FILOSOFISKA NOTISER

Årgång 3, Nr 2, Augusti 2016

ISSN: 2002-0198

Hemsida: www.filosofiskanotiser.com

Moral Value Property Projectivism

Ralph W. Clark

Abstract
I describe and then raise objections to what may be referred to as the Hume/Blackburn account of a projectivist view of the nature of moral values. One of the objections is that a category mistake is present in the claim that attitudes or feelings can be projected onto the world as a factor in the constitution of value properties. In order to avoid this and other objections, "moral value property projectivism" involves the projection of properties rather than attitudes or feelings. These properties are located in individual states of human consciousness. I focus on two such properties, which I designate goodness and badness, and which I argue are most immediately projected in such fashion as to give rise to general value properties. The constitution of moral value properties requires additional steps. The view I defend is a version of naturalism and moral realism.

1. Introduction

What are moral values? Answers can be placed in two broad categories that pertain to what may be called the "location" at a fundamental level for moral values. As regards the first location, in seeking an account of moral values we look primarily to the objects (actions, persons, intentions, outcomes, etc.) about which moral judgments are made. As regards the second location, we look primarily to the subjects (human beings, either individually or collectively) who make moral judgements, or in some other way respond evaluatively to objects. The first location is problematic because the properties of objects – at least such properties as science investigates, namely natural properties – seem bereft of some essential elements, such as the action-guiding character of moral values. Yet, if objects are said – a la Moore – to possess "nonnatural" properties linked to these essential elements, but which lie outside the purview of science, then these alleged properties raise questions that most philosophers would agree have not been satisfactorily answered. Because nonnatural properties by definition elude scientific investigation but are "out there in the world," yet are causally inert, how can we gain knowledge of them?

Filosofiska Notiser, Årgång 3, Nr. 2, Augusti 2016, 3–27

The second location is problematic as regards each of the two general types of answers that belong to it – cognitivist and noncognitivist. For cognitivist answers, the most widely defended claim is that moral values are best understood in terms of morally relevant facts about people's attitudes or desires. The main problem here is that describing attitudes or desires that people happen to have – saying, for example, that a certain group of people are in possession of attitudes opposed to a particular sort of inequality – does not seem to capture fully the action-guiding character of moral value assertions since it does not explain why members of this group, or anyone else, ought to have the attitudes in question.

For noncognitivist answers, the basic claim is that discourse regarding moral values is best understood as expressing in a nondescriptive fashion morally relevant attitudes, desires, or commitments. One of the main problems here is that moral discourse seems to be truth-apt, while the nondescriptive expression of attitudes or desires is not truth-apt, or at least is not straightforwardly or robustly truth-apt.

Noncognitivist positions are standardly described as "expressivist," but this label can be misleading since some of the advocates of expressivist views deny that their views are noncognitivist. An example is Simon Blackburn, who argues that morally relevant attitudes or desires are "projected" onto that to which moral discourse is addressed. To complicate the picture even further, it is possible – indeed desirable, as I argue in this paper – to defend a projectivist position that is cognitivist but does not have the "expressivist basis" that Blackburn gives to his version of projectivism.

The major answers and types of answers to which I have just referred, the major objections to these answers, and many of the responses to the objections are well-known in the literature. I spell out some of these answers, objections, and responses in what follows, in the course of defending what I believe to be a new version of the second broad type of answer (where we look primarily to human subjects in giving an account of moral values). The position that I defend is a version of moral realism (and thus also cognitivism), understood to be the view that judgments about moral values are, or are the contents of, assertions that are true or false – or at least are not automatically excluded from being true or false. Blackburn describes his position as being an example of quasi-realism. I will refer in what follows to some of the bases for debate as to exactly which senses of "true/false" are available, respectively, to the advocates of moral quasi-realism and the advocates of moral realism.

I begin by discussing the views of Blackburn, who is the best known current advocate of projectivism. I then defend a new version of projectivism that differs in some important respects from the position of Blackburn.

1. Blackburn's Projectivism

The core of Blackburn's position derives from a view made famous by Hume. Consider the following passage from Blackburn (1981, 470):

> ... in addition to judging the states of affairs the world contains, we may react to them. We form habits; we become committed to patterns of inference; we become affected, and form desires, attitudes, and sentiments. Such a reaction is 'spread on' the world, as Hume puts it in the *Treatise*, by talking and thinking as though the world contains states of affairs answering to such reactions.

Blackburn expresses equal approval for a passage from Hume's *Enquiry* where Hume famously describes the precursor to the projectivist position of Blackburn and others as involving "... gilding or staining all natural objects with the colours, borrowed from internal sentiment ..." As these quotations make clear, the core view of Hume and Blackburn is a variation on the "second location" view that I described above, according to which, in seeking a location for moral values, we look primarily to the subjects who make moral judgments or in some other way respond evaluatively to objects. Blackburn's version of projectivism is also a version of quasi-realism, as I have already mentioned, the intent of which is to show that projectivism, which for Blackburn begins life as a version of expressivism and thus noncognitivism, can be transformed into – or, perhaps better, can be seen as — a view that is not exactly noncognitivist. Below, I say more about Blackburn's quasi-realism. Before doing that, I will begin to present my case against Blackburn's position by arguing that its Humean foundation appears to be incoherent.

The category mistake objection to Hume/Blackburn. To assert that an object – which could be a state of affairs – has moral value is (or certainly seems to be) to ascribe a property to the object, or in other words to make a claim about *the way the object is.* By contrast, to have an attitude toward, a desire for, or a sentiment about an object is *to stand in a relation to the object*. Thus, to speak of "spreading" a sentiment on the world, or "gilding or staining" an object with a sentiment, or an aspect of a sentiment, is,

obviously, to speak metaphorically, but not in a way that advances the cause of philosophical clarification or argumentation. To express an attitude, desire, or sentiment is not to make any claim about the way an object is, or to make any claim at all for that matter, but rather to use language nondescriptively to convey information about the aforementioned relationship – to induce the listener to come to understand that, for example, the speaker favors the object.

The use of metaphor to which I refer in the previous paragraph has all of the advantages of theft over honest labor since it describes one thing (an attitude, desire, or sentiment–none of which is a property) in terms of another thing (namely something that is a property, or a quality; to become gilded or stained is to be changed qualitatively) in a situation where the heart of the relevant metaethical debate concerns the very question of whether or not moral discourse concerns the possession of a moral property by, for example, a state of affairs that is said to be just. If we say that a moral property is present, then we have a solid foundation for defending moral realism; if we deny that a moral property is present but say instead that there is present only an attitude, desire, or sentiment on the part of those who respond to the state of affairs in question, then we no longer have a solid foundation for defending moral realism. What the Humean projectivist metaphor, or metaphors, do is meld together the opposing sides in the debate. This would seem to be an illicit use of metaphor since, if the metaphor is allowed to stand as conveying philosophical content, then we will have gained a way to speak of moral properties from an expressivist perspective without having given a philosophical account of how it is that expressivism allows for "moral property talk." On its face, certainly, expressivism does not allow for moral property talk.

But perhaps there is a mechanism by which an attitude, desire, or sentiment can be transformed into a property. After all, not all projectivist positions are illegitimate. Lockean secondary quality projectivism, to which the moral projectivism of Hume/Blackburn is sometimes compared, does not suffer from the problem to which I just referred. For the Lockean view, it is said that phenomenal greenness, as an example, is projected onto the leaves of a tree at which I am looking; phenomenal greenness is projected out from my consciousness. While phenomenal greenness cannot exactly be spoken of as being a property which belongs to the conscious state that is present when I look at the leaves of the tree – to describe a conscious state as being green is itself a category mistake – nevertheless, phenomenal greenness is experience-

ed by me as being a "way that something is, or appears," which is to say that phenomenal greenness is either a property or at least is property-like. Therefore, in order to say that phenomenal greenness is projected onto the leaves, we do not need to employ the objectionable metaphorical leap that I described above.

Someone may argue that a color property such as greenness is not the best type of example to use in drawing a comparison between the projectivism of Hume/Blackburn and Lockean secondary quality projectivism because such a comparison does not involve attitudes or desires. A better example is the property of being delicious. The Lockean basis for saying that an item of food is delicious has got to be an attitude or desire, or some other type of evaluative response. What else could it be? But then it must be the case that the Lockean projectivist view (which for Locke applies to all five of the senses, including taste), or a view that closely resembles the Lockean view, *is* inclusive enough for moral property projectivism to fall within its scope. Lockean secondary quality projectivism, we have noted, does not suffer from the problem that I have been discussing, a problem that I have claimed does apply to the Hume/Blackburn version of projectivism.

My response is that, while the Lockean projectivist view does apply to all five of the senses, including taste, the examples pertaining to taste that Locke himself gives, such as sweetness, are not, strictly speaking, evaluative but instead are qualitative in a fashion that at least resembles the qualitative character of greenness. Just as phenomenal greenness is experienced as being a "way that something is, or appears," so also phenomenal sweetness is experienced as being a "way that something is, or appears." Phenomenal sweetness is either a property or is property-like. Thus, in saying that phenomenal sweetness is projected onto an item of food, we do not need to employ the objectionable metaphorical leap discussed above. By contrast, the property of being delicious is evaluative; there does not appear to be a qualitative phenomenal component of it that can be directly compared to phenomenal sweetness. Instead, what we have is a positive attitude on the part of someone who likes the food in question. An appeal to Lockean projectivism does not help in explaining how this attitude, which is neither a property nor property-like, is transformed via projection into a property.

Below, I defend a version of projectivism for moral properties that *is* close enough to Lockean secondary quality projectivism that, like the Lockean view, it does not need to employ the sort of objectionable metaphorical leap that I have been discussing.

In the meantime, I should mention that one way to reply to what I am calling the "category mistake objection" to the projectivism of Hume and Blackburn is to deny that the outcome of this projectivist picture is the claim that that onto which an attitude, say, is projected thereby acquires a property, or more specifically a moral property. However, if this reply is the direction that one goes in, then the resemblance between the moral projectivism of Hume and Blackburn, on the one hand, and the secondary quality projectivism of Locke, on the other hand, becomes even more tenuous than in the above discussion. Why, then, even refer to the former as projectivism? More to the point, regardless of what we may decide to call the view of Hume and Blackburn, for Blackburn there is a reason why he needs to preserve a link to the Lockean position, namely that Blackburn wants to claim that moral assertions are truth-apt. How can an assertion such as "Act A is moral" be truth-apt if "moral" as it occurs in this assertion does not name a property? If we back away entirely from the claim that "moral" names a property, then it would seem that we have retreated to the emotivist type of expressivism of Ayer and Stevenson, which is a position that many current philosophers view as discredited, and which, in any event, is a position that Blackburn himself wishes to distinguish from his own.

The normative arbitrariness objection to Hume/Blackburn. The root error that this objection capitalizes on can be traced back at least to Hobbes' claim that the definition of "good" is "object of desire" – which places the cart before the horse. Why would anyone ever desire something unless that thing was already good, or judged (or perceived) to be good, before the act of desiring came on the scene? Similarly, why would anyone have a positive attitude toward, or positive sentiment for, an object unless that object was already good, or judged (or perceived) to be good, before the attitude or sentiment came on the scene?[1] Indeed, it is difficult to understand how a desire (or positive attitude or sentiment) would even be present in a situation where no object was present that had, or was seen as having, a value property. After all, everyone agrees that desires motivate. If I desire X, then

[1] A sophisticated version of a Hobbes-like position is to found in Gibbard (1994): "To be good is to be desirable, and a thing is desirable if desiring it is warranted.... Non-cognitivists [such as Gibbard himself]... try to explain away the appearance of non-natural properties.... The expressivist's strategy is to [say] what state of mind is expressed by ascriptions of warrant.... This state of mind ... consists in *accepting norms* that say to do something." We can ask the same question of Gibbard that we ask of Hobbes: Why accept a norm that tells us to do something unless that something is good (or leads to what is good)?

I am motivated to gain or keep X unless a stronger, conflicting motivation on my part is present. How can I be motivated to gain or keep that which is, or is perceived to be, value neutral? Perhaps I can be thus "motivated" if "motivation" is understood to be wholly dissociated from deliberation or choice, and instead is no more than observed "movement toward" an object.[2] At best, under such circumstances, an act of desiring that underlies motivation is arbitrary. But surely, not all desires are arbitrary. At worst, acts of desiring that are claimed to be wholly arbitrary are not psychologically possible; they are discriminatory responses that occur in situations where there is no recognized basis for the discrimination.

Analogously, if the moral goodness of a state of affairs is said to be derived from – rather than being a basis for – the act, or acts, of desiring that state of affairs on the part of some person or group of persons, then the ascription of such moral goodness is arbitrary since the same person or persons could just as well have desired something quite different without there having been any change in normatively relevant factors pertaining to the state of affairs in question. For example, consider the moral value that resides in a person's having a morally praiseworthy character. We want to be able to say that someone has a morally praiseworthy character because – at least in part – that person consistently desires what is morally good or right, while someone else lacks such a morally praiseworthy character because, at least in part, this other person does not consistently desire what is morally good or right. In other words, we presuppose in these situations that the possession of the property of moral goodness by that which is desired precedes the desire. This does not seem to be possible for the projectivism of Hume and Blackburn.

The truth-aptness objection to Hume/Blackburn. This objection to Backburn's projectivism is from the perspective of ordinary ways of thinking and speaking. I have already mentioned that Blackburn claims that moral assertions are truth-apt, while also acknowledging that his projectivist account of moral assertions is a version of expressivism. Expressivism, as noted above, is normally classified as falling within the purview of

[2] This does seem to be Hobbes' view: "There be in animals, two sorts of *motions* peculiar to them: one called *vital* ... such as are the course of the *blood*, the *pulse* ... the other is *animal motion* ... as to *go*, to *speak*.... These small beginnings of motion, within the body of man, before they appear in walking, speaking, striking, and other visible actions, are commonly called *endeavor* ... This endeavor, when it is toward something that causes it, is called *appetite* or *desire*.... (from *Leviathan*, excerpted in Johnson and Reath (2007), 137–38)

noncognitivism, according to which discourse regarding moral values is to be understood as expressing in a nondescriptive fashion morally relevant attitudes, desires, or commitments – thus entailing that moral utterances are not truth-apt.

The objection that I am addressing here is an objection to which Blackburn himself has responded at length. In a review article, commenting on his own position, Blackburn (2006, 154) says the following:

> Quasi-realism was explained as trying to earn, on an expressivist basis, the features that tempt people to realism. In other words, it suggests that the realistic surface of the [moral] discourse does not have to be jettisoned. It can be explained and defended even by expressivists. Perhaps surprisingly, thoughts about fallibility, objectivity, independence, knowledge, and rationality, as well as truth and falsity themselves, would be available even to people thinking of themselves as anti-realists.

How is this to be accomplished? Michael Ridge (2006, 647) succinctly states Blackburn's answer:

> Blackburn's story ... is to give a deflationist account of truth and truth-aptness, according to which (*very* roughly) there is no robust property of truth and there is no real difference between saying 'p' and saying 'p is true.' On this way of thinking about truth, the fact that ethical sentences are truth-apt may well not reveal anything deep about the states of mind those sentences are conventionally used to express....

The main problem with this deflationary truth strategy as employed by Blackburn is that it begs the question. For mainstream advocates of the deflationary theory – that is, philosophers working mainly within the philosophy of language and metaphysics – the claim that there is no real difference between asserting "p" and asserting "p is true" is meant to apply to those situations, and only to those situations, where we already have good reason to suppose that asserting "p is true" is appropriate, as when we say, to use a textbook example, that "'Snow is white' is true" asserts no more than does "Snow is white." However, if the assertions on which we are focusing consist of the expression of attitudes, then we would not seem to have good reason to suppose that these assertions are such that assigning truth to them is

appropriate. Of course, people all the time *do* suppose that moral assertions are truth-apt; it is precisely this supposition from ordinary discourse that Blackburn wants to accommodate via his projectivist, quasi-realist account of moral discourse. But Blackburn cannot point to this supposition from ordinary discourse in order to avoid the charge of question-begging without having already shown that ordinary discourse, in addition to its presupposition that moral assertions are truth-apt, also presupposes that expressivism is correct, rather than presupposing that moral realism is correct. No worthwhile philosophical purpose is served by substituting one apparent instance of question-begging for another. Likewise, if we are willing to stipulate that the processes involved in moral discourse projectivism whereby – as Hume describes them, and Blackburn concurs – an attitude is "spread on" an object do in fact yield a moral property as belonging to the object, or at least something that is like a moral property, then again we would have good reason to suppose that asserting "p is true" in a moral context is appropriate – and we could then move on to the task of determining how a deflationary account of truth fits into this picture. Above, I argue that we should not accept the outcome of the processes involved in moral discourse projectivism that I have just described (processes said to yield moral properties as belonging to objects). Therefore, if I am on the right track here, there does not seem to be any way for Blackburn to show that his deflationist truth strategy applied to moral assertions is not question-begging.

2. A New Version of Projectivism; Some Preliminaries

I am giving the name "value property projectivism" to the basic view that I wish to defend in this paper. Unlike other versions of projectivism in metaethics, value property projectivism, as its name suggests, involves the projection of properties rather than attitudes, desires, or sentiments. Value property projectivism thus avoids the problems that attend upon the claim that attitudes, desires, or sentiments can be projected.

Necessarily, since projectivism of any type involves processes that must begin with human subjects insofar as they have experience of objects, if value properties are to be projected, they must primarily be located in, or primarily belong to, human subjects, and more specifically must primarily be located in individual states of consciousness – just as phenomenal greenness is located in individual states of consciousness. There are, I believe, two such value properties that are located in, or belong to, individual states of consciousness,

namely goodness and badness. When an individual state of consciousness possesses the property goodness, then that state of consciousness is experienced to *be good* by the person whose state of consciousness it is. When an individual state of consciousness possesses the property badness, then that state of consciousness is experienced to *be bad* by the person in question. Such instances of goodness and badness do not just by themselves have implications for moral values, the nature of which is more complex. I argue below that moral values (and moral disvalues) are "built up from" such instances of goodness and badness in conjunction with other factors; in themselves, such instances of goodness and badness are nonmoral value properties.

As an illustration, consider the state of consciousness that belongs to someone – let us call him Jim – during the time when this person is enjoying lunch. Jim's state of consciousness possesses the nonmoral property of being good but Jim is not likely to be aware of this fact, at least not as something on which he is focusing. While Jim's state of consciousness is at least partially accessible to him via introspection, ordinarily someone who is enjoying lunch would not engage in such introspection but instead would focus on his food or on some other activity that the person happened to be engaging in simultaneously with eating lunch. If asked about the quality of his food or how his lunch was going, Jim would, we are supposing, say something to the effect that the food was good, but would almost certainly not comment on any of the properties of his state of consciousness.

Now, something can be good for many different reasons. I want to focus on Jim's reasons for saying, if asked, that his lunchtime situation is good *here and now*, as opposed to its being good in relation to future benefits from, say, eating properly or spending money wisely. I submit that the fact that Jim's state of consciousness possesses the property of goodness underlies anything that he might say to the effect that his lunch is good here and now. As should be apparent, in claiming this I do not wish to comment on the meaning of "good" as used by Jim when he says that the food is good. Instead, I wish to say something about what, most fundamentally, *is good* in Jim's lunchtime situation as I am describing it.[3] Most fundamentally, it is Jim's state of consciousness that possesses the property of nonmoral goodness. I wish to say further that this property is a natural property but is not – at least at the present time – analyzable. It is a natural property because (1) while at

[3] The point that I am making here, and related points elsewhere, are intended to be in accord with what is described as "new wave moral semantics" by Horgan and Timmons (1992).

present it is discoverable only via experience of one's own inner states, where its standing as unanalyzable is not challenged, given an appropriate theory of mind there is no reason in principle why it could not be analyzed in terms of brain states or brain functions; (2) its existence can be invoked as part of a causal explanation of, for example, the fact that Jim desires his lunch.

Question: Should we not say that Jim's state of consciousness while eating his lunch is *good to* Jim in place of saying that his state of consciousness is good simpliciter? After all, no one but Jim could ever directly experience his state of consciousness as being good, or directly experience it in any other way for that matter (assuming as correct the widely held view that one's own states of consciousness are private in a relevant sense). What this shows, however, is that the distinction between *good to Jim* and *good simpliciter* does not arise regarding value properties ascribed to Jim's state of consciousness. A somewhat different but related problem does arise regarding the transition from the projection of nonmoral values to the projection of moral values: Are moral values relative to persons or groups in the sense of applying only *to* those persons or groups? I discuss this issue in a later section.

The position that I am defending here can fruitfully be compared to the position defended by Derek Parfit in "Normativity" and elsewhere even though Parfit defends a version of nonnaturalism and locates normative properties differently than I do. What Parfit's position and the position that I am defending have in common is the claim that normative *properties*, rather than attitudes or desires, reside at a fundamental level of analysis for value claims, including moral claims.[4] I locate the normative properties in states of consciousness, while Parfit locates them in objects and states of affairs that lie outside of states of consciousness. According to Parfit (2006, 331), "Normative concepts form a fundamental category." I agree. Parfit rejects the claim, as I do, that normative concepts must be restricted to the category

[4] The position that I am defending can also fruitfully be compared to the sort of objectivist position that is the focal point for the attack famously leveled by Mackie (1977) against what Mackie characterizes as the standard way that moral properties are treated in ordinary discourse: This is the view that moral motivation springs directly and completely from an awareness of moral properties. I am concerned with value properties (and derivatively with moral properties) - which, I wish to say, are such that motivation springs directly from them. For Mackie, such motivating properties, if there were any, would reside in external objects, and for that reason would be queer, he says. I am arguing that such motivating properties belong fundamentally to states of consciousness, and for that reason Mackie's charge of queerness does not apply to them.

of that which is moral; practical decisions, he says, need not involve moral thinking. Parfit says (332): "If we believe in irreducibly normative truths, we are what Korsgaard calls *dogmatic rationalists*." I accept this way of characterizing both Parfit's position and my own position. One of the consequences of accepting a "dogmatic rationalist" position in value theory is, as Parfit notes, that there is not a lot that is positive (as opposed to negative – criticizing other views) that one can say in defense of one's position. What an advocate does, essentially, is to bring to the reader's or listener's attention the sorts of instances of normative properties that the advocate believes exist. This is what Parfit does, and this is what I am attempting to do.

