Shoemaker starts with the commonsense belief that a necessary condition for rational anticipation and self-concern is identity (p. 60). If we believe that identity grounds anticipation and self-concern, we must find a theory of personal identity which will be capable of doing so.

Shoemaker believes that the psychological theory does a much better job at explaining the two concerns than the biological theory (p. 64, pp. 82-83). It seems that I can only rationally anticipate the experiences of my psychological descendants. If some future person won't be connected to my current psychological stream, then it's hard to see how I could rationally anticipate his experiences. And since self-concern presupposes anticipation, how could I have that special type of concern for his well-being?

I agree with the analysis of anticipation, but it seems to me that by making anticipation a necessary condition of self-concern we lose the possibility of ac-

counting for a large number of cases in which people are concerned for their own well-being in the absence of the possibility of anticipation of their experiences.

For lack of space I will offer one such example, a version of which has actually occurred.[1] Suppose that you need to undergo a minor operation that involves full anesthetics. When you have been given the drug and fall into deep sleep, you are sexually assaulted by the surgeons, who are careful enough not to leave any signs of their behavior. Then, they routinely perform the operation and in an hour or so you wake up, not having a clue about what has been going on. I believe that everyone has a reason to be concerned that a similar incident does not happen to them. But how could this form of self-concern possibly be legitimate if the necessary condition of self-concern is not met. Under the influence of anesthetics, one has no conscious experiences, so one cannot rationally anticipate them. But if anticipation is a necessary condition of self-concern, self-concern is irrational in this case. This seems quite incorrect to me. I believe that one's well-being is not exhausted by one's experiences, and, thus, self-concern is legitimate even in the absence of any experiences. Unfortunately, I have to leave aside an outline of an explanation of what relations I believe ground self-concern, if it is not psychological continuity. I do agree with Shoemaker, however, that anticipation is a wholly psychological matter.

The ground for responsibility

Another issue concerns the grounds for moral responsibility. Shoemaker believes that biological continuity cannot justify moral responsibility. He argues that it is not sufficient, because we would find it inappropriate to blame an individual with Alzheimer's disease for the crimes of the person he used to be, in spite of their biological continuity. The cerebrum transplant thought experiment further shows that biological continuity is not even necessary. Shoemaker then analyzes responsibility in terms of some subtle psychological capacities (p. 216).

What is striking is one of the conclusions that Shoemaker comes to after he exposes the limitations of the individual theories of personal identity. He states that maybe there isn't a single criterion of responsibility. Maybe we negotiate what criteria to use depending on the particular context. So in one context we may ground our judgment by biological continuity, such as in the case of me holding responsible my comatose father for his repeated humiliations of

[1] http://www.thestar.com/news/crime/2013/11/19/anesthesiologist_dr_george_doodnaught_guilty_of_sexually_assaulting_21_female_patients.html

me in the past. In another context it may be psychological continuity, such as in the case when a drunken person goes on an anti-Semitic tirade and does not remember it later.

But one must wonder how biological continuity could do the job in the comatose father case. Shoemaker has already concluded that biological continuity is not sufficient, because certain complex mental capacities are necessary. But the comatose father does not have these mental capacities: he isn't capable of executing intentions, receptive to blame, or able to judge the fairness of me blaming him. If we agreed that he could be responsible in the absence of these capacities, why require them in other cases at all?

These are just a few of the ideas that were inspired by reading Shoemaker's rich book. I can only repeat that I recommend that anyone with an interest in the intersection of metaphysics and ethics read this book.

Radim Bělohrad
belohrad@phil.muni.cz