Question: Why should we say that the goodness of Jim's lunchtime situation here and now is located fundamentally in Jim's state of consciousness*?*

Consider that if Jim were unconscious he would not be enjoying lunch even if he were somehow – perhaps in his (dreamless) sleep – going through the motions of eating. Consider next that some people seem to lose the capacity to enjoy things while retaining the capacity to go on doing those things. As far as Jim's lunchtime situation is concerned, this would mean that everything could remain the same as regards what Jim is doing and what, if anything, is being done to him – the same act of eating, the same food, the same atmosphere in the restaurant, the same degree of Jim's being nourished, etc. The only difference would be a lack of enjoyment on Jim's part, which accordingly would seem to be a conscious phenomenon. The essential variable, therefore, would seem to be Jim's state of consciousness. More specifically, we might ask: Is Jim *responding consciously* to his lunchtime situation in such fashion that his experience is good? If the answer is *yes*, then it would seem that we should predicate goodness of Jim's state of consciousness. Once we have done this, we can predicate goodness in a derivative fashion of other elements in the situation, such as the food Jim is eating or the conversation he is having with a companion. Below, I say more about such derivative applications of the predicate "goodness."

First, I will address the following question: In order to be good, must Jim's state of consciousness be pleasurable? Consider that it makes sense to say that Jim's state of consciousness is good in situations where references to pleasure are not necessary and perhaps are flat-out inappropriate, as for example a quiet evening at home for Jim where his state of mind is primarily relief that a stressful day is coming to an end. Hedonists such as Epicurus

sometimes appeal to the idea of modest or tranquil pleasure that may seem indistinguishable from someone's simply not experiencing pain, in order to include cases where, if ordinary language is our guide, we would probably not want to say that pleasure is present.

In addition to the ordinary language-based objection to the hedonist's attempt to include all instances of intrinsic good within the scope of the term "pleasure," hedonism is vulnerable on the basis of notorious difficulties that are present when any attempt is made via introspection to locate an element that can be characterized as pleasure on any and all occasions when someone can correctly be characterized as having a good, or positive, experience. These introspective difficulties for hedonism, to which I cannot do justice here, have been extensively discussed elsewhere.[5] Because value property projectivism, as I am defending it, does not appeal to the concept of pleasure, it is in no way compromised by any of these difficulties.

As regards Jim's lunchtime situation, the most important thing to be said in the context of the present discussion is that something about the situation is good. Most fundamentally, I submit, it is Jim's state of consciousness that is good. Generally speaking, nothing besides a person's state of consciousness is universally and noncontroversially present on any and all occasions when we want to say that the person's experience here and now is positive, or good. Moreover, this picture of the situation nicely fits with the link between goodness and motivation. Most straightforwardly, desires motivate because they are oriented toward goodness, as I observed earlier; we might just as well say that goodness motivates. Why, then, need goodness be said to motivate only via the intermediary of pleasure?

The view that I am defending here is consistent with a natural way of describing the phenomenology of motivation by values, moral or otherwise. We do not normally say that we are motivated by our desires (some of which, after all, we do not wish to allow to possess motivating power), but rather by our reasons, which are oriented to the properties that we apprehend desired

[5] A good treatment is Katz, Leonard. "Pleasure" *The Stanford Encyclopedia of Philosophy* (Fall 2007 Edition), Edward N. Zalta (ed.), http://plato.stanford.edu(archives)fall2007/entries/pleasure/ See especially 8–13. A good statement of one aspect of the problem is Feldman (2004, 79): "Consider the warm, dry, slightly drowsy feeling of pleasure that you get while sunbathing on a quiet beach. By way of contrast, consider the cool, wet, invigorating feeling of pleasure that you get when drinking some cold, refreshing beer on a hot day. Each of these experiences involves a feeling of pleasure - a sensory pleasure, in my terminology - yet they do not feel at all alike. ... I have come to the conclusion that they have just about nothing in common phenomenologically."

things as possessing. On the basis of such reasons, we may yield to a particular desire as regards motivation. On the basis of other reasons, we may dig in our heals and not yield to some other desire.

Question: If what I have been saying about the possession of goodness by states of consciousness is correct, then why is it that our beliefs do not ordinarily reflect the fact that this is so? Instead, we tend to believe that goodness, in many if not most instances, belongs fundamentally to objects, as opposed to states of ourselves. We believe that food is good. Or, if food is judged to be mainly a means to an end, then health is good or life is good. Jokes that are genuinely funny are good. Knowledge and beauty are good.

By way of answering the question posed at the beginning of the previous paragraph, I ask readers to consider that, in terms of the evolutionary development of human beings, there is little if anything to be gained toward a person's survival from the possession of an inclination to focus at all frequently on one's conscious states, while much is to be gained from the possession of an inclination to focus on the details of, say, what one is eating or the people with whom one is dealing. Knowing about the properties that we experience as being possessed by food will help us to obtain food and guide us when we eat it, and knowing about the abilities, dispositions, intentions, etc. of people whom we encounter will help us to benefit from our encounters. We have, therefore, developed a strong inclination to focus on external objects, not states of our consciousness.

At the same time, evolutionary considerations dictate that there be an inner mechanism that rewards or punishes us for, respectively, engaging with things that have survival value and engaging with things that do not have survival value. An *inner* mechanism is required since rewards and punishments that arise from circumstances outside of a person, such as from the existence of parental or social sanctions for harmful behavior, will not be constantly present, while potential benefits and injuries from the objects one encounters are constantly present. And even more important, there does not seem to be any way to conceive how external sanctions – those residing in circumstances outside of a person's consciousness – would actually function as sanctions except via the role of inner sanctions, namely rewards and punishments that are manifested within a person's consciousness. How, for example, would the threat of imprisonment as punishment function as a deterrent if the person being threatened did not perceive the experience of imprisonment, or some consequence of imprisonment upon future experiences, to be bad? The one constant accompaniment for all of our

experiences of objects in the world is our awareness of those objects, so it makes sense from an evolutionary perspective that rewards and punishments would be built into states of consciousness that are part and parcel with our states of awareness – as long as the proviso is attached that these inner rewards and punishments will play a background role in our experiences so as not to distract us from the all-important need to focus our attention on the objects that we encounter. I submit that the most general and basic way to characterize such inner rewards and punishments is to say that the states of consciousness associated with the rewards and punishments possess the properties of goodness and badness. An understanding of such properties, in conjunction with the evolutionary demand that we focus our attention outwardly yields value property projectivism.

A possible objection: Is it not the case that, even if we "defy evolution" and make a point of using our introspective powers to focus extensively on our states of consciousness rather than external objects, we will come up short in our attempts to focus our attention on the alleged properties of goodness and badness belonging to our states of consciousness? Just as hedonism faces introspective difficulties, will not a search for these alleged properties face introspective difficulties? My response is that value property projectivism does not require introspective success in fixing on the properties of goodness and badness as pertaining to our states of consciousness. Although it seems to me that such properties are accessible to introspection, I am willing to acknowledge that perhaps they are not, and if this should prove to be the most defensible view I would reframe my characterization of them. I would not treat these properties as phenomena open to introspection but instead as phenomena to be hypothesized. I would say that value property projectivism rests upon the hypothesis that these properties exist as belonging to states of consciousness.

In terms of the phenomenology of experiencing the evolutionary theory-oriented rewards and punishments to which I have been referring, it makes sense to say that we project out onto the things in the world certain value properties. Because the projection functions so seamlessly, as I am supposing that it does, we are usually not aware that it is taking place, but rather come to believe that it is objects such as items of food that are good or bad. In what follows, I say more about how projection works with moral values, and I compare projection as it pertains to values with projection as it pertains to secondary qualities such as colors. I discuss the link between projected values and motivation.

3. From Nonmoral Values To Moral Values

How do we get from the projection of nonmoral values to the projection of moral values? Let us begin by looking at the least problematic type of property projection, that involving Lockean secondary qualities.

Regarding color ascription, as I have noted, in the case of greenness it makes sense to say that "something in my mind" (phenomenal greenness) is projected by me – as a psychological act on my part – onto the leaves of the tree at which I am looking. What occurs is a private operation within my consciousness that happens to stand in a relation to an external object. Because other people (all who are normal color observers, like me) carry out essentially the same operation, there is an objective basis for saying that the correct answer to the question, what color are the leaves, is that they are green. Contrast this with how people experience the textbook substance pheno-thio-urea, which only a minority of the human population perceive to be bitter. This minority only, we can say, project phenomenal bitterness onto instances of experienced pheno-thio-urea. At the same time, this minority is large enough that the concept of normal observer has no application as regards human sensory experience of pheno-thio-urea, and as a consequence, there is no objective basis for saying that the correct answer to the question, what is the taste of pheno-thio-urea, is either that this substance is bitter or that it is not bitter. Is there, then, an objective basis for saying that pheno-thio-urea is *bitter to* someone who belongs to the relevant minority of perceivers? The term "objective" is broad enough that an affirmative answer to this question is certainly acceptable. After all, it is a fact that a particular person who belongs to the population in question does experience pheno-thio-urea to be bitter. Moreover, if the human population changed in its makeup so that all or most people came to experience pheno-thio-urea to be bitter, then we (the human population as a whole) would have acquired an objective basis for saying that pheno-thio-urea is bitter simpliciter, just as we can say of the leaves mentioned earlier that they are green simpliciter.

Now, let us look again at the example of Jim who finds his lunch to be good here and now. I have argued that he is projecting the property of goodness onto his lunch, but of course we cannot say that Jim's lunch is good simpliciter since other people – with as much claim as Jim to being normal observers, or evaluators, of lunch-type foods – may eat the same food that Jim has eaten and not like it. Therefore, the situation is similar to that of pheno-thio-urea, which is *bitter to* some people but not to others.

We have, then, the basis for what may look like common ground between projectivism applied to certain Lockean secondary qualities and projectivism applied to values such as *goodness to P* (where "P" stands for some person or group). However, this common ground does not extend to moral values. Consider, first, that for *nonmoral* values, in order to progress from *goodness to P* to *goodness simpliciter*, all we need is a population with the right composition. Consider the obvious value of relaxation (on appropriate occasions). Relaxation on an appropriate occasion is good to Jim, let us say, but it is also good to virtually every other human being; accordingly, it is correct to say that it is good simpliciter – but only as long as we do not mean that it is morally good, or (if there is reason to prefer "moral rightness" to "moral goodness" as a basic value predicate) that it is morally right to pursue relaxation on the occasions in question. Perhaps one's moral duty will not be fulfilled if one relaxes. At most – in terms of certain ethical viewpoints – relaxation is only prima facie morally good or morally right to pursue. We certainly cannot say without qualification that relaxation is morally good or morally right, while we can say without qualification that relaxation is good in a nonmoral sense of being good – and in this respect relaxation is like a great many other nonmoral objects or states of affairs.

Why the difference between the moral and nonmoral cases? To locate the basis for an answer, let us consider colors again. In the case of greenness, if my color perception is out of sync with that of the large majority of observers (I suffer from a type of red-green color blindness), then we might say, not that the leaves are red, but that they are "red to me." Correspondingly, if I were a masochist it might be said that certain types of pain are "good to me" (I project goodness onto them). However, "moral to me" makes no sense unless we accept a radical version of individual moral relativism that no actual advocates of metaethical projectivism at the present time, or advocates of any other major metaethical theory for that matter, would accept. Thus, the analogy with colors clearly breaks down at this point, and this is instructive because it indicates that common ground between nonmoral and moral values also breaks down at a certain point. One of the lessons here is that we cannot think of moral values as simply being "built up from" nonmoral values via the mechanism of arriving at population-wide convergence as regards what – without that convergence – would simply be instances of *goodness to P*. Something more is needed.

If the situation were otherwise, moral values would be wholly conventional, the consequence of whatever (initially nonmoral) value

convergence a population happened to experience. For example, starting from the observation that neglecting all considerations of justice is good to Jim (on the supposition that Jim is the sort of person for whom this is true; he projects goodness onto all relevant situations where justice is neglected), we would be able to proceed to the conclusion that neglecting all considerations of justice is morally good simpliciter if the requisite population convergence were to come about or to be discovered (which would encompass a situation where most people in the population turned out to be like Jim, not caring at all about justice), which is an unacceptable conclusion. Whatever else we may want to say about the projection of value properties, we want to be able to say that not all projections of value, even where populations are unanimous in embracing them, can count as projections of moral value.

Hume recognized this problem, and attempted to solve it by stipulating that only "disinterested approbation" qualified as the sort of attitude that, when projected, constituted a moral response. If I have more sympathy for my friend than for a stranger, when both are being treated with the same degree of injustice, then I need to correct for my personal bias, according to the view of Hume, in order to exhibit an appropriately disinterested sympathy if my response is to be considered a moral response. The problem with Hume's position is that the resulting determination that act A (reflecting disinterested approbation) is moral, while act B (reflecting biased approbation) is not moral, becomes a matter of definition, and thus convention: Ordinary usage – as Hume understands it – assigns "moral" in the one case but not the other. Aside from taking note of the attitudes or sentiments that we may happen to find ourselves and our fellows possessing on a given occasion, we have no more reason to act morally than to act nonmorally – which apparently was Hume's own view, but not something that he worried about because, for one reason, he apparently believed in the existence of a de facto uniformity, or near uniformity, of the human population as regards the possession of disinterested approbation in relevant situations.[6]

We know now that the overall human population is much less homogeneous than Hume imagined. Of course, "morally homogeneous societies" do exist and have existed throughout history, especially within primitive cultures. An examination of such societies will allow us to construct a simplified picture of the transition from nonmoral values to moral

[6] "The minds of all men are similar in their feelings and operations." (from the *Treatise*, quoted in Johnson and Reath (2007, 175))

values from the perspective of value property projectivism. After doing that, we can move to a discussion of moral values in our contemporary society, which is not morally homogeneous.

A morally homogeneous society is one in which few, if any, individuals raise serious questions about the moral values of the society. Essentially, the concept of moral progress is absent in such societies: There can be progress regarding compliance with moral norms, but not regarding the norms themselves. (Compare: No sense can be attached to the idea of "color perception progress" in situations where "normal color observer" comes to range over a different population than previously.) Thus, from the perspective of the large majority of people in the morally homogeneous society whom we judge to be moral individuals (virtually every society has criminals and misfits, and among these individuals some bona fide amoralists, all of whom we exclude for present considerations), the Humean picture is satisfactory if we interpret it from the perspective of value property projectivism. A moral individual will, first, project nonmoral values, just as Jim does in the lunchtime situation mentioned above, and will act on these values when they do not conflict with moral values. When there are conflicts, a moral individual will – at least usually or largely – be motivated more strongly by moral considerations. What this means, first, is that a moral individual P projects the property of goodness more strongly onto those specific items (actions, intentions, states of affairs, rules – whatever) to which P's society assigns moral recognition than onto any other items; moral recognition simply reflects unanimity, or near unanimity, as regards the set of items onto which members of P's society project goodness. (For Hume's society, if we suppose that Hume described it correctly, the primary such items were actions undertaken in response to disinterested approbation as projected onto the same set of items by most members of the society.) Another way to describe the situation is to say that P projects *overriding goodness* (which is the strongest version of goodness available to P), onto the items in question – at least usually or largely.

The qualification "at least usually or largely" needs to be added because few individuals – even those whom we would describe as being thoroughly moral – can be expected always to embrace doing the exact morally right thing as determined by the values of the society to which these individuals belong. Therefore, we need to add a second condition that P must meet in order to be described as moral, namely that P projects goodness, or overriding goodness, onto a general inclination on P's part to conform to his

society's moral values. The possession of this general inclination makes up for P's occasional moral lapses as regards specific actions, while allowing P to be a "thoroughly moral" individual. This second condition involves projecting value properties onto the projecting of value properties. Another way to describe this second condition is to say that P possesses a moral self-image which is a more or less constant element in P's psychological makeup, and serves to guide and prod P even when P may not project goodness or overriding goodness onto specific items to which P's society assigns moral recognition. P thinks of himself as being a moral individual overall, and he is correct in thinking this.

4. Defending Moral Value Property Projectivism (First Round)

I wish to defend the following: For morally homogeneous societies, if we accept the perspective of value property projectivism then there will be no problems or complications in saying that moral assertions are truth-apt, that moral realism, naturalism, and cognitivism are correct, that the perspective of the amoralist can be accommodated, and that moral beliefs motivate (except for the case of amoralists, but they do not pose a problem, as I explain below). For purposes of discussion, let us suppose that for a morally homogeneous society S we have a suitable definition for "just act" and that acting justly is the primary moral value in S. Let us suppose also that an individual member M of S asserts the following:

(1) Acting justly is morally right.

Moral properties are natural properties. "Morally right" is a moral property. It *means* that which one is obligated to do. It *is* that property which is projected onto a class of acts that, for a morally homogeneous society, are determined by the moral orientation of that society. Thus, moral members of the society find themselves projecting goodness, or overriding goodness, onto whatever belongs to the class of acts in question. Just as greenness is a natural property, although a projected one, so also is moral rightness.

(1) is true. As understood by M and other members of S, (1) is true: Just acts *are* acts which possess the property of goodness, or overriding goodness, and they are experienced as such by most members of S. Most members of S (because they are moral) will simply find themselves projecting goodness, or overriding goodness, onto a certain set of acts, just as most people (because

they have normal color perception) find themselves projecting phenomenal greenness onto a certain set of objects.

Amoralists are accommodated. Amoral members of S are individuals who do not find themselves projecting goodness, or overriding goodness, onto many, if any, of the appropriate items, but they may nevertheless understand enough about how the term "morally right" is used extensionally by members of S to be able to assent with understanding to the claim that (1) is true. In a somewhat similar fashion, some blind or color-blind people can learn to determine for some occasions that it is appropriate to say that a particular object is green.

Moral realism and cognitivism are accommodated. If value property projectivism for morally homogeneous societies makes moral rightness a natural property, and (1) true, then value property projectivism supports moral realism and cognitivism for morally homogeneous societies.

5. Defending Moral Value Property Projectivism (Second Round)

I now turn to the task of defending value property projectivism for societies that are not morally homogeneous. The crucial difference between societies that are, and societies that are not, morally homogeneous is that the latter, but not the former, allow for genuine moral progress wherein not only can there be improvement as to degree of compliance with existing moral values, but there can also be improvement in the set of moral values against which degrees of compliance are measured.

The key for moral realists who are projectivists is to recognize that projecting value properties is a natural process that necessarily is constrained by the nature of that of which the process is a function – namely, human beings, who have a variety of needs, both individual and social. At any given time, there will be in place on the part of various members of society the projection of goodness, or overriding goodness, onto various objects and states of affairs. Some of these projections will be coordinated with other projections, some will not be coordinated with other projections. An example of situations involving the former could be states of promise-keeping, while an example of situations involving the latter could be states of seizing and eating any food a person happens to encounter, regardless of who owns the food. On the whole, coordinated projections will be the basis for successful actions (actions where the projected value property is realized) more often than uncoordinated projections. For obvious reasons, disinterested approbation (to cite Hume's category) can be expected to have a relatively high

success rate. A high success rate translates into the perception that being good, or being good in an overriding fashion, as projected, really is good, or good in the strongest sense.

The question, are actions of type A *really* moral, can be seen as a question about the likelihood that actions of type A will in the future prove to be such that they can be coordinated with other actions onto which the members at large of a society are projecting goodness, or overriding goodness. Such a likelihood will necessarily be calculated from the perspective of a given time and location where various value projections are already in place – so it is in this sense context-bound, and has a conventional element. In other words, moral progress is always a progression from what happens to exist at a given time; it is "progress from" what exists as opposed to "progress toward" some already-determined "moral ideal." For the view that I am defending in this paper, no such moral ideal exists – indeed, there is no way even to conceive what it might mean for there to be such an ideal.

For the view that I am defending, we do not *start* by, say, analyzing human nature a la Aristotle, and then *conclude* that, say, the morally right course of action involves promise-keeping (or some other example of a socially cooperative action). Instead, we start with whatever examples of goodness, or overriding goodness, are in fact being projected; then we predict that a higher level of coordination among projected values will be realized if there is a greater degree of promise-keeping; finally, if we observe that the prediction is bourne out, we can then conclude that moral progress lies in the direction of having in place for a society the enhanced degree of premise-keeping. Prior to the time when the prediction is actually bourne out, we may have good reason to believe that it will be bourne out. And indeed, we do have reason to believe that moral progress would occur if there existed a higher degree of promise-keeping than presently exists within virtually any society; we may conclude, if we wish to do so, that we are thereby provided with an insight into human nature.

Likewise, for the view that I am defending we do not start with the "moral intuition" that there would be moral progress if more people acted with the intention to achieve the "greatest good for the greatest number." Instead, we may predict that there would be moral progress if more people acted with the intention to achieve the "greatest good for the greatest number." As it happens, this is a prediction for which there does not appear to be much support, and for several reasons that contemporary normative theorists have noted, such as that the utilitarian criterion is too demanding upon individuals

– which translates into saying that nothing close to unanimity is ever likely to occur within any society as regards the projection of goodness, or overriding goodness, onto actions intended to promote the greatest good for the greatest number. For the view that I am defending, the utilitarian criterion does not allow sufficiently for coordination among all the values that the typical individual living in today's societies does now, and will likely in the foreseeable future, be projecting.

Is there a general moral criterion that might fare better than the utilitarian criterion? I suggest that we specify in a somewhat bland fashion a moral criterion that calls for what can be described as a balance of self-interest against the interests of others (the "Welfare-Balancing Criterion," or WBC); the main goal in appealing to this criterion is to avoid the objection against utilitarianism that it is too demanding upon individuals. I do not have space enough here to expand WBC so as to spell out how WBC might meet other objections to utilitarianism or to other normative views.

Support for WBC would mean that a prediction is bourne out to the effect that, the more that humans know about themselves and the world, and the greater the range of their experiences, the more likely it is for most individual human beings to be more strongly motivated by WBC than by other general moral criteria in situations where these other criteria yield different value projections. Keep in mind that for value property projectivism, motivation (all types, including moral motivation) is purely a matter of the presence within the consciousness of the person who is motivated by an instance of goodness, or overriding goodness. As far as motivation is concerned, these nonmoral properties do all the heavy lifting. We describe the resulting motivation as being moral when one or more properties resulting from projection of these nonmoral properties is a moral property, that is, one involving a high level of coordination as regards value projection for members of a society.

6. Conclusion

The conclusion that I wish to reach in this paper is that there exists a projectivist view of moral value properties that is cognitivist but does not have the problematic "expressivist basis" that Blackburn gives to his version of projectivism. In the view defended here, what is projected is a property, not an attitude, desire, or sentiment. Most fundamentally, it is a value property, which may become a moral value property within the right context. The projection of value properties is sufficiently similar to the situation with

Lockean secondary quality projection that major strengths of the latter carry over to the former. The transition from value property projection to moral value property projection can be accomplished satisfactorily, as regards at least the preliminary presentation of this transition that is given in the present paper.

Acknowledgements
An earlier draft of this paper, presented at a conference arranged for members of the West Virginia University Philosophy Department, benefitted greatly from comments from several of my colleagues. Ted Drange provided many helpful comments at a still earlier stage in my thinking about the issues in this paper. An anonymous referee for this journal provided many helpful comments on the penultimate draft of this paper.

References

Blackburn, Simon. (1981). Reply: Rule-Following and Moral Realism. In Stephen Holtzman and Christopher Leich, (eds.) *Wittgenstein: To Follow a Rule*, pp. 163–187. London: Routledge. Reprinted in Andrew Fisher and Simon Kirchin, (eds.). *Arguing About Metaphysics*, pp. 470–488. London: Routledge.

_____. (2006). Antirealist Expressivism and Quasi-Realism. In Copp. (2006), pp. 147–62.

Copp, David, (ed.). (2006). *The Oxford Handbook of Ethical Theory*. Oxford: Oxford University Press.

Feldman, Fred. (2004). *Pleasure and the Good Life*. Oxford: Oxford University Press.

Fisher, Andrew and Kirchin, Simon, (eds.). (2006). *Arguing About Metaethics*. London: Routledge.

Gibbard, Alan. (1994). Meaning and Normativity, *Philosophical Issues*, Vol. 5, pp. 95–115.

Horgan, Terence and Timmons, Mark. (1992). Troubles for New Wave Moral Semantics: The Open Question Argument Revived. *Philosophical Papers*, 21, pp. 153–75. Reprinted in Fisher and Kirchin. (2006), pp. 179–199.

Johnson, Oliver and Reath, Andrews. (2007). *Ethics: Selections From Classical and Contemporary Writers*, (tenth edition). Belmont, CA: Thomson.

Katz, Leonard. Pleasure. *The Stanford Encyclopedia of Philosophy*, (Fall 2007 Edition. Edward N. Zalta, (ed.), http://plato.stanford.edu(archives) fall2007/entries/pleasure/

Mackie, J. (1977). *Ethics: Inventing Right and Wrong*. New York: Penguin.

Parfit, Derek. (2006). Normativity. In Shaffer-Landau, (ed.). (2006), pp. 325–380.

Ridge, Michael. (2006). Saving the Ethical Appearances. *Mind*, Vol. 115, pp. 633–650.

Sayre-McCord, Geoffrey, (ed.). (1988). *Essays in Moral realism*. Ithaca: Cornell University Press.

Shaffer-Landau, Russ. (2006). *Oxford Studies in Metaethics*, Vol. 1. Oxford: Oxford University Press.

Swinburne, R. G. (1976). The Objectivity of Morality. *Philosophy* 51, pp. 5–20.

Ralph W. Clark
West Virginia University
256 Stansbury Hall
Ralph.Clark@mail.wvu.edu

Remarks on Thomas Reid's allegedly Newtonian Science of the Human Mind

Robert Callergård

Abstract
It is commonly assumed that Thomas Reid advocated a Newtonian approach to the study of mental phenomena. I argue to the contrary that there are few good philosophical reasons for such a characterization. Reid is highly critical of attempts to model the study of mind on the model of physics. Typical features of physical theory that Reid rejects for the study of mind are measurement of quantities, multi-layered axiomatic structure, and any analogy between mental and material phenomena. The only similarity there is between the study of material phenomena and mental phenomena is that both, according to Reid, are concerned with laws of nature. But quite unlike physics, in which laws serve as the backbone of theory (description, explanation), laws have an almost negligible part to play in Reid's treatment of mental phenomena. A main reason for this, I suggest, is that most operations of the mind typically involve exercise of active power, that is, we take part in them as agents, we engage in them.

0. Introduction

Routine has it that the Scottish philosopher Thomas Reid (1710–1796) advocated a Newtonian approach to the study of mental phenomena.[1] I argue to the contrary that there are few good philosophical reasons for such a characterization, however useful it may be as a merely historical classification. In fact, not only is Reid critical towards attempts to model the study of mind on the model of physics, including such features of physical theory that were associated at the time with Newton, in addition Reid's own efforts to reform the study of the mind bears little resemblance to anything of the sort of theories that might reasonably be associated with Newton. There is, however, a stronger claim to be made: Even if Reid regards mental phenomena as largely speaking natural phenomena, and as such should be

1 See for instance Yaffe (2004), p. 3.

Filosofiska Notiser, Årgång 3, Nr. 2, Augusti 2016, 29–51

studied with the same rigour and in the same spirit as material phenomena, and even if he affirms, as he does, that laws of (mental) nature is something to be looked for in the study of mind, there are some fundamental reasons why a science of the mind, in fact, will not be much concerned with laws of nature. One important reason for this may very well be that Reid's description of the mind is based on the concept of mental acts, as opposed to, typically, Hume's approach in which the mind is seen as a stream of perceptions or mental states, without unity and connection. Another important reason, which Reid notes occasionally, is that insofar as the human mind has the capacity to exert active power, which for Reid implies possessing a power to act freely, it seems impossible to describe this aspect of the mind accurately in terms of deterministic laws of nature on the model of physics. Since Reid, however, never argues that, *because* mental phenomena are best described as mental acts rather than as a stream of Humean perceptions, *or because* we can act genuinely freely, *therefore* laws of nature are quite irrelevant to the study of the mind, I will suggest a deeper and more interesting feature of Reid's conception of human nature which explains why the science of the mind will not be much concerned with laws of nature. The reason is that most operations of the mind are such that we take part in them as agents; they are not events that merely happen to us, instead, we engage in them.

In what follows I will (1) set forth some of the reasons why Reid's science of the mind might be assumed to be a typically "Newtonian" enterprise, show (2) why Reid rejected various imitations of physical theory, explain (3 & 4) the limited role of law-like statements in Reid's reformed science of the mind, and lastly (5) suggest a reason why mental phenomena, they way Reid understands them, will only rarely and effectively be subsumed under laws of nature.

1. The science of the mind as a Newtonian enterprise

Newton's take-over of much of the natural philosophy scene in the 18th century inspired many attempts to transfer something of Newton's handling of scientific questions in mechanics, astronomy, and optics, not only to neighbouring subjects like electricity and chemistry, but also to the study of the phenomena of politics and morals, and life. Calls were also made to apply more rigorous methods to the study of mental phenomena, but what role Newtonian procedures ought to play in a reformed science of the mind was not always altogether clear. What are those excellent features of

Newton's treatment of physics that the scientist of the mind should learn from and try to repeat, adapt, or copy for his own field of study? Many things conspire to make it seem that Reid's quest for a new solid science of the mind had a lot to do with Newton. Indeed, as anyone familiar with his writings can see, Reid repeatedly refers to Newton with greatest admiration, and among the themes he specifically and repeatedly brings to the forefront is Newton's conception of science, specifically the *Regulae Philosophandi* that Newton set down in the third book of *Principia*.

The simplest way to link Newton with Reid's ambition to reform the science of the mind is by pointing to similar efforts of George Turnbull (1698–1748) and David Hume (1710–1776), whose writings explicitly and implicitly contained overt references to Newton and Newtonian ideas, and both of which certainly had an impact on Reid's own thinking in its earlier phases. Turnbull was Reid's main teacher at Marischal College, Aberdeen, his *Regent*, which means he taught all subjects except mathematics through three years of study. At the time Reid was in his early teens and Turnbull in his mid twenties. In 1740 Turnbull published *Principles of Moral Philosophy*, a work that contains an account of the mind based on the association of ideas and explicitly inspired by Newton's assertion in the *Opticks* according to which the very same method used successfully by Newton in natural philosophy should also work for moral sciences.[2] Reid studied Hume's *Treatise of Human Nature* (1739–1740) at the time of its publication, and Reid's whole oeuvre might very well be seen as a life-long engagement with Hume's philosophy.

Reid's background was enriched also by other Newtonian influences. Reid's teacher of mathematics at Marischal college was non less than the famous mathematician Colin Maclaurin, author of *Treatise of Fluxions* (1742) and *Account of Newton's Discoveries* (1748), and in addition Reid belonged on his mothers side to the famous Gregory family which contributed to the early introduction of Newtonian ideas in Scotland. These are well-known facts about Reid's philosophical and scientific background and they suffice to show that Reid could profit from a variety of sources of allegedly Newtonian thinking.

But we would do wrong to assume that Reid would not assimilate these influences in his own way. For one thing, he knew Newton's writings first

2 Turnbull quoted from *Opticks*: "And if natural Philosophy in all its Parts, by pursuing this Method, shall at length be perfected, the Bounds of Moral Philosophy will be enlarged." Newton (1979 [1704]). *Opticks*, Dover Publications: New York, p. 405.

hand. He had studied Newton's *Principia* together with John Stewart, his friend and class mate from Marischal College, who at young age had succeeded MacLaurin at Marischal in 1727 as professor of mathematics.[3] Later, in the 1750ies Reid taught physics at King's College, Aberdeen, and his occupation with the meaning of Newton's teaching increased, which can be seen from his published books and in unpublished writings.[4] For another thing, it will become clear as we proceed that Reid rejected the associationist models of the mind of Turnbull and Hume. Indeed, Reid explicitly rejected the idea that mental items can be treated in analogy with material bodies which are subjected to general laws of attraction.

2. Structural imitations of physics

Since alignment with Newton and his work was almost a national sport at the time in Britain and since we have every reason to think that Reid would rather form his own judgement of and from Newton's writings than merely succumb to second-hand sources, which by the way would only rarely play the same tune, we do best to look at some of the principled objections that Reid aired against misconceived adoptions of typical traits of physical theory in the study of mind. I will look at three aspects of physical theorizing that may very well be associated with Newton's physics: quantitative methods, axiomatic form, and mind-matter analogies.

2.1 Quantity and measurement

From remarks made in Reid's first published paper "An Essay on Quantity" (1748) it is clear that Reid did not believe that mathematics is the key to unfold the secrets of the human mind. In this small piece on measurement theory Reid criticises Francis Hutcheson's attempt, in his own words on the title page of the first edition of *An Inquiry into the Original of our Ideas of Beauty and Virtue*, "to introduce a Mathematical calculation in Subjects of Morality."[5] The idea is to be able to "compute the Morality of any Actions" and for this purpose Hutcheson sets forth a number of "Proposition of Axioms" all of which are expressed in mathematical style. For instance, the

3 Broadie, Alexander (2004). "Reid in context", in *The Cambridge Companion to Thomas Reid*, Cambridge University press: Cambridge. Wood, Paul (1993). *The Aberdeen Enlightenment*, Aberdeen university Press: Aberdeen 1993, p. 19f.

4 See Reid (1995).

5 Subtitle of the first edition (1725) of *An Inquiry into the Original of our Ideas of Beauty and Virtue*. This subtitle was omitted in later editions, while the sections referred to remained.

formula "M = BA" expresses an equivalence said to hold between "the moment of good" (M) and the product of the "benevolence" (B) and the ability (A) of an agent. In other words, to get a measure of the quantity of good produced by an agent, you multiply the benevolence of the agent with his practical ability to exert this benevolence. In similar style Hutcheson defines quantities of hatred, interest, and moral evil, in an array of formulas. Reid's criticism is principled and hard. He points out that, while it is intelligible to talk about different degrees of virtue, taste, pleasure, beauty, intelligence, etc., it is quite another thing to be able to measure them, that is, to assign numerical values according to an accepted standard of units.

> To talk intelligibly of the Quantity of Pain, we should have some Standard to measure it by, some known Degree of it so well ascertained that all Men, when they talked of it, should mean the same thing; we should also be able to compare other Degrees of Pain with this, so as to perceive distinctly, not only whether they exceed or fall short of it, but how far, or in what proportion. Whether by a half, a fifth, or a tenth.[6]

In Reid's view there is nothing wrong in principle to introduce an "improper quantity" if it is defined in terms of "proper quantities". Speed for instance is an improper quantity which is defined in terms of distance and time. A proper quantity is a quantity which "is measured by its own Kind; or which of its own Nature is capable of being doubled or tripled, without taking in any Quantity of a different Kind as a Measure of it." Reid mentions extension, duration, number and proportion as proper quantities. Speed and quantity of motion, on the other hand, are improper quantities, and they can be measured because they are defined in terms of proper quantities.

For Reid an improper quantity is an invention or an "artifice" that works if and only if it is defined in terms of proper quantities. Reid's solution to the so called Vis-Viva controversy, that is, the controversy whether quantity of motion should be defined as the product of mass and velocity or as the product of mass and the square of the velocity, is that this is not a matter of finding out that true concept of quantity of motion which corresponds to reality. Instead, it is a matter of choosing between alternative definitions, something that should be decided by considering matters of convenience and simplicity in discourse. Reid does not mind therefore that Hutcheson

6 "An Essay on Quantity", section 1.

introduces quantities we never heard of, because there is a sort of instrumentalism to Reid's conception of improper quantities.

It is otherwise with proper quantities. In order to be measurable on their own they need to meet at least two conditions. First, they need to be structured in such a way as to allow that it is intelligible to say that some instance of the quantity is so and so many times bigger than another instance, and secondly, we need to have some criteria to tell us what proportions are instantiated, that is, a method of measurement. Now, it is clear that Reid has no confidence whatsoever that Hutcheson or anyone could meet these conditions.

> Although attempts have been made to apply mathematical Reasoning to some of these Things, and the Quantity of Virtue and Merit in Actions has been measured by simple and compound Ratio's, yet I do not think that any real Knowledge has been struck out this Way: It may perhaps, if discretely used, be a Help to Discourse on these Subjects, by pleasing the Imagination, and illustrating what is already known; but until our Affections and Appetites shall themselves be reduced to Quantity, and exact Measures of their various Degrees be assigned, in vain shall we essay to measure Virtue and Merit by them. This is only to ring Changes upon Words, and to make a Shew of mathematical reasoning, without advancing one Step in real Knowledge.[7]

Fours decades later Reid repeated the verdict:

> This may perhaps, in the way of analogy, serve to illustrate what was before known; but I do not think any truth can be discovered in this way. There are, no doubt, degrees of benevolence, self-love, and other affections; but, when we apply ratios to them, I apprehend we have no distinct meaning.[8]

2.2 Axiomatic deductive structure

The old idea that scientific knowledge ideally has a multi-layered axiomatic structure, which sprung from Euclid's *Elementa* and Aristotelian schemes of definition, gained momentum as a viable idea for natural philosophy with the

7 "An Essay on Quantity", section 4.

8 *Intellectual Powers*, p. 546.

gradual development of the mechanical sciences in early modern times. New and more effective applications of mathematics gave scientists new powers of prediction and explanation of astronomical and earthly phenomena, which in turn made it handy to structure expositions of theory in axiomatic style. Newton's derivation of Kepler's laws, and the incorporation of some work of Galileo and Huygens into the theoretical structure of the *Principia* also suggested that there might be many layers of laws of nature in which some are more basic than others. This is also Reid's vision of physics. Laws are general facts, but some laws are more general than others and they can be used to demonstrate laws that are less general. The most general laws we know we usually call laws of nature and they are the ones that end up as principles or axioms in treatises. Thus Reid explains the law that bodies falling towards the centre of the earth accelerate, by showing that this law is the necessary consequence of the laws of gravity and inertia together. Thus he also interprets Newton's work on ether theories as the search for a set of more general laws from which to demonstrate the law of gravity, and thus he discusses the question whether or not the law of gravity is deducible from the three laws of motion.

Now, Reid repeatedly stresses the importance for any science to get clear about its first principles. He argues that agreement in principles is requisite to conduct successful scientific arguments and that it is mark of a "mature" science that its first principles have been settled to the satisfaction of the scientific community.[9] Beyond that, however, there is little to suggest that Reid envisioned an axiomatic and multi-layered structure for a mature science of the mind. Indeed, one of the faults Reid found in Hume's science of man was the ambition to rest a "complete system" on just a few principles (i.e. the copy principle and the laws of association). The very idea that the number of principles ought to be very small for a successful account of mental phenomena is, as we have already seen, prejudicial. In contrast, Reid's system contains a considerable number of principles that outnumber any theorems deduced from them or is explained by them. The contrast in this respect between Reid's own approach and Hume's was wholly clear to Reid, and it appears that he was sensitive to a critical remark made by Joseph Priestley – a stout advocate of associationist psychology – to the point that he (Reid) tended to explain the mind by an abundance of original and unaccountable principles when it would be more scientific, in Priestley's view, to reduce the number of principles as far as possible; preferably to

9 *Intellectual Powers*, p. 62.

principles of the association of ideas.[10] It is true that a considerable part of Reid's efforts consists in detailed descriptions of mental phenomena which often end in the identification of "original principles" or "ingredients" that make up mental acts. In the *Intellectual Powers* Reid apparently responded to Priestley:

> I believe the original principles of the mind, of which we can give no account, but that such is our constitution, are more in number than is commonly thought. But we ought not to multiply them without necessity.[11]

A conclusion to make so far is that, with no urge whatsoever for a mathematical approach of measurement, and with no apparent desire for a foundation for the science of the mind consisting of a small set of principles or axioms, there is little reason for Reid to strive for a multi-layered axiomatic deductive structure based on laws of mental nature. We will see later that the principles that Reid tries to settle for the mind are only rarely of a law-like nature.

A short digression before going on: In view of the fact that Reid emphasized the importance for any science to settle "its principles" in order to be "mature" it should be asked what kind of relation between principles and superstructure he envisioned.[12] Unfortunately, it is difficult to find a detailed answer in Reid's writings. In the case of ethics, however, he specifically deny that it is a relation of evidence.

> A system of morals is not like a system of geometry, where the subsequent parts derive their evidence from the preceding parts, and one chain of reasoning is carried on from the beginning; so that, if the arrangement is changed, the chain is broken, and the evidence is lost. It resembles more a system of botany, or mineralogy, where the subsequent parts depend not for their evidence upon the preceding, and the arrangement is made to facilitate apprehension and memory, and not to give evidence.[13]

---

10 Joseph Priestley, *An Examination of Dr. Reid´s Inquiry* etc. p. 18ff.

11 *Intellectual Powers*, p. 349.

12 *Intellectual Powers*, p. 62, p. 457f. See also Callergård (2006) ch. 1, pp. 11ff for a discussion.

13 *Active Powers*, v, ii.

When Reid speaks about the principles of natural sciences it is to a great extent methodological principles that he is thinking of, such as to be found in the works of Bacon, Newton and the best scientists. Indeed, not even in Reid's famous epistemology of *first principles of common sense* is there anything to suggest that the principles to be settled will work as axioms or laws from which theorems, propositions and corollaries might be demonstrated. And even if Reid would think of these epistemic first principles as axioms from which singular everyday common sense beliefs might be derived, it is not clear that *that* derivation is part of the science of the mind. Indeed, the main business of the science of the mind seems to be to identify first principles, not to use them.[14]

2.3 Mind-matter analogies

Other features of physical theory that tempted Newtonian copycats were concepts of matter, motion, bodies, and forces. Reid's libertarianism about free will is explanation enough perhaps why he would resist any confounding between the realms of matter and mind, and between nomological necessity and freedom. His remarks against weak and bad analogies stand, however, independent of such worries. The problem with analogical reasoning is that the reliability of its conclusions depends on the similarity of the things compared. Reid therefore advices that analogy should be avoided in the study of mind and be replaced by careful introspective reflection, because no two kinds of phenomena seem to be more different than matter and mind.

Reflection is difficult however and that for two reasons. First, it is not an easy thing to attend to mental operations. Our mental capacities are suited to be used in dealing with everyday issues, such as external objects. The science of the mind is not a very natural pursuit for the mind. Secondly, however, even when we reflect successfully on mental operations, we are at loss to describe them accurately, that is, we do not have a scientifically suitable and established terminology by which to describe mental phenomena. What philosophers do in this situation, according to Reid, is to model their accounts of the mind on models of the behaviour of external bodies in motion. But this prejudice is not only due to sloppy reasoning and bad analogies. Philosophers share the prejudice with natural languages.

> Almost all the words, by which we express the operations of the mind, are borrowed from material objects. To understand, to conceive, to

14 *Inquiry*, p. 216. *Intellectual Powers*, p. 452ff.

> imagine, to comprehend, to deliberate, to infer, and many others, are words of this kind; so that the very language of mankind with regard to the operations of our minds, is analogical.[15]

All in all then, analogies are difficult to avoid and a menace to the science of the mind. Reid therefore finds ample ground for criticism of contemporary theorists of the mind whenever they are misled by the analogies of natural language or when they deliberately model mental phenomena similarly to our experience of material phenomena. In the conclusion of the *Inquiry* Reid highlights this feature of modern philosophy, which he traces from Descartes to Hume.

> They acknowledge that nature hath given us various simple ideas: These are analogous to the matter of Descartes' physical system. They acknowledge likewise a natural power by which ideas are compounded, disjoined, associated, compared: This is analogous to the original quantity of motion in Descartes' physical system. From these principles they attempt to explain the phaenomena of the human understanding, just as in the physical system the phaenomena of nature were to be explained by matter and motion.[16]

Consequently Reid attacked the conception of the mind as a *camera obscura*, as a *sensorium*, and as a container of ideas.[17] Similarly, Reid attacks *ideas* conceived as entities in their own right and as objects of thought.[18] By rejecting, finally, *association of ideas* as the fundamental operation of judgement and thinking the service of laws describing the general patterns of mental phenomena, such as Hume's three principles of association, will, quite understandably, not be much asked for in Reid's science of the mind.

3. Laws of nature

So far we have only shown that Reid must have conceived the general structure of mental phenomena to be quite different from the structure of

15 *Intellectual Powers*, p. 54f. See also *Inquiry*, p. 14 & 204f.

16 *Inquiry*, p. 212.

17 *Intellectual Powers*, p. 21f, p. 91f. *The Philosophical Orations of Thomas Reid*, ed. D.D. Todd, transl. by Shirley Darcus Sullivan, Southern Illinois University Press, Third Oration (1759), p. 61f.

18 See for instance *Inquiry*, chapter V and *Intellectual Powers*, Essay II.

physical phenomena. It may still be asked, however: does not laws or law-like statements make a salient part of Reid's science of the mind? If that is the case, this would certainly be ground for some affinity with Newtonian style study of the phenomena of nature.

To answer this question, let us first decide what is meant by law in the context of Reid's philosophy. If 'law' is taken in the general sense of 'principles' or 'original principles' – expressions that abound in Reid's writings – the answer must be yes, indeed. But such a broad notion of law does not justify a connection with Newton and early modern physics any more than a connection with the whole of the philosophical and scientific tradition since ancient times. Better therefore to stick to a notion of law that sits well with Newton's physics and Reid's understanding of laws of nature. Reid's concept of laws of nature is most easily characterized as "constant conjunctions" between events. He expresses appreciation for Hume's analysis of causation provided it is understood as an analysis of *physical causation* (and not of causation proper, i.e. the active power of agents). Constant conjunctions are general contingent empirical propositions that are either true or false. They do not express any necessary connection between cause and effect, and they do not reveal the efficient causes of change.[19]

With this notion of law, the answer to our initial question in this section will still be yes: laws do play a role in Reid's science of the mind. In the early pages of the *Inquiry* (1764) Reid had stated quite programmatically that:

> The man who first discovered that cold freezes water, and that heat turns it into vapour, proceeded on the same general principles, and in the same method, by which Newton discovered the law of gravitation and the properties of light. His *regulae philosophandi* are maxims of common sense, and are practised every day in common life; and he who philosophizes by other rules, either concerning the material system, or concerning the mind, mistakes his aim.[20]

And this view seems to have been reaffirmed some twenty years later:

19 I discuss this in section 3 of "Thomas Reid's Newtonian Theism: his differences with the classical arguments of Richard Bentley and William Whiston", *Studies in history and philosophy of science*, 41 (2010), pp. 109–119.

20 *Inquiry*, p. 12.

> The constitution of the human mind, and all that necessarily flows from its constitution, though it does not belong to what is now called *Natural* Philosophy, may justly be considered as part of the great volume of Nature. Being, therefore, the work of Nature, its powers, and faculties, their extent and limits, their growth and decline, and their connection with the state of the body, may, not improperly, be called phaenomena of Nature. And as far as these phaenomena can, by just induction, be reduced to general laws, such laws may properly be called laws of Nature.[21]

By such statements we should perhaps expect Reid to look for mental laws in his study of mind. The situation is, however, quite different. Reid's attempt to provide a more scientific and accurate account of the mind is not much concerned at all with laws, but rather with descriptions of the structure of a variety of mental operations. Specifically, he is concerned with phenomenological description of the components of mental operations, the specific type of objects mental acts are directed to, and the interrelations and dependencies in-between operations.[22] In addition he tries to identify the typical concepts and beliefs that specific operations evoke under normal circumstances, and to identify those first principles of common sense that are implicit in their role as sources of evidence.

Another thing to note is that Reid's science of the mind is a deliberately eclectic science. It welcomes evidence from any scientific discipline if it casts light on the nature and workings of the human mind.[23] His approach is a pioneering multi-discipline combination of (to use modern terms now) *cognitive science* and *philosophy of mind* combined with anything that *linguistic*, *physiological*, *anthropological* and *ethological* observations may

21 *Thomas Reid on the Animate Creation*, ed. Paul Wood, Edinburgh University Press: Edinburgh, (1995), p. 185.

22 Reid tends to speak of "ingredients" of mental operations when analysing them. The operation of perception for instance has the ingredients of a conception of a thing and the irresistible belief in its present existence. At times Reid uses the expression "constant concomitant" to picture how ingredients are related. See Callergård (2006) p. 61f.

23 This, I would suggest, is at least part of the meaning of the subtitle of his first book – "on the principles of common sense" – and this point of view is also displayed in the chapter "Principles taken for granted" in *Intellectual Powers*.

offer.[24] This also explains why Reid finds so many occasions to remind his readers of Newton's *Regulae Philosophandi*: As much as there are laws of nature to be discovered within the bounds of this eclectic science, and as much as, in Reid's view, there are many inferior theories put forward by theorists to explain mental phenomena (such as by Descartes, Locke, Hume, Hartley and Priestley) it is only to be expected that Reid feels that he must remind us of what is meant by explaining natural phenomena by laws of nature. These *regulae* encapsulate in his view a correct understanding of what search for laws of nature amounts to, and they will therefore be particularly appropriate in the criticism of inferior theories. This explains why Reid often comes back to the *Regulae* in his writings; more often, however, because they are needed in the criticism of theories, than because they are vehicles for positive discoveries of mental laws of nature.

It might be objected, however, that Reid in the midst of his science of the mind insists on and reports a type of mental law of nature of tremendous importance for his whole project, and that laws might in fact play an important systematic role in Reid's science of the mind. This is when Reid, in the course of his critical investigation of the theory of ideas, establishes that there is a law-like connection between certain sensations and the concepts and beliefs they evoke. Whenever we have a specific sensation, say of hardness, this immediately evokes the thought of the existence of a particular property of the body touched. The relation between the sensation of hardness and the concept and belief of hardness is, Reid claims, a law of nature. The sensation is a 'cause' and the conception and belief that arises is an 'effect'. The relation is a 'constant conjunction' and the truth of its holding is a contingent fact about the way we happen to be constructed. This is the way we happen to be "hard-wired" and it is conceivable at least that we might have been hard-wired differently, like the sensation of sweetness leading us to think of the property of the hardness of a body. It is not true then that the conception and belief of hardness is necessarily triggered by that particular sensation, since there might have been a different hard-wiring. The

24 The kind of *first philosophy* that Reid endorses, which consists in a program for establishing "first principles of common sense", should be fairly acceptable for *first philosophy* critic Quine, insofar as Reid's search for first principles of common sense is explicitly guided by considerations of a wide set of sciences such as logic, epistemology, linguistics, anthropology etc. Reid, like Quine, does not believe in an evidentially privileged point of view from which first philosophy can be established. See *Intellectual Powers*, pp. 459ff.

relation is a contingent constant conjunction and bears nothing of a necessary connection beyond that.

Human nature, then, consists of a lot of hard-wiring. The question is: Is it a primary aim for Reid to map the hard-wiring of our constitution, to describe law-like connections between particular sensations and corresponding conceptions of and beliefs about properties of bodies? Such mapping would indeed encourage the perspective of sorting out the relations in between these laws and to produce a theory of the mind the backbone of which would be the interrelations between laws of different generality. The answer is, I think, that Reid certainly would welcome any useful knowledge about this kind of hard-wiring, if it can be had. Here and there he points out law-like relations he thinks fundamental and worth noting in the course of his investigations. Sometimes it is because the issue at hand concerns psychophysical aspects of perception (such as the study of squinting, the parallel motion of the eyes, and double-seeing in chapter VI of the *Inquiry*). But often enough his primary concern is different. When Reid draws our attention to the law-like relation between the sensation of hardness and the correspondent conception hardness as a property of material bodies his interest resides wholly in the nature of the relation, which is a philosophical issue. This is because he believes that modern philosophers generally have misunderstood the nature of the connection, believing for instance that our conceptions are produced from, or explained by, or copied from, our sensory experiences. In addition, some theorists (Descartes, Hartley, Priestley) have, without solid evidence, put forward hypotheses to explain the efficient causal processes that leads us to form conceptions and beliefs about the properties of bodies. Reid's concern is about what sort of relation this is. He is not the least interested in making his discovery the starting point of an empirical research programme for establishing similar hard-wired connections of human nature. The empirical question about what connections of the sort there would be to map is not really on the table at all.

Although laws, as we have seen, do not serve as the backbone of Reid's science of the mind, he still reckoned mental phenomena to be largely speaking natural phenomena. That is, mental phenomena are part of created nature and should be studied as such according to the established methods of empirical investigation. Mental phenomena (or some subset of them) are therefore at least potentially susceptible to be explained or described in terms of laws of nature. There is, however, a more principled restriction to the usefulness of laws in science of the mind which has to do with Reid's

conviction that human beings have a capacity of freely exerting active power. In the two last sections of this paper I will shortly comment on this restriction, in the next section insofar as Reid discusses this himself, and in the last section I will suggest a specific reason, which Reid does not air himself explicitly, why his study of the human mind would never be much concerned with laws of nature at all.

4. Reid's later reconsideration of the relevance of the *Regulae* to the science of the mind

It is worth noting first that Reid never thought it necessary in his published writings to guard against his readers making the mistake of thinking that he was in the pursuit of trying to reduce all mental phenomena to law-like connections, the way a stout materialist or a necessitarian would do. It was only when he was prompted by Priestley's materialist and necessitarian conception of human nature, which was supported by what Reid took to be a serious misinterpretation of the *Regulae Philosophandi*, that Reid had to reconsider the relation between the mere search for laws of nature and the study of mind.[25] As we have already seen, he reaffirmed that mental phenomena are natural phenomena, adding, as we also saw, that "as far as these phaenomena can, by just induction, be reduced to general laws, such laws may properly be called laws of Nature". Clearly then, there is a recognition here that some mental phenomena is not susceptible to such treatment, and what Reid specifically had in mind was volition. He had to reconsider the range of the *Regulae* and the subject matter of the science of the mind:

> Whether Sir Isaac Newton, in his rules of Philosophizing, had in his view the natural phaenomena of the mind, or not, does not appear; but, it is evident, that the reason of them extends to these, as well as to the phaenomena of the material system; and therefore they may be applied to both with equal propriety, and ought to be adhered to with equal strictness. But it is to be observed, that the voluntary actions of men can in no case be called natural phaenomena, or be considered as

25 Reid´s writings on Priestley is found in some manuscripts titled "Some Observations on the Modern System of Materialism" which are now to be found in *Thomas Reid on the Animate Creation*, ed. Paul Wood, Edinburgh University Press: Edinburgh, 1995, pp. 173ff.

> regulated by the physical laws of Nature. Our voluntary actions are subjected to moral, but not physical laws.[26]

And a little later he writes,

> There are many important branches of human knowledge, to which Sir Isaac Newton's rules of Philosophizing have no relation, and to which they can with no propriety be applied. Such are Morals, Jurisprudence, Natural Theology, and the abstract Sciences of Mathematics and Metaphysics; because in none of those Sciences do we investigate the physical laws of Nature. There is therefore no reason to regret that these branches of knowledge have been pursued without regard to them.[27]

Apparently then, the restriction imposed on the applicability of the *Regulae* on mental phenomena implies that sometimes they apply, sometimes they don't, depending on the specific issue at hand. Should we conclude that the eclectic field of study called the science of the human mind is such that whenever there are some mental phenomena that can be traced to a general law-like connection the *Regulae* will be relevant, while in other parts they will not. Does the science of the mind has distinguishable parts some of which are nomological in character, and some of which are not? Indeed, does the mind itself has distinguishable parts, some of which run regularly according to laws and others in which human volition breaks in? The only sure thing to say so far is that laws of nature has a considerably smaller role to play in the science of the mind compared with the role played in such sciences in which laws serve as the backbone of a theoretical structure.[28]

---

26 *Thomas Reid on the Animate Creation,* p. 185.

27 *Thomas Reid on the Animate Creation*, p. 185f.

28 I disagree therefore with Wolterstorff (2001) and with Copenhaver (2006). They take for granted that laws of nature is the central piece of Reid's science of the mind. And they differ in their interpretations of what extent we are able to penetrate into the constitution of the mind understood as a basic set of laws. I agree with Rebecca Copenhaver that laws of nature in the science of the mind are no different from laws in physics and that, therefore, there is no reason to think that there are any particular 'mysteries' involved in our comprehension of laws of the mind, such as has been suggested by Nicholas Wolterstorff. The methodology and epistemology of nomological investigations are exactly the same. I suppose that I differ from Copenhaver in thinking that such investigations has a small part to play in Reid´s science of the mind. See Rebecca Copenhaver, "Is Reid a mysterian?", in *Journal of the History of Philosophy*, vol. 44,

5. The active mind: attention and the rearing of faculties

That a science of the mind is no easy thing to accomplish successfully everyone can agree on. But where in lies the chief difficulty? According to Hume the major difficulty – to only one which he cares to mention in the *introduction* to the *Treatise* – is that the experiments that the researcher wishes to conduct on his own mind will be disturbed by the presence of the mind of the experimentalist himself.[29] Quite undisturbed by this concern Hume finds it easy enough to distinguish the basic furniture of the mind: On the first page he establishes that they are all perceptions of the mind and that they are all either impressions or ideas. It takes Hume another two or three pages to find out his first law, the copy principle, and yet another page or two to come up with three principles of association that will do most of the explanatory work throughout the *Treatise*. Reid in contrast discusses methodological issues in all his three books. If Reid, as we saw earlier, had complained that philosophers had pictured mental phenomena to be as conceptually simple as matter in motion governed by laws of nature, his description of whatever he thinks encounters anyone who wishes to explore the human mind shows a more complicated subject matter to work on. Mental operations, he writes in the *Inquiry*...

> ...are so mixed, compounded, and decompounded, by habits, associations, and abstractions, that it is hard to know what they were originally.[30]

> it is extremely difficult for the mind to return upon its own footsteps, and trace back those operations which have employed it since it first began to think and act.[31]

Now, this might of course merely be the bad luck of the scientist of the mind compared with the more lucky astronomer or zoologist, whose objects of study are easily identifiable; no reason to reject the grand idea of a fully nomological account of mental phenomena (with due considerations, of

---

no. 3 (2006), 449-466, and, Nicholas Wolterstorff, *Thomas Reid and the story of epistemology*, Cambridge university press: Cambridge, 2001.

29 David Hume, *A Treatise of Human Nature*, bk. 1, introduction, p. 6.

30 *Inquiry*, p. 14.

31 *Inquiry*, p. 15.

course, made for the irregularities of the liberty of the will). I hope to indicate in this last section that this is not just a practical complication, like an entangled yarn that just happens to be somewhat hard to disentangle. The crux of the matter is that there is a particular entanglement that creeps in almost everywhere in our mental operations, namely, our active engagement in these operations; they are not events that merely happen to us, instead, we engage in them. A stark contrast can be made here between Hume's model of the mind and Reid's. In Hume's model what the model describes is what happens to a subject under certain circumstances, and the explanations sought for are laws of mental phenomena. Hume minimizes the role of the agent in accounting for the mental phenomena. Reid shares the scientific aim of exploring the basic laws, principles and dependencies which accurately describe and explain mental phenomena. In addition, however, he must also take into account the fact that we participate in our mental life as agents. Indeed, because his scientific aim is to describe the human mind accurately, he cannot leave out this feature of our mental life. Let's see how this works by way of two examples.

5.1 'Active' and 'Intellectual' powers – the case of attention

In the introduction to the *Essays on the Active Powers of the Human Mind* Reid writes that "it is evidently the intention of our Maker, that man should be an active and not merely a speculative being" and that we for this reason have been endowed with active power.

> Our business is to manage these powers, by proposing to ourselves the best ends, planning the most proper systems of conduct that is in our power, and executing it with industry and zeal. This is true wisdom; this is the very intention of our being.[32]

There is nothing to indicate, as far as I have seen, that what Reid has in mind here is only our visible, physical and publicly accessible actions in society, as opposed to a purely speculative philosophical being. Indeed, it turns out that the very distinction that separates Reid's two major *Essays* – the *Intellectual* and the *Active* powers of the mind – is more conventional than real:

> The faculties of the understanding and will are easily distinguished in thought, but very rarely, if ever, disjoined in operation. In most,

32 *Active Powers*, p. 5.

> perhaps all the operations of mind for which we have names in language, both faculties are employed, and we are both intellective and active.[33]

Here, I suggest, is the crux of the matter. There is no way the scientist of the mind can pretend we are not there ourselves as agents among the mental phenomena to be studied. Hume's worries about the experimenter influencing the experiment turns out to be, from a Reidian point of view, wholly misconceived. So called mental phenomena are to a too great extent a matter of us performing mental actions.

Reid goes on to single out *attention*, *deliberation*, and what he calls *a fixed purpose or resolution* as operations which are commonly classed under the intellect but which, in his view, involve the will and therefore might as well be classed under our active powers.[34] The case of *attention* is of particular interest because Reid highlights the importance of this ability for all thinking and acting to the extent of claiming that "a great part of wisdom and virtue consists in giving a proper direction to our attention."[35] A closer look on this aspect of Reid's philosophy of mind reveals that just to think about something and to think about something as something requires the exercise of attention.

> … so great is the effect of attention, that, without it, it is impossible to acquire or retain a distinct notion of any object of thought.[36]

This has far-reaching consequences for the very character of Reid's science of the mind. According to Reid, to perceive, to remember and to conceive is to be mentally directed to some item which is distinct from the operation at

33 *Active Powers*, p. 59f.

34 Reid opposes the view of *Deliberation* according to which it is a quasi-mechanical process where different motives struggle and the strongest prevails. What decides a conflict in between motives is a judgment made by a rational agent that weighs the importance of different principles and motives. See *Active Powers*, p. 216f.

A *fixed purpose or resolution with regard to our future conduct* is, for instance, when you decide to always be a good person, or when you decide to go to London next winter. Such resolutions are exercises of will and they last over long time. Both deliberation and resolutions can clearly have mental actions as their objects, as for instance when weighing evidence, and when resolving always to go by clear and distinct ideas in scientific matters.

35 *Active Powers,* p. 63.

36 *Active Powers,* p. 60.

hand, such as a material object, a past event, or a conception. In doing so we are *ipso facto* attending to the object at hand, and to do so successfully we need to exert our will. Because acts of attention pervade our mental life the subject matter of Reid's science of the mind is quite different from Hume's nomological approach.

Our next example concerns cases in which we are engaged with trying to improve our inborn mental capacities.

5.2 Rearing our faculties

In the *Inquiry* Reid writes:

> Of the various powers and faculties we possess, there are some which nature seems both to have planted and reared, so as to have left nothing to human industry. Such are the powers which we have in common with the brutes, and which are necessary to the preservation of the individual, or the continuance of the kind. There are other powers, of which nature hath only planted the seeds in our minds, but hath left the rearing of them to human culture. It is by the proper culture of these, that we are capable of all those improvements in intellectuals, in taste, and in morals, which exalt and dignify human nature; while, on the other hand, the neglect or perversion of them makes its degeneracy and corruption.[37]

Reid uses this observation to insist on the difficulties involved in the study of the human mind. It gives expression at the same time of a view of the human mind as something intrinsically active. What we are and become depend on what we do with our faculties, our basic mental equipment. The ability to rear whatever we got from start does not mean that we can change our constitution or our nature, but it means that the basic powers of our constitution is there for us to be used the best we can, and this includes being developed and cultivated. Reid uses an agricultural metaphor.

> The earth is left by nature in such a state as to require cultivation for the accommodation of man. […] By clearing, tilling and manuring the ground, by planting and sowing, by building cities and harbours, draining marshes and lakes, making rivers navigable, and joining them by canals, by manufacturing the rude materials which the earth, duly

37 *Inquiry*, p. 13.

> cultivated, produces in abundance, by the mutual exchange of commodities and of labour, he may make the barren wilderness the habitation of rich and populous states.[38]

It quickly turns out, however, that we receive our higher mental faculties pretty much in a similar "rude and barren" state.

> His animal faculties are sufficient for the preservation of the species; they grow up of themselves, like the trees of the forest, which require only the force of nature and the influence of Heaven. His rational and moral faculties, like the earth itself, are rude and barren by nature, but capable of a high degree of culture; and this culture he must receive from parents, from instructors, from those with whom he lives in society, joined with his own industry.[39]

Now, as much as it would be strange to describe the human species without any mention of our capacity to work on and change our physical environments and conditions of life, it would be a serious omission in a science dealing with our mental equipment to leave out our ability to relate to and developing this equipment. Indeed, to make the best of our inborn "rude and barren" capacities is our duty and is, as we saw before, what Reid calls true wisdom.

All things considered, mental operations are not, on Reid's view, events that merely happens to us. They are also events in which we engage actively. This explains why Reid's science of the mind is not a characteristically nomological science, and therefore also, on account of this and other considerations we have made, not particularly "Newtonian".

References

Broadie, Alexander (2004). Reid in context. In *The Cambridge Companion to Thomas Reid*, Cambridge University Press: Cambridge.

Callergård, Robert (2006). *An Essay on Thomas Reid's Philosophy of Science*, Acta Universitatis Stockholmiensis: Stockholm.

38 *Active Powers*, p. 43.

39 *Active Powers*, p. 43.

Callergård, Robert (2010). Thomas Reid's Newtonian Theism: his Differences with the classical Arguments of Richard Bentley and William Whiston. In *Studies in History and Philosophy of Science* 41, pp.109–119.

Copenhaver, Rebecca (2006). Is Reid a mysterian? *Journal of the History of Philosophy*, vol. 44, no. 3, 449–66.

Harris, James (2003). Reid on the Character of a Science of the Mind. unpublished paper read at a Boston conference in 2003.

Hume, David (2010 [1739–40]). *A Treatise of Human Nature*, ed. Norton & Norton, Oxford University Press: Oxford.

Hutcheson, Francis (2004 [1726]). *An Inquiry into the Original of our Ideas of Beauty and Virtue*, Liberty fund inc.: Indianapolis.

Newton, Isaac (1979 [1704]). *Opticks*, Dover Publications: New York.

Newton, Isaac (1999 [1726, 3rd ed.]). *Mathematical Principles of Natural Philosophy*, (transl. I.B. Cohen and A. Whitman), University of California press: Berkeley.

Priestley, Joseph (1775). *An Examination of Dr. Reid's Inquiry*...[etc.], 2nd, ed., London.

Reid, Thomas (1989). *The Philosophical Orations of Thomas Reid*, ed. D.D. Todd, transl. by Shirley Darcus Sullivan, Southern Illinois University Press.

Reid, Thomas (1997 [1764]). *An Inquiry into the Human Mind on the Principles of Common Sense*, ed. Derek. R. Brookes, Edinburgh University Press: Edinburgh.

Reid, Thomas (2002 [1785]). *Essays on the Intellectual Powers of Man*, ed. Derek R. Brookes, Edinburgh University Press: Edinburgh.

Reid, Thomas (1995). *Reid on the Animate Creation*, ed. Paul Wood, Edinburgh University press: Edinburgh.

Reid, Thomas (2010 [1788]). *Essays on the Active Powers of Man*, ed. Knud Haakonssen and James A. Harris, Edinburgh University Press: Edinburgh.

Reid, Thomas (2002). *The Correspondence of Thomas Reid*, ed. Paul Wood, Edinburgh University Press: Edinburgh.

Reid, Thomas (1748). An Essay on Quantity. *Transactions of the Royal Society of London*, vol. XIV.

Turnbull, George (1740 [2005]). *Principles of Moral Philosophy*, ed. Alexander Broadie, Liberty fund inc.: Indianapolis.

Wood, Paul B. (1993) *The Aberdeen Enlightenment – the arts curriculum in the eighteenth century*, Aberdeen university Press: Aberdeen.

Wolterstorff, Nicholas (2001). *Thomas Reid and the Story of Epistemology*, Cambridge university press: Cambridge.
Yaffe, Gideon (2004). *Manifest Activity – Thomas Reid's Theory of Action*, Oxford University Press, Oxford.

Robert Callergård
Department of Philosophy
Uppsala university
robert.callergard@filosofi.uu.se

Den Gyllene Regeln och Substitutionsfunktioner

Daniel Rönnedal

Abstrakt
Enligt den s.k. gyllene regeln bör vi behandla andra så som vi själva vill bli behandlade. Det här är en princip som kan uttryckas och tolkas på många olika sätt. ”*Allt* ni vill att andra gör för er det skall ni också göra för dem” är en annan formulering. Denna sats antyder att regeln *kan* förstås på ett sådant sätt att den handlar om *alla* (typer av) handlingar. Men vad innebär det? I den här uppsatsen undersöker jag några olika preciseringar. Jag går igenom hur man kan förstå uttrycket ”alla handlingar” med hjälp av substitutionsfunktioner som ersätter handlingspredikat med välformade formler. Jag visar hur man med denna tolkning som utgångspunkt kan härleda en mängd normer som alla tycks följa ur den gyllene regeln. Enligt denna läsning är principen potentiellt mycket kraftfull och användbar. Tolkningen leder emellertid även till vissa tekniska problem som talar för att den är alltför stark. Jag visar därefter hur dessa problem kan lösas.

1. Introduktion

Enligt den s.k. gyllene regeln bör vi behandla andra så som vi själva vill bli behandlade. Den gyllene regeln är en av världens mest spridda, historiskt inflytelserika och accepterade normer. Alla världsreligioner, och många andra religioner, tycks innehålla någon variant av denna regel[1], och mängder av filosofer har också accepterat den i en eller annan form[2]. Det är emellertid inte alls uppenbart hur denna norm bäst uttrycks och tolkas. Det finns en mängd olika formuleringar av den gyllene regeln och varje uttryck i denna princip kan preciseras på många olika sätt.[3] Detta leder till bokstavligt talat tusentals olika möjliga tolkningar av denna välkända princip. Jag kommer i den här uppsatsen att undersöka *en* av de många frågor som uppstår då man

[1] Neusner och Chilton (red.) (2008).
[2] Se t.ex. Hobbes (1985), Kapitel XIV, s. 190, Pufendorf (1964), Bok 2, 3:13, Mill (1987), Kapitel 2, s. 28, och Gensler (1996), särskilt Kapitel 5.
[3] Rönnedal (2015) tar upp ett antal frågor som är relevanta då man försöker förstå den gyllene regeln. Se också Rönnedal (2016).

försöker förstå denna regel, nämligen huruvida den handlar om *alla* (typer av) handlingar eller bara *vissa* (typer av) handlingar. Uppsatsens huvudsyfte är att försöka förstå den logiska formen hos den gyllene regeln, inte att diskutera olika argument för eller emot denna princip. Personligen tror jag att det finns tolkningar av den gyllene regeln som är rimliga, även om det också finns många preciseringar som har problematiska konsekvenser. Oavsett om man är intresserad av att försvara eller kritiskt granska denna regel eller någon särskild precisering, bör man försöka förstå vad den innebär och fokusera på de bästa möjliga tolkningarna.[4]

Uppsatsen är indelad i sju avsnitt. Avsnitt 2 tar upp några interpretationer av den gyllene regeln och nämner ett antal slutsatser som tycks följa ur denna princip. Avsnitt 3 handlar om substitutionsfunktioner. I Avsnitt 4 visar jag hur teorin om substitutionsfunktioner kan användas för att härleda alla de satser som introduceras i Avsnitt 2 och som intuitivt följer ur GR. Avsnitt 5 innehåller ett problem för GR4, en viss tolkning av GR, som innebär att man kan härleda ett antal kontraintuitiva satser om man antar att man kan tillämpa vilka substitutionsfunktioner som helst på denna regel. Avsnitt 6 tar upp några möjliga lösningar på problemet ifråga. Jag argumenterar för att man kan undvika den aktuella svårigheten om man begränsar substitutions-funktionernas värden till ”universella” predikat som inte innehåller några individkonstanter. Avsnitt 7 sammanfattar uppsatsen.

2. Tolkningar av den gyllene regeln

Enligt den kanske enklaste formuleringen av den gyllene regeln säger denna princip att om du vill att en individ x utför handling H mot dig, så bör du utföra H mot x. Och om vi generaliserar detta får vi följande resultat.

> **(GR)**. Det gäller för alla individer x och y att: Om x vill att y utför handling H mot x, så bör x utföra H mot y.

[4] För mer historisk information om den gyllene regeln, se t.ex. Wattles (1996) och Gensler (2013), Kapitel 5. För en diskussion om den gyllene regelns förhållande till olika religioner, se Neusner och Chilton (red.) (2008). Filosofiska introduktioner till den gyllene regeln finner man bl.a. i Carson (2010), Kapitel 6, Carson (2013), Gensler (1996), särskilt Kapitel 5, och Gensler (2013). För mer information om den gyllene regeln se t.ex. Blackstone (1965), Bruton (2004), Cadoux (1912), Duxbury (2009), Gensler (1986), (2013), Gewirth (1978), Gould (1980), Hare (1963), Hertzler (1934), Hirst (1934), Hoche (1978), Huang (2005), Reinikainen (2005), Rönnedal (2015), Singer (1963), Wattles (1996) och Weiss (1941).

Detta ger oss dock inte i sig något svar på vad ”H” står för. GR formuleras ibland på följande sätt. *Allt* ni vill att andra gör för er bör ni också göra för dem. Denna formulering antyder att GR *kan* och kanske också *bör* tolkas på ett sådant sätt att den handlar om ”alla” (typer av) handlingar. Då skulle vi kunna uttrycka GR på följande sätt.

> **(GRH)**. Det gäller för alla handlingar H och individer x och y att: Om x vill att y utför handling H mot x, så bör x utföra H mot y.

Men vad innebär det? Och vad menas med ”alla handlingar” i detta sammanhang? Inbegriper det alla handlingar över huvud taget eller alla handlingar av en viss typ? Låt oss nämna fyra av de intressantaste tolkningarna av detta uttryck.

Tolkning 1. ”alla handlingar” betyder alla handlingar som kan beskrivas med enkla handlingspredikat. Gör vi denna tolkning får vi följande variant av GR.

> **(GR1)**. Det gäller för alla individer x och y att: Om x vill att y utför handling H mot x, så bör x utföra H mot y, där ”H” står för en handling som kan beskrivas med ett enkelt handlingspredikat.

Tolkar vi GR på detta sätt tycks regeln medföra bl.a. följande satser.

> (K1) Om du vill att din vän är ärlig mot dig, så bör du vara ärlig mot din vän.
>
> (K2) Om du vill att din partner är trogen mot dig, så bör du vara trogen mot din partner.
>
> (K3) Om du vill att denna främling håller sina löften till dig, så bör du hålla dina löften till denna främling.

Tolkning 2. ”alla handlingar” betyder alla handlingar som kan beskrivas med enkla handlingspredikat eller negationer av enkla handlingspredikat. Givet denna läsning, får vi följande variant av GR.

> **(GR2)**. Det gäller för alla individer x och y att: Om x vill att y utför handling H mot x, så bör x utföra H mot y, där ”H” står för en handling som kan beskrivas med ett enkelt handlingspredikat eller en negation av ett enkelt handlingspredikat.

Tolkar vi GR på detta sätt tycks regeln medföra den s.k. silverregeln eller den ”negativa” formen av GR. Silverregeln kan uttryckas på följande sätt:

> **(SR)**. Det gäller för alla individer x och y att: Om x vill att y *inte* utför handling H mot x, så bör x *inte* utföra H mot y.

Dessutom tycks bl.a. följande satser vara konsekvenser av GR enligt denna tolkning.

> (K4) Om du vill att din granne *inte* stjäl från dig, så bör du *inte* stjäla från din granne.
> (K5) Om du vill att din arbetskamrat *inte* ljuger för dig, så bör du *inte* ljuga för din arbetskamrat.
> (K6) Om du vill att din ovän *inte* dödar dig, så bör du *inte* döda din ovän.

Tolkning 3. ”alla handlingar” betyder alla handlingar som kan beskrivas med enkla handlingspredikat, negationer av enkla handlingspredikat eller sådana handlingspredikat kopplade till ett villkor. Läser vi ”alla handlingar” på detta sätt, får vi följande variant av GR.

> **(GR3)**. Det gäller för alla individer x och y att: Om x vill att y utför handling H mot x, så bör x utföra H mot y, där ”H” står för en handling som kan uttryckas med ett enkelt handlingspredikat, en negation av ett enkelt handlingspredikat eller ett handlingspredikat av sådant slag kopplat till ett villkor.

Tolkar vi GR på detta sätt tycks regeln medföra bl.a. följande satser.

> (K7) Om du vill att din bror hjälper dig om du behöver hjälp, så bör du hjälpa din bror om han behöver hjälp. (Om du vill att om du behöver hjälp så hjälper din bror dig, så bör det vara fallet att om din bror behöver hjälp så hjälper du honom.)
> (K8) Om du vill att din syster tackar dig om du hjälper henne, så bör du tacka din syster om hon hjälper dig. (Om du vill att om du hjälper din syster så tackar hon dig, så bör det vara fallet att om din syster hjälper dig, så tackar du henne.)

(K9) Om du vill att denna familjemedlem ber om ursäkt till dig om hon har svikit dig, så bör du be om ursäkt till denna familjemedlem om du har svikit henne. (Om du vill att om denna familjemedlem har svikit dig så ber hon om ursäkt till dig, så bör det vara fallet att om du har svikit denna familjemedlem så ber du om ursäkt till henne.)

Tolkning 4. GR talar enligt denna tolkning om alla handlingar över huvud taget (oavsett vilka handlingspredikat vi använder för att beskriva våra handlingar). Dvs. GR tolkas på följande sätt.

(GR4). Det gäller för alla individer x och y att: Om x vill att y utför handling H mot x, så bör x utföra H mot y, där ”H” står för en handling som kan uttryckas med vilket handlingspredikat som helst.

Med denna tolkning tycks alla satserna (K1)–(K9) följa ur GR. Dessutom tycks t.ex. följande sats vara en konsekvens av GR4.

(K10) Om du vill att x behandlar alla dina barn med respekt, så bör du behandla alla x’s barn med respekt.

Låt mig nämna ytterligare några möjliga tolkningar. Alla varianter av GR ovan kan begränsas på så sätt att de endast handlar om vissa (typer av) handlingar, som t.ex. kan definieras genom en uppräkning. Den första varianten av GR skulle då kunna uttryckas på följande sätt.

(GR1′). Det gäller för alla individer x och y att: Om x vill att y utför handling H mot x, så bör x utföra H mot y, där ”H” står för handlingen (i) …, (ii) … osv.

Där ”(i) …, (ii) … osv.” fylls i med en lista på de (typer av) handlingar man antar att GR uttalar sig om, t.ex. ”(i) att tala sanning, (ii) att hålla sina löften… osv.”. Silverregeln skulle kunna formuleras på följande sätt givet denna tolkning.

(SR′). Det gäller för alla individer x och y att: Om x vill att y inte utför handling H mot x, så bör x inte utföra H mot y, där H står för handlingen (i) …, (ii) …, (iii) … osv.

Där ”(i) …, (ii) …, (iii) … osv.” återigen fylls i med en lista på de (typer av) handlingar man antar att SR uttalar sig om, t.ex. ”(i) att ljuga, (ii) att stjäla, (iii) att döda… osv.”.

I de ovanstående tolkningarna av GR har vi explicit talat om ”handlingar”. Men GR kan också tolkas på sådant sätt att den inbegriper t.ex. attityder, känslor och förhållningssätt, förutom handlingar i en mer snäv mening. En konsekvens av GR tolkad på detta sätt skulle t.ex. kunna vara följande. Om du vill att din arbetskamrat *känner uppskattning* när du har hjälpt henne, så bör du *känna uppskattning* när din arbetskamrat har hjälpt dig.

Regeln kan också tolkas på ett sådant sätt att den även handlar om *hur* vi vill bli behandlade av andra och *hur* vi bör behandla andra. Det tycks ofta vara fallet att vi bryr oss om *hur* andra behandlar oss och inte bara *att* de utför en viss (typ av) handling. En konsekvens av GR tolkad på detta sätt skulle t.ex. kunna vara följande. Om du vill att x håller sitt löfte till dig *utan att klaga och vara allmänt otrevlig*, så bör du hålla ditt löfte till x *utan att klaga och vara allmänt otrevlig*.

Alla varianter av GR ovan ((GR1)–(GR4)) kan tolkas brett, så att de inkluderar olika (typer av) handlingar, attityder, känslor, förhållningssätt osv. Men de kan också begränsas till uttalanden om ”handlingar” i en mer snäv mening.

Tolkning 4 ger oss den starkaste formen av GR, tolkning 1 den svagaste. GR4 medför GR3, som medför GR2, som medför GR1. I en viss mening är alltså GR4 den intressantaste tolkningen. Inom vetenskapsteorin brukar man ofta betrakta en mer generell teori som bättre än en mindre generell teori (allt annat lika). Samma sak tycks gälla för normativa teorier. Men det finns också risk för att starka teorier blir alltför starka, så att vi kan härleda orimliga konsekvenser. Därför kan det vara rimligt att börja med att undersöka så starka teorier som möjligt och sedan i ljuset av eventuella kontraintuitiva slutsatser revidera dem. I en viss mening är det därför rimligast att börja med att undersöka tolkning 4.

Enligt tolkning 4 tycks alltså alla satserna (K1)–(K10) följa ur den gyllene regeln. Men *hur*? Det är givetvis enkelt att *påstå* att de följer. Men kan vi *bevisa* detta?

Kantianer hävdar ofta att det följer ur det kategoriska imperativet att vi bör hålla våra löften. Men det är ofta oklart exakt hur detta antas följa. De argument som åberopas är ofta informella. Och det är inte uppenbart om de

kan omformuleras till deduktivt giltiga argument. Gäller detta även den typ av ”konsekvenser” av GR som vi har nämnt ovan?

Om vi använder tolkning 4, kan den logiska formen hos GR anges på följande sätt:

(FGRH) $\forall H \forall x \forall y(V_x Hyx \rightarrow OHxy)$.

Denna sats läses ”Det gäller för alla handlingar H och alla individer x och y att: Om x vill att y utför handling H mot x, så bör x utföra handling H mot y”. Vi kvantifierar här explicit över handlingar. En alternativ formulering är den följande:

(FGRS) $\forall x \forall y(V_x Hyx \rightarrow OHxy)$,

där H tolkas schematiskt, som en variabel som kan bytas ut mot vilket predikat som helst. I praktiken säger detta schema samma sak som $\forall H \forall x \forall y(V_x Hyx \rightarrow OHxy)$.

Använder vi någon av dessa symboliseringar följer alla satserna (K1)–(K10), eller – mer precist – adekvata formaliseringar av dessa satser. Gör vi det uppstår emellertid en del problem, bl.a. samma problem som om vi antar att vi kan ersätta handlingspredikatet H i den gyllene regeln med vilken formel som helst (se Avsnitt 5 nedan). Dessutom krävs det, åtminstone för den första symboliseringen, att vi utvecklar en högre ordningens logik. Det är intressant att se om vi kan undvika det.

Vi skall i den här uppsatsen istället visa hur alla satserna (K1)–(K10) kan härledas med hjälp av substitutionsfunktioner från handlingspredikatet H till (godtyckliga) formler. Närmare bestämt kommer vi att begränsa oss till substitution av formler som inte innehåller några individkonstanter för att undvika det problem med GR4 som diskuteras i Avsnitt 5. Det här leder till en tolkning av den gyllene regeln, nämligen GR5, som är något svagare än GR4. GR4 medför denna tolkning, men inte tvärtom. För att förstå detta lösningsförslag måste vi först veta lite mer om substitutionsfunktioner. Vi börjar gå igenom den nödvändiga teorin i nästa avsnitt. Den ”teori” för dessa funktioner vi skall använda har utvecklats av S. C. Kleene (1971) och Gerhard Schurz (1997). Vi följer Schurz framställning nästan ordagrant, men vi kommer att göra några kompletteringar.

3. Substitutionsfunktioner

För att kunna uttrycka oss exakt kommer vi att utveckla ett formellt språk, som vi sedan använder för att formalisera den gyllene regeln och härleda satserna (K1)–(K10).

Syntax

Konventioner

σ och π denoterar substitutionsfunktioner för icke-logiska predikat, den första ersätter predikat med formler (”komplexa” predikat), den andra ersätter predikat med predikat. Notera att σ tar den formel som förekommer omedelbart till höger som argument. $\sigma Fu_1 \ldots u_n$ står t.ex. för $\sigma(Fu_1 \ldots u_n)$, och $\sigma A \rightarrow B = \sigma(A) \rightarrow B$, som inte är identisk med $\sigma(A \rightarrow B) = \sigma(A) \rightarrow \sigma(B) = \sigma A \rightarrow \sigma B$. Om vi vill uttrycka att σ refererar till ett predikat, kan vi skriva σF, σ(F), $(\sigma F)[u_{1\text{-}n}/x_{1\text{-}n}]$ eller $(\sigma(F))[u_{1\text{-}n}/x_{1\text{-}n}]$.

$\mathcal{L}$, $\mathcal{V}$, $\mathcal{C}$, $\mathcal{T}$, $\mathcal{R}$… denoterar mängden av alla formler, variabler, konstanter, termer, respektive relationer. A, B, C… betecknar godtyckliga formler.

Schurz använder x, y, z inte endast som individvariabler i objektspråket utan också som metavariabler i metaspråket, som varierar över alla individvariabler; vi skall också låta s, t, u, v generellt variera över termer. Detsamma gäller predikatvariabler F, G, H… och satsvariabler p, q, r… Schurz använder även objektspråkets satslogiska konnektiv och kvantifikatorer i metaspråket. ”:=” står för identitet per definition. **N** står för mängden av alla naturliga tal. Vi använder samma konventioner i den här uppsatsen.

Vokabulär

Schurz (1997, s. 34) introducerar endast en mängd individvariabler utan en extra mängd individkonstanter. Han klarar sig därmed utan distinktionen mellan öppna och slutna formler. Givet de vanliga distinktionerna mellan fria och bundna variabler, spelar de fria individvariablerna i Schurz system samma roll som individkonstanter i språk med en distinktion mellan individvariabler och individkonstanter. Det grundläggande språk Schurz beskriver innehåller följande vokabulär.

(1) en uppräkningsbart oändlig mängd $\mathcal{V}$ av individvariabler u, v, ... , x, y, z (möjligtvis med index).

(2) För varje n ≥ 0, en uppräkningsbart oändlig mängd $\mathcal{R}^n$ av n-ställiga predikat F, G, H … (möjligtvis med index); i synnerhet gäller det att $\mathcal{P} := \mathcal{R}^0$ står för mängden av satsvariabler p, q, … (noll-ställiga predikat). $\mathcal{R} := \cup_{n \geq 0} \mathcal{R}^n$ denoterar mängden av alla predikat. …

(3) De logiska symbolerna (konnektiven och operatorerna) ¬ (negation), ∨ (disjunktion), ∀ (universell kvantifikator), □ (aletisk nödvändighets operator), O (deontisk plikt operator), och parenteser), och (.

Vi skall emellertid utvidga detta alfabet på två sätt. Vi introducerar

(4) en uppräkningsbart oändlig mängd $\mathcal{C}$ av individkonstanter a, b, c ... (möjligtvis med index); och

(5) en (logisk) satsoperator V (en viljeoperator).

Anledningen till detta är att vi vill kunna referera till specifika individer med hjälp av en mängd konstanter och att vi vill kunna uttrycka att en eller flera individer *vill* någonting. Variablerna och konstanterna kallas tillsammans för ”termer”. Mängden av alla termer betecknas $\mathcal{T}$.

Definition 1 (Icke primitiva symboler). Symbolerna → (materiell implikation), ∧ (konjunktion), ↔ (materiell ekvivalens), ⊤ (Verum), ⊥ (Falsum), ∃ (existens kvantifikator), ◇ (aletisk möjlighets operator) och P (deontisk tillåtelse operator) definieras på vanligt sätt. (A → B) := (¬A ∨ B), (A ∧ B) := ¬(¬A ∨ ¬B), (A ↔ B) := ((A → B) ∧ (B → A)), ⊤ := (p ∨ ¬p), ⊥ := ¬⊤, ∃xA:= ¬∀x¬A, ◇A:= ¬□¬A, PA:= ¬O¬A. Notera att varje individvariabel som är bunden räknas som en logisk symbol, medan fria individvariabler och predikat räknas som icke-logiska symboler. ■

Språk

Schurz (1997, ss. 34–35) identifierar ett språk $\mathcal{L}$ med mängden av alla formler som genereras av följande regler:

(i) $F \in \mathcal{R}^n, x_1, \ldots, x_n \in \mathcal{V} \Rightarrow Fx_1 \ldots x_n \in \mathcal{L}$ (de atomära formlerna);
(ii) $A, B \in \mathcal{L} \Rightarrow \neg A, (A \vee B), \Box A, OA \in \mathcal{L}$;
(iii) $A \in \mathcal{L}, x \in \mathcal{V} \Rightarrow \forall xA \in \mathcal{L}$; och
(iv) Ingenting annat är en formel.

Vi skall emellertid modifiera dessa regler på följande sätt. Vi byter ut (i) mot (i′), (iv) mot (iv′), och lägger till (v).

(i′) $F \in \mathcal{R}^n$, $t_1, \ldots, t_n \in \mathcal{T} \Rightarrow Ft_1 \ldots t_n \in \mathcal{L}$ (de atomära formlerna);
(iv′) $A \in \mathcal{L}$, $t \in \mathcal{T} \Rightarrow V_t A \in \mathcal{L}$;
(v) Ingenting annat är en formel.

Kommentar 2. De satslogiska konnektiven läses på vanligt vis. $\Box A$ läses ”Det är nödvändigt att A”, OA ”Det bör vara fallet att A” (eller ”Det är obligatoriskt att A”), och $V_t A$ ”t vill att A”.

Parenteser utelämnas i regel om ingen mångtydighet uppstår. Predikatens ställighet bestäms av kontexten. I ”Fx” är F ett ett-ställigt predikat och i ”Fxy” ett två-ställigt predikat, etc. $\forall x_{1\text{-}n} A$ är en förkortning av $\forall x_1 \ldots \forall x_n A$; och likadant med $\exists x_{1\text{-}n} A$.

Om $\mathbf{\Phi}$ är en term, formel eller mängd formler i $\mathcal{L}$, så betecknar $\mathcal{V}(\mathbf{\Phi})$, $\mathcal{C}(\mathbf{\Phi})$, $\mathcal{T}(\mathbf{\Phi})$, $\mathcal{R}(\mathbf{\Phi})$, $\mathcal{R}^n(\mathbf{\Phi})$, $\mathcal{P}(\mathbf{\Phi})$ mängden variabler, konstanter, termer, predikat, n-ställiga predikat, respektive satsvariabler som förekommer i $\mathbf{\Phi}$. $\mathcal{L}(\mathbf{\Phi})$ denoterar $\mathbf{\Phi}$’s språk, dvs. den mängd formler som kan konstrueras från $\mathcal{R}(\mathbf{\Phi}) \cup \mathcal{T}(\mathbf{\Phi})$ plus de logiska konnektiven och operatorerna. ■

Definitioner av viktiga begrepp

Begreppen fri och bunden variabel, alfabetisk variant etc. förklaras som vanligt. Om x är en variabel som förekommer i en formel A, så är det viktigt att skilja mellan variabeln x själv och en eller flera förekomster av x i A. En förekomst av x i A är en viss del av formeln A, en del som har den syntaktiska formen ”x”. Satsen $A := \forall xFx \wedge Gx$ innehåller t.ex. tre förekomster av variabeln x: den längst till vänster, omedelbart till höger om kvantifikatorn, är bunden, den i mitten är bunden, och den längst till höger är fri. På samma sätt måste vi skilja mellan en delformel B av A och förekomsterna av B i A. Satsen $A := B \wedge (B \rightarrow C)$ innehåller två förekomster av delformeln B.

Nu kan vi kalla en förekomst av variabel x i en formel A fri om och endast om (omm) den inte förekommer inom räckvidden för en förekomst av $\forall x$ i A (annars är den bunden). Vi kallar variabeln x själv fri i A omm den har åtminstone en fri förekomst i A; annars är x bunden i A. $\mathcal{V}_f(A)$ denoterar mängden av fria variabler i A och $\mathcal{V}_b(A)$ mängden av bundna variabler i A. Uppenbarligen gäller det att $\mathcal{V}_f(A) \cap \mathcal{V}_b(A) = \varnothing$ och att $\mathcal{V}_f(A) \cup \mathcal{V}_b(A) = \mathcal{V}(A)$. Begreppet variabel-substitution kan nu förklaras på vanligt sätt. Vi säger att y är fri för x i A (där $x, y \in \mathcal{V}$, $A \in \mathcal{L}$) omm x inte förekommer i A inom räckvidden för en kvantifikator som binder y. Givet att y är fri för x i A, så är A[y/x] – resultatet av den korrekta substitutionen av y för x i A – den

formel som resulterar från A genom att ersätta varje fri förekomst av x med y. Om y är fri för x i A, så säger vi också att A[y/x] är definierad, annars är den odefinierad. På liknande sätt gäller det att $A[y_1/x_1, \ldots, y_n/x_n]$ denoterar resultatet av den korrekta simultana substitutionen av y_i för x_i (för alla $1 \leq i \leq n$; där x_i är parvis distinkta, men y_i inte behöver vara det). $A[y_{1\text{-}n}/x_{1\text{-}n}]$ är en förkortning av $A[y_1/x_1, \ldots, y_n/x_n]$. Notera att operationen "[y/x]" tar hela formeln "A" i ett uttryck av formen A[y/x] som argument. Schurz använder ibland parenteser för att belysa detta: t.ex. $\exists xA[y/x]$ läses $(\exists xA)[y/x]$, vilket är skilt från $\exists x(A[y/x])$ om A innehåller x fri (då är $\exists x(A[y/x]) \neq \exists xA$, medan $(\exists xA)[y/x] = \exists xA$) (Schurz (1997, s. 36)). Vi skall här göra likadant. Låt oss definiera detta lite mera exakt.

Definition 3 (fria och bundna variabler, öppna och slutna formler).

(1) En förekomst av en variabel v i A är fri i A omm den inte förekommer inom räckvidden för ett kvantifikator-uttryck av typen $\forall$v eller $\exists$v.

Rekursiv definition

(i) Om A är atomär, så är varje förekomst av varje v i $\mathcal{V}$ fri i A.

(ii) Om $A = \neg B$, $\Box A$ eller OA, så är en v-förekomst fri (bunden) i A omm den är fri (bunden) i B.

(iii) Om $A = B \vee C$, så är en v-förekomst i B (C) fri (bunden) i A omm den är fri (bunden) i B (C).

(iv) Om $A = V_xB$, så är en v-förekomst i B fri (bunden) i A omm den är fri (bunden) i B; x-förekomsten omedelbart till höger om V i A är (alltid) fri i A. Om c är en konstant och $A = V_cB$, så är en v-förekomst fri (bunden) i A omm den är fri (bunden) i B.

(v) Om $A = \forall xB$, så är en v-förekomst fri i A omm v inte är identisk med x och v är fri i B, annars är v bunden i A.

(2) En variabel är fri i A omm den har åtminstone en fri förekomst i A. $\mathcal{V}_f(A)$ = mängden av variabler som är fria i A.

(3) En formel i $\mathcal{L}$ som inte har några fria variabler kallas en sluten formel eller sats. Annars är det en öppen formel eller sats. ■

Vi vill att A[t/x] skall vara den formel som uppstår från en substitution av t för alla fria förekomster av x i A. Vi vill alltså att A[t/x] skall vara en formel som säger samma sak om t som A säger om x.

Exempel 4. Om A = Fx → Hyx, så är A[t/x] = Ft → Hyt. Om A = Fxy → ∀yHyx, så är A[t/x] = Fty → ∀yHyt. Om A = Fxy → ∀xHyx, så är A[t/x] = Fty → ∀xHyx. ■

För att uppnå detta måste vi dock undvika att t blir bunden av någon kvantifikator ∀t eller ∃t i A efter substitution. Vi måste undvika ”inkorrekta” ”substitutioner” av följande slag.

Exempel 5. Om A = Fxy → ∀yHyx, så är A[y/x] = Fyy → ∀yHyy. Om A = ∀x∀y(Fxy → Gxy), så är A[y/x] = ∀x∀y(Fyy → Gyy). Dessa ”substitutioner” är inkorrekta. Alla understrukna variabelförekomster är bundna av en kvantifikator i A. ■

Låt oss nu gå igenom några fler definitioner.

Definition 6 (Fri substitution).

(1) t är fri (att substitueras) för x i A omm ingen fri förekomst av x i A ligger inom räckvidden för en kvantifikator som binder t.

Rekursiv definition

(i) Om A är atomär, så är t fri för varje x i $\mathcal{V}$i A.

(ii) t är fri för x i ¬A, □A, OA och V_sA om t är fri för x i A.

(iii) t är fri för x i A ∨ B om t är fri för x i A och i B.

(iii) t är fri för x i ∀zA om t inte är z och t är fri för x i A.

(2) Om t är fri för x i A, så denoterar (A)[t/x] resultatet av den simultana substitutionen av t för alla fria x-förekomster i A. I annat fall är (A)[t/x] odefinierad. ■

Kommentar 7. Notera att (A)[t/x] är en metalingvistisk notation. Parenteser kring A används för att undvika mångtydighet. Men vi skall också ofta utelämna dem. A[t/x] opererar då på hela A, som vi tidigare påpekat. ■

Alla modala predikatlogiska system som Schurz beskriver är slutna under två substitutionsoperationer: substitution av fria individvariabler, och substitution av predikat (Schurz (1997, s. 45)). Dvs. i alla system Schurz beskriver gäller det att om A är ett teorem i logiken L, så är $A[y_{1\text{-}n}/x_{1\text{-}n}]$ ett teorem i L, givet att $A[y_{1\text{-}n}/x_{1\text{-}n}]$ är definierad. Schurz utvecklar en ny substitutionsoperation för predikat. En komplikation är att predikat inte endast kan ersättas av ordinära (atomära) predikat utan också av komplexa predikat, dvs. komplexa formler.

Notationen för uniform substitution av predikat har utvecklats av Kleene för icke-modal predikatlogik (Kleene (1971, ss. 155–162)), och utvidgas till modal predikatlogik av Schurz (1997, s. 45). Betrakta ett n-ställigt predikat F

som följs av vissa parvis distinkta variabler $x_1, \ldots, x_n$; $x_1, \ldots, x_n$ i $Fx_1 \ldots x_n$ kallas för ”namn form” variabler (Kleene (1971, s. 156)). Uniform substitution av formeln B för $Fx_1 \ldots x_n$ innebär att varje förekomst av $Fu_1 \ldots u_n$ (där u_i är variabler i $\mathcal{V}$, inte nödvändigtvis distinkta) ersätts av den korresponderande B-substitutions-instansen $B[u_{1\text{-}n}/x_{1\text{-}n}]$ givet att vissa restriktioner är uppfyllda som förhindrar sammanblandning av variabler. Kom ihåg att $B[u_{1\text{-}n}/x_{1\text{-}n}]$ står för den sats som är resultatet av att samtidigt ersätta varje fri förekomst av x_1 mot u_1, varje fri förekomst av x_2 mot u_2 osv. De fria variabler i B som inte är namn form variabler, kallas anonyma variabler i B (Kleene (1971, s. 156f)). Notera att det är möjligt att B inte innehåller några fria förekomster av x_i ($1 \leq i \leq n$). I detta fall ersätts varje $Fu_1 \ldots u_n$ av samma formel B. Eftersom valet av namn form variabler är godtyckligt, så använder Schurz en oändlig mängd parvis distinkta namn form variabler $x_1, \ldots, x_i, \ldots$.

Definition 8 (Substitution av predikat).
En substitutionsfunktion för predikat är en funktion σ: $\mathcal{R} \rightarrow \mathcal{L}$. En formel B är fri för predikat F i $\mathcal{R}^n$ i en formel A omm det för varje $Fu_1 \ldots u_n$ i A gäller att:

(i) $B[u_{1\text{-}n}/x_{1\text{-}n}]$ är definierad, och

(ii) $Fu_1 \ldots u_n$ inte förekommer inom räckvidden för en kvantifikator som binder en förekomst av en anonym variabel i B (Kleene (1971, s. 156)). En formel A är fri för σ om det för varje F i $\mathcal{R}$(A) gäller att σF är fri för F i A. Om A är fri för σ, så säger vi också att σA är definierad, annars är σA odefinierad. Om A är fri för σ, så denoterar σA resultatet av den simultana substitutionen av varje förekomst av $Fu_1 \ldots u_n$ med $(\sigma F)[u_{1\text{-}n}/x_{1\text{-}n}]$ i A för varje F i $\mathcal{R}^n$(A) och n i **N**. (Schurz (1997, s. 46)) ■

Definition 9 (Rekursiv definition av ”A är fri för σ” och ”σA”).

(1) Om $A = Fu_1 \ldots u_n$ är atomär, så är A fri för σ omm $u_1, \ldots, u_n$ är fria för $x_1, \ldots, x_n$ i σF, resp.; och om detta är fallet, så är $\sigma A = (\sigma F)[u_{1\text{-}n}/x_{1\text{-}n}]$.

(2) Då $A = \neg B$, $B \vee C$, $\Box B$, OB, eller $V_t B$ gäller följande: Om $A = \neg B$, $\Box B$, OB eller $V_t B$, så är A fri för σ omm B är fri för σ; i det kvarvarande fallet om B och C är fria för σ; och givet att A är fri för σ, så är $\sigma A = \neg \sigma B$, $(\sigma B \vee \sigma C)$, $\Box \sigma B$, $O\sigma B$, respektive $V_t \sigma B$.

(3) A = ∀zB. Då är A fri för σ om B är fri för σ och för varje G i $\mathcal{R}^n$(B) (och varje n i **N**), z inte är en anonym variabel i σG; och om detta är fallet, så är σA = ∀zσB. (Schurz (1997, s. 46)) ■

Kommentar 10. Om $Fu_1 \ldots u_n$ är atomär, så är $\sigma(Fu_1 \ldots u_n) = (\sigma(F))[u_{1\text{-}n}/x_{1\text{-}n}]$, vilket kan förenklas till följande uttryck $\sigma(Fu_1 \ldots u_n) = (\sigma F)[u_{1\text{-}n}/x_{1\text{-}n}]$ eller $\sigma F[u_{1\text{-}n}/x_{1\text{-}n}]$.

Notera att genom att ersätta definierade symboler med primitiva begrepp så är det enkelt att se att induktiva villkor av typ (2) också gäller för ∧, →, ↔, ◇ och P, och villkor av typ (3) för ∃.

Substitutionsfunktioner för propositioner är ett särskilt fall av substitutionsfunktioner för predikat begränsade till noll-ställiga predikat. ■

Några exempel

Diskussionen har hittills varit ganska abstrakt. Vi skall nu ta upp några exempel på konkreta substitutionsfunktioner för att belysa innebörden i de olika begreppen. Alla exempel, utom det första, är lånade från Schurz.

Låt F och H vara ett-ställiga predikat, och låt π(F) = πF = G, π(H) = πH = H och A := ∀x(Fx → Hx). Här är π en substitutionsfunktion som tar oss från predikat till predikat. Då är π(A) = πA = π(∀x(Fx → Hx)) = π∀x(Fx → Hx) = ∀x(Gx → Hx). Vi kan stegvis komma fram till detta på följande sätt. π(∀x(Fx → Hx)) = ∀xπ((Fx → Hx)) = ∀x(π(Fx) → π(Hx)) = ∀x(π(F)x → π(H)x) = ∀x(Gx → Hx).

Låt F, och G vara ett-ställiga predikat, H, I, J, och K två-ställiga predikat; vidare, låt σ(F) = σF = Hx_1z, σ(G) = σG = Izx_1, σ(J) = σJ = (∃zHx_1z → Kx_1x_2), och A := ∀y((Fy ∧ Gy) → □∃uJyu). Notera att i σF = Hx_1z är x_1 en namn form variabel och z en anonym variabel, i σG = Izx_1 är x_1 en namn form variabel och z en anonym variabel, och i σJ = (∃zHx_1z → Kx_1x_2) är x_1 och x_2 namn form variabler och z en anonym variabel. Då är A fri för σ, och σ(A) = σA = ∀y((Hyz ∧ Izy) → □∃u(∃zHyz → Kyu)). Men B := Fu → Jzu är inte fri för σ eftersom σJ[z/x_1, u/x_2] = (∃zHx_1z → Kx_1x_2)[z/x_1, u/x_2] är odefinierad. (∃zHx_1z → Kx_1x_2)[z/x_1, u/x_2] = (∃zH<u>z</u>z → Kzu), och i (∃zH<u>z</u>z → Kzu), blir den understrukna förekomsten av z bunden av ∃z. C := ∀z(Gu ∧ ¬Fu) är inte heller fri för σ eftersom den anonyma variabeln z i σG och σF blir bunden av ∀z i σC (= ∀z(I<u>z</u>u ∧ ¬Hu<u>z</u>)). Schurz nämner även ett exempel på en ”degenererad” substitution. Om F och G är tvåställiga predikat och σF = Hx_1, σG = p, så är σ∀x∃y(Fxy → Gyy) = ∀xEy(Hx → p) (som är ekvivalent med ∃xHx → p).

Genom att använda den rekursiva definitionen av substitutionsfunktioner kan man stegvis komma fram till att σA ovan är $\forall y((Hyz \wedge Izy) \rightarrow \Box\exists u(\exists zHyz \rightarrow Kyu))$ på följande sätt.

$\sigma(\forall y((Fy \wedge Gy) \rightarrow \Box\exists uJyu)) =$
$\forall y\sigma(((Fy \wedge Gy) \rightarrow \Box\exists uJyu)) =$
$\forall y(\sigma((Fy \wedge Gy)) \rightarrow \sigma(\Box\exists uJyu)) =$
$\forall y((\sigma(Fy) \wedge \sigma(Gy)) \rightarrow \Box\sigma(\exists uJyu)) =$
$\forall y((\sigma(Fy) \wedge \sigma(Gy)) \rightarrow \Box\exists u\sigma(Jyu)) =$
$\forall y(((\sigma(F))[y/x_1] \wedge (\sigma(G))[y/x_1]) \rightarrow \Box\exists u(\sigma(J))[y/x_1,u/x_2]) =$
$\forall y(((Hx_1z)[y/x_1] \wedge (Izx_1)[y/x_1]) \rightarrow \Box\exists u(\exists zHx_1z \rightarrow Kx_1x_2)[y/x_1,u/x_2])$
$=$
$\forall y((Hyz \wedge Izy) \rightarrow \Box\exists u(\exists zHyz \rightarrow Kyu))$

Om man utelämnar den yttersta parentesen runt en sats när ingen mångtydighet uppstår blir $\forall y\sigma(((Fy \wedge Gy) \rightarrow \Box\exists uJyu))$ istället $\forall y\sigma((Fy \wedge Gy) \rightarrow \Box\exists uJyu)$ och $\forall y(\sigma((Fy \wedge Gy)) \rightarrow \sigma(\Box\exists uJyu))$ istället $\forall y(\sigma(Fy \wedge Gy) \rightarrow \sigma(\Box\exists uJyu))$. $\forall y(((\sigma(F))[y/x_1] \wedge (\sigma(G))[y/x_1]) \rightarrow \Box\exists u(\sigma(J))[y/x_1,u/x_2])$ kan förenklas till $\forall y(((\sigma F)[y/x_1] \wedge (\sigma G)[y/x_1]) \rightarrow \Box\exists u(\sigma J)[y/x_1,u/x_2])$, vilken i sin tur kan förenklas till $\forall y((\sigma F[y/x_1] \wedge \sigma G[y/x_1]) \rightarrow \Box\exists u\sigma J[y/x_1,u/x_2])$.

Vi är nu redo att visa att satserna (K1)–(K10) i Avsnitt 2 är härledbara från GR4.

4. Lösningar av hur GR medför satserna (K1)–(K10) i Avsnitt 2

I Avsnitt 2 nämnde vi några satser ((K1)–(K10)) som alla tycks följa från den gyllene regeln, åtminstone om vi tolkar denna regel som (GR4). I det här avsnittet skall vi visa hur man kan förstå detta. Antag att den gyllene regeln symboliseras på följande sätt.

(FGR) $\forall x\forall y(V_xHyx \rightarrow OHxy)$

$\forall x\forall y(V_xHyx \rightarrow OHxy)$ läses ”Det gäller för alla x och y att: Om x vill att y utför H mot x, så bör x utföra H mot y”. Om vi antar att predikatet H i (FGR) kan ersättas med vilken formel som helst, kan vi härleda adekvata symboliseringar av alla satserna (K1)–(K10) i Avsnitt 2. Detta antyder att följande formella framställning av (GR4) är rimlig. Den gyllene regeln tolkad som (GR4) medför (FGR) och alla satser som kan fås från (FGR) med hjälp av en substitutionsfunktion σ, givet att (FGR) är fri för σ.

Låt oss i detalj gå igenom hur man kan härleda (K1), (K5), (K7) och (K10) om man tolkar den gyllene regeln på detta sätt.[5]

Härledning av (K1)
(K1) Om du vill att din vän är ärlig mot dig, så bör du vara ärlig mot din vän.

Låt Rxy stå för x är ärlig mot y och låt d vara en singulär term som refererar till dig och v en singulär term som refererar till din vän. Låt $\sigma(H) = Rx_1x_2$. x_1 och x_2 är namn form variabler, det finns ingen anonym variabel i uttrycket, och (FGR) är fri för σ. Då kan (K1) härledas med hjälp av nedanstående steg.

$\sigma(\forall x\forall y(V_xHyx \rightarrow OHxy)) = \forall x\sigma(\forall y(V_xHyx \rightarrow OHxy)) =$
$\forall x\forall y\sigma((V_xHyx \rightarrow OHxy)) = \forall x\forall y(\sigma(V_xHyx) \rightarrow \sigma(OHxy)) =$
$\forall x\forall y(V_x\sigma(Hyx) \rightarrow O\sigma(Hxy)) =$
$\forall x\forall y(V_x(\sigma(H))[y/x_1, x/x_2] \rightarrow O(\sigma(H))[x/x_1, y/x_2]) =$
$\forall x\forall y(V_x(Rx_1x_2)[y/x_1, x/x_2] \rightarrow O(Rx_1x_2)[x/x_1, y/x_2]) =$
$\forall x\forall y(V_xRyx \rightarrow ORxy)$

Slutsatsen i denna ”deduktion” kan läsas på följande sätt. ”Det gäller för alla x och y att: Om x vill att y är ärlig mot x, så bör x vara ärlig mot y”. Från detta följer:

$V_dRvd \rightarrow OHdv$

om vi instanserar x med d och y med v. $V_dRvd \rightarrow OHdv$ läses: ”Om du vill att din vän är ärlig mot dig, så bör du vara ärlig mot din vän” = (K1). V.S.B.

Härledning av (K5)
(K5) Om du vill att din arbetskamrat inte ljuger för dig, så bör du inte ljuga för din arbetskamrat.

Låt Lxy stå för x ljuger för y och låt d vara en singulär term som refererar till dig och a en singulär term som refererar till din arbetskamrat. Låt $\sigma(H) = \neg Lx_1x_2$. x_1 och x_2 är namn form variabler, det finns ingen anonym variabel i uttrycket, och (FGR) är fri för σ. (K5) kan nu härledas på följande sätt.

[5] I en strikt mening visar vi hur ett antal *symboliseringar* av (K1), (K5), (K7) och (K10) följer från (FGR) med hjälp av substitution av predikatet H. Men givet att dessa symboliseringar är ”korrekta”, kan vi dra slutsatsen att dessa satser själva följer ur GR.

$\sigma(\forall x\forall y(V_xHyx \rightarrow OHxy)) = \forall x\sigma(\forall y(V_xHyx \rightarrow OHxy)) =$
$\forall x\forall y\sigma((V_xHyx \rightarrow OHxy)) = \forall x\forall y(\sigma(V_xHyx) \rightarrow \sigma(OHxy)) =$
$\forall x\forall y(V_x\sigma(Hyx) \rightarrow O\sigma(Hxy)) =$
$\forall x\forall y(V_x(\sigma(H))[y/x_1, x/x_2] \rightarrow O(\sigma(H))[x/x_1, y/x_2]) =$
$\forall x\forall y(V_x(\neg Lx_1x_2)[y/x_1, x/x_2] \rightarrow O(\neg Lx_1x_2)[x/x_1, y/x_2]) =$
$\forall x\forall y(V_x\neg Lyx \rightarrow O\neg Lxy)$

Den sista satsen läses: ”Det gäller för alla x och y att: Om x vill att y inte ljuger för x, så bör det vara fallet att x inte ljuger för y”. Det följer att:

$V_d\neg Lad \rightarrow O\neg Lda$

om vi instanserar x med d och y med a. $V_d\neg Lad \rightarrow O\neg Lda$ läses: ”Om du vill att din arbetskamrat inte ljuger för dig, så bör du inte ljuga för din arbetskamrat” = (K5). V.S.B.

Härledning av (K7)
(K7) Om du vill att din bror hjälper dig om du behöver hjälp, så bör du hjälpa din bror om han behöver hjälp.
(Om du vill att om du behöver hjälp så hjälper din bror dig, så bör det vara fallet att om din bror behöver hjälp så hjälper du honom.)

Låt Bx stå för x behöver hjälp, Hxy för x hjälper y och låt d vara en singulär term som refererar till dig och b en singulär term som refererar till din bror. Låt $\sigma(H) = Bx_2 \rightarrow Hx_1x_2$. x_1 och x_2 är namn form variabler, det finns ingen anonym variabel i uttrycket, och (FGR) är fri för σ. Följande deduktion visar hur (K7) kan härledas ur (FGR) med hjälp av σ.

$\sigma(\forall x\forall y(V_xHyx \rightarrow OHxy)) = \forall x\sigma(\forall y(V_xHyx \rightarrow OHxy)) =$
$\forall x\forall y\sigma((V_xHyx \rightarrow OHxy)) = \forall x\forall y(\sigma(V_xHyx) \rightarrow \sigma(OHxy)) =$
$\forall x\forall y(V_x\sigma(Hyx) \rightarrow O\sigma(Hxy)) =$
$\forall x\forall y(V_x(\sigma(H))[y/x_1, x/x_2] \rightarrow O(\sigma(H))[x/x_1, y/x_2]) =$
$\forall x\forall y(V_x(Bx_2 \rightarrow Hx_1x_2)[y/x_1, x/x_2] \rightarrow O(Bx_2 \rightarrow Hx_1x_2)[x/x_1, y/x_2]) =$
$\forall x\forall y(V_x(Bx \rightarrow Hyx) \rightarrow O(By \rightarrow Hxy))$

Dvs. ”Det gäller för alla x och y att: Om x vill att y hjälper x om x behöver hjälp, så bör x hjälpa y om y behöver hjälp”. Det följer att:

$V_d(Bd \rightarrow Hbd) \rightarrow O(Bb \rightarrow Hdb)$

om vi instanserar x med d och y med b. $V_d(Bd \rightarrow Hbd) \rightarrow O(Bb \rightarrow Hdb)$ är en rimlig formalisering av (K7) = ”Om du vill att om du behöver hjälp så hjälper din bror dig, så bör det vara fallet att om din bror behöver hjälp så hjälper du honom”.

(K7) kan alltså fås genom substitution från den gyllene regeln: Det gäller för alla individer x och y att: Om x vill att y utför handling H mot x, så bör x utföra H mot y. V.S.B.

Härledning av (K10)

(K10) Om du vill att x behandlar alla dina barn med respekt, så bör du behandla alla x’s barn med respekt.

Låt Bxy stå för x är ett barn till y, Rxy för x behandlar y med respekt, och låt d vara en konstant som refererar till dig. Låt $\sigma(H) = \forall z(Bzx_2 \rightarrow Rx_1z)$. $\sigma(H)$ läses ”Det gäller för alla z att om z är ett barn till x_2, så behandlar x_1 z med respekt”. Uttrycket saknar anonyma variabler, x_1 och x_2 är namn form variabler, och (FGR) är fri för σ. Följande härledning visar hur (K10) kan härledas ur (FGR) med hjälp av σ.

$\sigma(\forall x\forall y(V_xHyx \rightarrow OHxy)) = \forall x\sigma(\forall y(V_xHyx \rightarrow OHxy)) =$
$\forall x\forall y\sigma((V_xHyx \rightarrow OHxy)) = \forall x\forall y(\sigma(V_xHyx) \rightarrow \sigma(OHxy)) =$
$\forall x\forall y(V_x\sigma(Hyx) \rightarrow O\sigma(Hxy)) =$
$\forall x\forall y(V_x(\sigma(H))[y/x_1, x/x_2] \rightarrow O(\sigma(H))[x/x_1, y/x_2]) =$
$\forall x\forall y(V_x(\forall z(Bzx_2 \rightarrow Rx_1z))[y/x_1, x/x_2] \rightarrow$
$O(\forall z(Bzx_2 \rightarrow Rx_1z))[x/x_1, y/x_2]) =$
$\forall x\forall y(V_x\forall z(Bzx \rightarrow Ryz) \rightarrow O\forall z(Bzy \rightarrow Rxz))$

Den sista satsen i denna härledning läses: ”Det gäller för alla x och y att om x vill att det gäller för alla z att om z är ett barn till x så behandlar y z med respekt, så bör det vara fallet att det gäller för alla z att om z är ett barn till y så behandlar x z med respekt. Från denna sats kan vi sluta oss till följande formel om vi instanserar x med d och y med x.

$V_d\forall z(Bzd \rightarrow Rxz) \rightarrow O\forall z(Bzx \rightarrow Rdz)$

Denna formel läses: ”Om du vill att det gäller för alla z att om z är ett barn till dig så behandlar x z med respekt, så bör det vara fallet att det gäller för alla z att om z är ett barn till x så behandlar du z med respekt”. Och denna sats är synonym med (K10). V.S.B.

(K10) är en ganska komplicerad konsekvens av GR4 och visar att inte endast enkla kategoriska plikter följer från denna princip, utan också t.ex. olika sorters villkorliga normer. Det här gör GR4 till en mycket kraftfull moralisk maxim, som potentiellt kan användas i härledningen av många andra normer.

Notera också att alla konsekvenser av GR4 i sin tur kan användas i olika härledningar av olika moraliska påståenden. Betrakta t.ex. följande argument.

1. Den gyllene regeln.
2. Du vill att x skall behandla dina barn med respekt.
3. (Det är ett historiskt faktum att) Lisa är ett barn till x.

Alltså

4. Du bör behandla Lisa med respekt.

Detta argument är giltigt om vi antar att den gyllene regeln tolkas på det sätt vi har föreslagit ovan, vi använder en vanlig möjlig värld-semantik för vårt aletiskt-deontiska system, alla historiska fakta är historiskt nödvändiga, vi kvantifierar över alla möjliga individer, och alla deontiskt tillgängliga världar är aletiskt tillgängliga. Detta innebär inte nödvändigtvis att GR *faktiskt* är sann. Men det innebär att denna regel tillsammans med övriga premisser medför slutsatsen. Och alla andra premisser är plausibla.

Vi kan visa detta på följande sätt. 1 medför (i) $V_d\forall z(Bzd \rightarrow Rxz) \rightarrow O\forall z(Bzx \rightarrow Rdz)$. 2, $V_d\forall z(Bzd \rightarrow Rxz)$, tillsammans med (i) medför (ii) $O\forall z(Bzx \rightarrow Rdx)$. (ii) tillsammans med 3, $\Box Blx$, medför 4, ORdl. Att (ii) och 3 medför 4 kan visas på följande sätt. (Vi antar en vanlig möjlig värld-semantik i följande argument.) Antag att (1) $O\forall z(Bzx \rightarrow Rdx)$ och (2) $\Box Blx$ är sanna i den möjliga världen w_0, och att (3) ORdl är falsk i w_0. Då gäller det att (4) det finns en möjlig värld w_1 sådan att w_1 är deontiskt tillgänglig från w_0 [Från 3], och (5) Rdl är falsk i w_1 [Från 3]. Alltså, (6) $\forall z(Bzx \rightarrow Rdz)$ är sann i w_1 [Från 1 och 4]. (7) $Blx \rightarrow Rdl$ är sann i w_1 [Från 6]. (8) w_1 är aletiskt tillgänglig från w_0 [Från antagandet att alla deontiskt tillgängliga världar också är aletiskt tillgängliga]. Alltså, (9) Blx är sann i w_1 [Från 2 och

8]. Det följer att, (10) Rdl är sann i w_1 [Från 7 och 9], och att (11) Rdl $\wedge \neg$Rdl är sann i w_1 [Från 5 och 10]. Men detta är absurt.

5. Problem för GR4: substitutionsargumentet

Den gyllene regeln är alltså en potentiellt mycket användbar moralisk princip. Vi skall emellertid nu undersöka ett argument som talar för att GR4 är en alltför stark tolkning av GR. Enligt den aktuella preciseringen av GR är alla satser som kan fås från (FGR) med hjälp av en substitutionsfunktion σ, givet att (FGR) är fri för σ, konsekvenser av den gyllene regeln. Men betrakta nu följande exempel.

Låt Fxy stå för x är ärlig mot y, d referera till dig och a till någon godtycklig, konkret person (t.ex. din partner). Antag att $\sigma(H) = (x_1 = a \wedge Fx_1x_2) \vee (\neg x_1 = a \wedge \neg Fx_1x_2)$. x_1 och x_2 är namn form variabler, och a är en individkonstant. Uttrycket innehåller inga anonyma variabler. Då är följande argument giltigt, om vi gör vissa antaganden som förefaller vara mycket plausibla (se nedan).

1. V_dFad	[Antag]
2. $V_d\sigma(H)[a/x_1, d/x_2]$	[Från 1, ”def av” $\sigma(H)[a/x_1, d/x_2]$ etc.]
3. $O\sigma(H)[d/x_1, a/x_2]$	[Från 2 med GR4]
4. $O\neg Fda$	[Från 3, ”def av” $\sigma(H)[d/x_1, a/x_2]$ etc.]

Här följer en informell läsning av detta argument.

Du vill att a skall vara ärlig mot dig.
Alltså vill du att a skall utföra σH mot dig.
Om du vill att a skall utföra σH mot dig, så bör du utföra σH mot a.
Alltså bör du utföra σH mot a.
Alltså bör du inte vara ärlig mot a.

Men denna slutsats är kontraintuitiv. Låt oss kalla detta argument för ”substitutionsargumentet”. Från det faktum att du vill att a skall vara ärlig mot dig och att GR4 är sann, följer det då att du *inte* bör vara ärlig mot a. Vi kan härleda liknande resultat för alla typer av handlingar. Om du vill att a är trogen mot dig, så bör du *inte* vara trogen mot a. Om du vill att a håller sina löften till dig, så bör du *inte* hålla dina löften till a osv. Detta är moraliskt orimligt.

Och inte nog med det. Vi kan dessutom härleda en direkt motsägelse om GR4 är sann och vi antar att det inte finns några genuina moraliska dilemman. För vi kan också härleda OFda på följande sätt. (SL innebär att steget följer med hjälp av vanlig satslogik.)

5. V_dFad	[Antag]
6. $V_dFad \rightarrow OFda$	[Från GR4]
7. OFda	[5, 6, SL]

Och från detta är det lätt att härleda en kontradiktion.

8. $OFda \wedge O\neg Fda$	[4, 7, SL]
9. $\neg(OFda \wedge O\neg Fda)$	[OD]
10. $(OFda \wedge O\neg Fda) \wedge \neg(OFda \wedge O\neg Fda)$	[8, 9, SL]

Men 10 är en motsägelse. De enda regler, förutom GR4, vi behöver anta i steg 5–10 är (OD), $\neg(OA \wedge O\neg A)$, som utesluter förekomsten av explicita moraliska dilemman, och olika grundläggande satslogiska regler. Och alla dessa principer tycks vara rimliga. Om argumentet är giltigt och alla övriga premisser är sanna, måste vi förkasta GR4. Det här argumentet är därför djupt problematiskt för denna tolkning av den gyllene regeln. Kan en anhängare av GR4 möjligtvis attackera steg 2, 3, eller 4 i argumentet ovan? Vi skall nu se hur dessa steg kan försvaras.

Steg 2 $V_dFad \Rightarrow V_d\sigma(H)[a/x_1, d/x_2]$. För att visa steg 2, måste vi visa att $V_dFad \Rightarrow V_d\sigma(H)[a/x_1, d/x_2]$. Notera att $\sigma H = (x_1 = a \wedge Fx_1x_2) \vee (\neg x_1 = a \wedge \neg Fx_1x_2)$. Alltså är $\sigma(H)[a/x_1, d/x_2] = (a = a \wedge Fad) \vee (\neg a = a \wedge \neg Fad)$. Det följer att $V_d\sigma(H)[a/x_1, d/x_2] = V_d((a = a \wedge Fad) \vee (\neg a = a \wedge \neg Fad))$. Så, för att visa att $V_dFad \Rightarrow V_d\sigma(H)[a/x_1, d/x_2]$, måste vi visa att $V_dFad \Rightarrow V_d((a = a \wedge Fad) \vee (\neg a = a \wedge \neg Fad))$. Detta följer om vi antar att V_xB följer ur V_xA om A och B är logiskt ekvivalenta (kalla denna regel (VE)). För Fad är logiskt ekvivalent med $(a = a \wedge Fad) \vee (\neg a = a \wedge \neg Fad)$. Regeln att V_xB följer ur V_xA om A och B är logiskt ekvivalenta är knappast sann för alla personer i alla situationer. Men det är rimligt att anta att alla personer som är fullständigt rationella satisfierar denna princip. Och om vi antar att den gyllene regeln gäller också för fullständigt rationella personer, så har vi fortfarande ett problem. Om vi antar att V fungerar som en normal modal operator, kan vi även bevisa att V_dFad medför $V_d\sigma(H)[a/x_1, d/x_2]$ på följande sätt.

$V_dFad \Rightarrow V_d\sigma(H)[a/x_1, d/x_2] = V_d((a = a \wedge Fad) \vee (\neg a = a \wedge \neg Fad))$

(1) V_dFad, 0
(2) $\neg V_d((a = a \wedge Fad) \vee (\neg a = a \wedge \neg Fad))$, 0
(3) 0s1
(4) $\neg((a = a \wedge Fad) \vee (\neg a = a \wedge \neg Fad))$, 1
(5) $\neg(a = a \wedge Fad)$, 1
(6) $\neg(\neg a = a \wedge \neg Fad)$, 1
(7) Fad, 1
↙ ↘
(8) $\neg a = a$, 1 (9) $\neg$Fad, 1
(10) * (11) *

Steg 3 $V_d\sigma(H)[a/x_1, d/x_2] \Rightarrow O\sigma(H)[d/x_1, a/x_2]$. Vi har redan visat att $V_d\sigma(H)[a/x_1, d/x_2] = V_d((a = a \wedge Fad) \vee (\neg a = a \wedge \neg Fad))$. $\sigma H = (x_1 = a \wedge Fx_1x_2) \vee (\neg x_1 = a \wedge \neg Fx_1x_2)$. Alltså, $\sigma(H)[d/x_1, a/x_2] = (d = a \wedge Fda) \vee (\neg d = a \wedge \neg Fda)$. Således gäller det att $O\sigma(H)[d/x_1, a/x_2] = O((d = a \wedge Fda) \vee (\neg d = a \wedge \neg Fda))$. För att visa att $V_d\sigma(H)[a/x_1, d/x_2] \Rightarrow O\sigma(H)[d/x_1, a/x_2]$, måste vi alltså visa att $O((d = a \wedge Fda) \vee (\neg d = a \wedge \neg Fda))$ följer från $V_d((a = a \wedge Fad) \vee (\neg a = a \wedge \neg Fad))$ med hjälp av GR. Vi visar först att $\forall x\forall y(V_x((y = a \wedge Fyx) \vee (\neg y = a \wedge \neg Fyx)) \rightarrow O((x = a \wedge Fxy) \vee (\neg x = a \wedge \neg Fxy)))$ följer från (FGR) om vi tillämpar σ på denna sats. Denna sats medför i sin tur $V_d((a = a \wedge Fad) \vee (\neg a = a \wedge \neg Fad)) \rightarrow O((d = a \wedge Fda) \vee (\neg d = a \wedge \neg Fda))$ med hjälp av vanlig predikatlogik. Från detta följer omedelbart vårt resultat. Här är vårt bevis.

$\sigma H = (x_1 = a \wedge Fx_1x_2) \vee (\neg x_1 = a \wedge \neg Fx_1x_2)$.
$\sigma(\forall x\forall y(VxHyx \rightarrow OHxy)) = \forall x\sigma(\forall y(VxHyx \rightarrow OHxy)) =$
$\forall x\forall y\sigma((VxHyx \rightarrow OHxy)) = \forall x\forall y(\sigma(VxHyx) \rightarrow \sigma(OHxy)) =$
$\forall x\forall y(V_x\sigma(Hyx) \rightarrow O\sigma(Hxy)) =$
$\forall x\forall y(V_x(\sigma(H))[y/x_1,x/x_2] \rightarrow O(\sigma(H))[x/x_1,y/x_2]) =$
$\forall x\forall y(V_x((x_1 = a \wedge Fx_1x_2) \vee (\neg x_1 = a \wedge \neg Fx_1x_2))[y/x_1,x/x_2] \rightarrow$
$O((x_1 = a \wedge Fx_1x_2) \vee (\neg x_1 = a \wedge \neg Fx_1x_2))[x/x_1,y/x_2]) =$
$\forall x\forall y(V_x((y = a \wedge Fyx) \vee (\neg y = a \wedge \neg Fyx)) \rightarrow O((x = a \wedge Fxy) \vee (\neg x = a \wedge \neg Fxy)))$.

$V_d((a = a \wedge Fad) \vee (\neg a = a \wedge \neg Fad)) \rightarrow O((d = a \wedge Fda) \vee (\neg d = a \wedge \neg Fda))$ är en instans av den sista satsen i härledningen ovan (låt x = d, och y = a).

Steg 4 $O\sigma(H)[d/x_1, a/x_2] \Rightarrow O\neg Fda$. $\sigma H = (x_1 = a \wedge Fx_1x_2) \vee (\neg x_1 = a \wedge \neg Fx_1x_2)$. Det innebär att $\sigma(H)[d/x_1, a/x_2] = (d = a \wedge Fda) \vee (\neg d = a \wedge \neg Fda)$ och att $O\sigma(H)[d/x_1, a/x_2] = O((d = a \wedge Fda) \vee (\neg d = a \wedge \neg Fda))$. För att visa $O\sigma(H)[d/x_1, a/x_2] \Rightarrow O\neg Fda$, måste vi alltså visa att $O((d = a \wedge Fda) \vee (\neg d = a \wedge \neg Fda))$ medför $O\neg Fda$. För att visa detta lägger vi till premissen att du inte är identisk med a, dvs. $\neg d = a$, och antar att all identitet är nödvändig. All identitet är nödvändig om vi antar att individkonstanterna tolkas som rigida designatorer och refererar direkt till individer. Här följer ett bevis av steg 4.

$O\sigma(H)[d/x_1, a/x_2], \neg d = a \Rightarrow O\neg Fda$

(1) $O((d = a \wedge Fda) \vee (\neg d = a \wedge \neg Fda))$, 0
(2) $\neg d = a$, 0
(3) $\neg O\neg Fda$, 0
(4) $P\neg\neg Fda$, 0
(5) 0s1
(6) $\neg\neg\neg Fda$, 1
(7) $(d = a \wedge Fda) \vee (\neg d = a \wedge \neg Fda)$, 1

(8) $d = a \wedge Fda$, 1	(9) $\neg d = a \wedge \neg Fda$, 1
(10) $d = a$, 1	(11) $\neg d = a$, 1
(12) Fda, 1	(13) $\neg Fda$, 1
(14) $d = a$, 0	(15) *
(16) *	

6. Möjliga svar

Är det möjligt att undvika slutsatsen i detta argument? Så vitt jag kan se finns det i teorin sex möjliga svar på problemet i Avsnitt 5. Vi kan (i) Förkasta satslogiken (och/eller predikatlogiken); (ii) Överge (OD), $\neg(OA \wedge O\neg A)$; (iii) Ge upp antagandet att V_dFad; (iv) Förkasta uppfattningen att identitetsrelationen är nödvändig; (v) Överge regeln (VE), att om B är logiskt ekvivalent med A och V_xA, så V_xB; eller (vi) Förkasta GR4. Låt oss kort undersöka dessa alternativ.

(i) Om man överger den klassiska satslogiken (och/eller den klassiska predikatlogiken) och t.ex. accepterar att det kan finnas sanna motsägelser, så kan man undvika slutsatsen att GR4 är falsk. Men det här tycks vara ett desperat förslag som man endast i yttersta nödfall bör ta till. Både den klassiska satslogiken och den klassiska predikatlogiken är extremt tilltalande och väletablerade.

(ii) Om vi förkastar OD kan vi undvika steg 9 i substitutionsargumentet ovan och kan då inte härleda en direkt motsägelse. Och det finns filosofer som har ifrågasatt denna princip. Personligen tycker jag dock att den är rimlig, och jag har försökt försvara den i andra arbeten (se Rönnedal (2012, ss. 73–96)). Även om man skulle överge (OD) så följer fortfarande steg 8 i argumentet. Och detta tycks vara en nästan lika orimlig konklusion som 10. Notera också att vi kan härleda ett liknande moraliskt dilemma för varje H, sådant att du vill att a utför H mot dig. Vidare gäller det att om du vill att a är ärlig mot dig, hjälper dig, håller sina löften till dig osv., så följer det att du *inte* bör vara ärlig mot a, *inte* bör hjälpa a, *inte* bör hålla dina löften till a osv. Och detta är orimligt, oavsett om resonemanget ger upphov till en motsägelse eller inte.

(iii) Om det inte är sant att du vill att a är ärlig mot dig, så följer inte våra problematiska slutsatser. Kanske vill inte alla personer att andra skall vara ärliga mot dem. Men det är nog ett rimligt antagande att det gäller för åtminstone nästan alla personer. Och för att argumentet skall gå igenom räcker det med att det finns en enda person som vill att någon annan är ärlig mot henne. Notera också att samma typ av argument kan användas även om F antas representera någon annan typ av handling. Så även om det skulle vara sant att det inte är fallet att det finns något x och något y sådana att x vill att y är ärlig mot x, så finns det säkert något x och något y sådana att x vill att y utför F mot x, för någon (problematisk typ av) handling F. Det här svaret tycks därför inte vara rimligt.

(iv) Om vi förkastar antagandet att identitetsrelationen är nödvändig, så kan vi inte längre visa steg 4 på samma sätt som ovan. Och det är inte uppenbart att all identitet är nödvändig. Om vi betraktar individkonstanterna d och a som rigida designatorer, om dessa refererar till samma personer i varje möjlig värld, och de refererar direkt till konkreta individer, följer det dock att deras icke-identitet är nödvändig. Och i vårt argument tycks det vara rimligt att anta att d och a är rigida designatorer och refererar direkt till dig och din partner. Om detta är riktigt, är (iv) inte ett svar som kan användas för att undvika substitutionsargumentet.

(v) Som vi redan har påpekat är regeln att V_xB följer ur V_xA om A och B är logiskt ekvivalenta knappast sann för alla personer i alla situationer. Men det är rimligt att anta att alla personer som är fullständigt rationella satisfierar denna princip. Och GR borde gälla för fullständigt rationella personer om några. Då har vi fortfarande ett problem.

(vi) Det enda rimliga alternativ som återstår tycks vara att överge GR4. Men om GR4 inte är en plausibel princip, betyder det också att den gyllene regeln inte är förnuftig? Innebär det att vi bör överge denna moraliska princip? Inte nödvändigtvis. Det finns andra tolkningar av GR. Vi nämnde några i Avsnitt 2. Vi skall nu emellertid undersöka ytterligare en interpretation. Betrakta följande läsning av GR.

> **(GR5).** Det gäller för alla individer x och y att: Om x vill att y utför handling H mot x, så bör x utföra H mot y, där ”H” står för en handling som kan uttryckas med vilket handlingspredikat som helst som inte innehåller några individkonstanter.

Låt $\underline{\sigma}$ vara en substitutionsfunktion som tar oss från predikat till formler som inte innehåller några individkonstanter. Då gäller det att den gyllene regeln tolkad som (GR5) medför (FGR) och alla satser som kan fås från (FGR) med hjälp av en substitutionsfunktion $\underline{\sigma}$, givet att (FGR) är fri för $\underline{\sigma}$. Om vi tolkar GR på detta sätt, så kan vi undvika problemet i Avsnitt 5. $\sigma H = (x_1 = a \wedge Fx_1x_2) \vee (\neg x_1 = a \wedge \neg Fx_1x_2)$. Men $(x_1 = a \wedge Fx_1x_2) \vee (\neg x_1 = a \wedge \neg Fx_1x_2)$ är en formel som innehåller individkonstanten a. σ kan därför inte tillämpas på (FGR) enligt GR5. Tolkar vi GR som GR5 är steg 3 i substitutionsargumentet inte tillåtet. Detta argument kan därför inte användas emot GR5.

GR5 är nästan lika stark som GR4. GR5 medför GR3, GR2 och GR1 och alla intuitivt riktiga konsekvenser av GR som vi nämnde i Avsnitt 2, dvs. (K1)–(K10) kan också härledas ur GR5 på samma sätt som vi gjorde i Avsnitt 4. Att GR5 är något svagare än GR4 tycks därför inte vara något allvarligt problem för en anhängare av den gyllene regeln.

Faktum är att det kan finnas andra skäl att föredra GR5 framför GR4. Begränsningen till substitutionsfunktioner vars värden inte innehåller några individkonstanter reflekterar väl den spridda uppfattningen att moralen *är* eller *bör vara* opartisk. Vad olika personer bör och inte bör göra och hur de bör och inte bör behandlas tycks inte vara beroende av vilka specifika individer de råkar vara, utan tycks endast bero på deras ”universella” egenskaper. Om GR tolkas på det här sättet är det möjligt att regeln kan härledas från en mer generell universaliserbarhets- eller opartiskhetsprincip. Att argumentera för detta ligger dock utanför den här uppsatsens ramar.

Konklusionen är att den gyllene regeln tolkad som GR5 inte kan vederläggas av substitutionsargumentet i Avsnitt 5 och att det förefaller vara helt rimligt att föredra GR5 framför GR4.

7. Slutsats

Enligt den gyllene regeln bör vi behandla andra så som vi själva vill bli behandlade. Jag har i den här uppsatsen undersökt några olika tolkningar av denna maxim och jag har framför allt intresserat mig för frågan om GR uttalar sig om *alla* handlingar eller inte och vad det innebär. Vi såg hur uttrycket ”alla handlingar” kan preciseras med hjälp av teorin för substitutionsfunktioner. Jag visade hur man med hjälp av tolkning 4 av GR kan härleda en mängd normer som alla tycks vara konsekvenser av den gyllene regeln. Enligt denna läsning är den gyllene regeln potentiellt mycket kraftfull och användbar. Vi såg emellertid att det s.k. substitutionsargumentet talar för att vår ursprungliga formulering, GR4, tycks vara alltför stark. Jag visade hur detta argument kan bemötas och hur man kan formulera en något svagare variant, GR5, som förefaller vara rimligare än GR4. Om resonemangen i den här uppsatsen är riktiga, så kan substitutionsargumentet inte användas för att vederlägga GR5. Och eftersom GR5 tycks vara en rimlig tolkning av den gyllene regeln, så innebär substitutionsargumentet inte något allvarligt problem för en anhängare av denna välkända princip. Det finns emellertid en mängd övriga uttryck i den gyllene regeln som också kan tolkas på många olika sätt. För att visa hur man kan bemöta olika potentiella invändningar, bör man nog säga något mer också om dessa uttryck. Jag hoppas kunna återvända till ämnet vid något annat tillfälle.[6]

Referenser

Blackstone, W. T. (1965). The Golden Rule: A Defense. *Southern Journal of Philosophy*, ss. 172–177.

Bruton, S. V. (2004). Teaching the Golden Rule. *Journal of Business Ethics*, Vol. 49, Nr. 2, ss. 179–187.

Cadoux, A. T. (1912). The Implications of the Golden Rule. *International Journal of Ethics*, Vol. 22, Nr. 3, ss. 272–287.

Carson, T. L. (2010). *Lying and Deception: Theory and Practice*. Oxford: Oxford University Press.

Carson, T. L. (2013). Golden Rule. I Hugh LaFollette (red.) *The International Encyclopedia of Ethics*, ss. 2186–2192.

Duxbury, N. (2009). Golden Rule Reasoning, Moral Judgement and Law. *Notre Dame Law Review* 84, ss. 1529–1605.

[6] I Rönnedal (2016) diskuterar jag några potentiella problem med vissa formuleringar av den gyllene regeln. Jag visar också hur dessa problem kan bemötas.

Gensler, H. J. (1986). Ethics is Based on Rationality. *The Journal of Value Inquiry* 20, ss. 251–264.

Gensler, H. J. (1996). *Formal Ethics*. London and New York: Routledge.

Gensler, H. J. (2013). *Ethics and the Golden Rule*. New York and London: Routledge.

Gewirth, A. (1978). The golden rule rationalized. *Midwest Studies in Philosophy*, I11, ss. 133–147.

Gould, J. A. (1980). Blackstone's Meta-Not-So-Golden-Rule. *The Southern Journal of Philosophy*, Vol. 18, Issue 4, ss. 509–513.

Hare, R. M. (1963). *Freedom and Reason*. Oxford: Oxford University Press.

Hertzler, J. O. (1934). On Golden Rules. *International Journal of Ethics*, Vol. 44, Nr. 4, ss. 418–436.

Hirst, E. W. (1934). The Categorical Imperative and the Golden Rule. *Philosophy*, Vol. 9, Nr. 35, ss. 328–335.

Hobbes, T. (1985). *Leviathan*. Penguin Books. (red. C. B. Macpherson). (Ursprungligen publicerad 1651.)

Hoche, H.-U. (1978). Die Goldene Regel. Neue Aspekte eines alten Moralprinzips. *Zeitschrift für philosophische Forschung*, Bd. 32, H. 3, ss. 355–375.

Huang, Y. (2005). A Copper Rule versus the Golden Rule: A Daoist-Confucian Proposal for Global Ethics. *Philosophy East and West*, Vol. 55, Nr. 3, ss. 394–425.

Kleene, S. C. (1971). *Introduction to Metamathematics*. Groningen: Wolters-Noordhoff Publishing.

Mill, J. S. (1987). *Utilitarianism*. Buffalo, New York: Prometheus Books. (Ursprungligen publicerad 1863).

Neusner, J. och Chilton, B. (red.) (2008). *The Golden Rule: The Ethics of Reciprocity in the World Religions*. Continuum.

Pufendorf, S. (1964). *On the Law of Nature and Nations*. New York: Wildy and Sons. (Ursprungligen publicerad 1672).

Reinikainen, J. (2005). The Golden Rule and the Requirement of Universalizability. *The Journal of Value Inquiry* 39, ss. 155–168.

Rönnedal, D. (2012). *Extensions of Deontic Logic: An Investigation into some Multi-Modal Systems*. Department of Philosophy, Stockholm University.

Rönnedal, D. (2015). The Golden Rule and The Platinum Rule. *The Journal of Value Inquiry*, Volume 49, Issue 1, ss. 221–236.

Rönnedal, D. (2016). Den Gyllene Regeln och Intra- och Interpersonella Viljekonflikter. *Filosofiska Notiser*, Årgång 3, Nr 2, Augusti, ss. 81–106.

Schurz, G. (1997). *The Is-Ought Problem: An Investigation in Philosophical Logic*. Springer.
Singer, M. G. (1963). The Golden Rule. *Philosophy*, Vol. 38, Nr. 146, ss. 293–314.
Wattles, J. (1996). *The Golden Rule*. New York, Oxford: Oxford University Press.
Weiss, P. (1941). The Golden Rule. *The Journal of Philosophy*, Vol. 38, Nr. 16, ss. 421–430.

Daniel Rönnedal
Filosofiska institutionen
Stockholms universitet
daniel.ronnedal@philosophy.su.se

Den Gyllene Regeln och Intra- och Interpersonella Viljekonflikter

Daniel Rönnedal

Abstrakt
Enligt den s.k. gyllene regeln bör vi behandla andra så som vi själva vill bli behandlade. Denna historiskt mycket inflytelserika princip är fortfarande en av de mest spridda och allmänt accepterade normer som någonsin formulerats. Alla världsreligioner tycks innehålla någon variant av denna maxim och mängder av filosofer har också accepterat den i en eller annan form. Men regeln har också kritiserats och det tycks finnas många möjliga tolkningar som har problematiska konsekvenser. I den här uppsatsen diskuterar jag en mängd argument emot den gyllene regeln som alla i någon mening går ut på att principen tillsammans med vissa andra rimliga antaganden är inkonsistent. Därefter visar jag hur dessa argument kan besvaras och hur den gyllene regeln kan försvaras. I ljuset av vår diskussion tycks vi kunna dra följande slutsatser. (1) Det finns tolkningar av den gyllene regeln som har problematiska konsekvenser och som vi därför troligtvis bör undvika. (2) Inte alla tolkningar av denna princip drabbas av de argument som diskuteras i den här uppsatsen. (3) Då man försöker avgöra värdet hos denna norm bör man försöka fokusera på de bästa möjliga tolkningarna.

1. Introduktion

Enligt den s.k. gyllene regeln bör vi behandla andra så som vi själva vill bli behandlade. Denna historiskt inflytelserika princip är en av världens mest spridda och accepterade normer. Alla världsreligioner, och många andra religioner, tycks innehålla någon variant av denna regel[1], och åtskilliga filosofer med olika moralfilosofiska teorier har också accepterat den i en eller annan form[2]. Men regeln har också kritiserats och det tycks finnas många möjliga tolkningar som har problematiska konsekvenser.[3] I den här uppsatsen diskuterar jag en mängd argument emot den gyllene regeln som alla i någon

[1] Neusner och Chilton (red.) (2008).
[2] Se t.ex. Hobbes (1985), Kapitel XIV, s. 190, Pufendorf (1964), Bok 2, 3:13, Mill (1987), Kapitel 2, s. 28, och Gensler (1996), särskilt Kapitel 5.
[3] Se t.ex. Gensler (2013) för en sammanfattning av några potentiella invändningar. Gensler försvarar en särskild tolkning av den gyllene regeln mot dessa invändningar.

Filosofiska Notiser, Årgång 3, Nr. 2, Augusti 2016, 81–106

mening går ut på att principen tillsammans med vissa andra rimliga antaganden är inkonsistent. Därefter visar jag hur dessa argument kan besvaras och hur den gyllene regeln kan försvaras. I ljuset av vår diskussion tycks vi kunna dra följande slutsatser. (1) Det finns tolkningar av den gyllene regeln som har problematiska konsekvenser och som vi därför troligtvis bör undvika. (2) Inte alla tolkningar av denna princip drabbas av de argument som diskuteras i den här uppsatsen. Med andra ord, det finns tolkningar av den gyllene regeln som undviker de problem vi tar upp i den här artikeln. (3) Då man försöker avgöra värdet hos denna maxim bör man försöka fokusera på de bästa möjliga tolkningarna.

Den här uppsatsen är indelad i fem avsnitt. Avsnitt 2 innehåller en kort introduktion till den gyllene regeln. I Avsnitt 3 och 4 diskuterar jag sammanlagt sju olika argument emot olika versioner av denna princip. Avsnitt 3 innehåller tre argument som utgår ifrån *intrapersonella* viljekonflikter och Avsnitt 4 innehåller fyra argument som utgår ifrån *interpersonella* viljekonflikter. Alla argument bygger i någon mening på att man kan härleda en kontradiktion från någon version av den gyllene regeln om vi antar att det *finns* eller *kan finnas* intra- eller interpersonella viljekonflikter. Även om de olika argumenten liknar varandra, så skiljer de sig åt på väsentliga punkter, vilket motiverar att vi diskuterar dem alla. Problemet med interpersonella viljekonflikter omnämns bl.a. av Richard Whately (1856), och Don Locke (1981). Harry Gensler tar upp en variant av problemet i Gensler (2013), ss. 211–212. Men de få diskussioner som återfinns i litteraturen är relativt skissartade och jag känner inte till någon annan som har diskuterat dem närmare. Flera varianter av de argument jag behandlar tycks vara helt nya. Diskussionen i den här uppsatsen fyller därför ett gap i litteraturen. Avsnitt 5 innehåller en sammanfattning av uppsatsen och några slutsatser.[4]

2. Den gyllene regeln

Enligt den s.k. gyllene regeln bör vi behandla andra så som vi själva vill bli behandlade. Men hur skall man förstå detta och vad betyder de olika uttrycken i denna princip?[5] Den gyllene regeln kan preciseras på en mängd

[4] För mer information om den gyllene regeln, se t.ex. Blackstone (1965), Bruton (2004), Cadoux (1912), Carson (2010), Kapitel 6, Carson (2013), Duxbury (2009), Gensler (1986), Gensler (1996), särskilt Kapitel 5, (2013), Gewirth (1978), Gould (1980), Hare (1963), Hertzler (1934), Hirst (1934), Hoche (1978), Huang (2005), Neusner och Chilton (red.) (2008), Reinikainen (2005), Rönnedal (2015), Singer (1963), Wattles (1996) och Weiss (1941).

[5] Rönnedal (2015) tar upp ett antal frågor som är relevanta då man försöker förstå den gyllene regeln. Se också Rönnedal (2016).

olika sätt och vi kommer i den här uppsatsen att nämna flera olika alternativ. Vi skall börja med att utgå från följande bokstavliga formulering:

> **(□BGR)**. Det är nödvändigt att: Det gäller för alla individer x och y och alla handlingar H att: Om x vill att y utför handling H mot x, så bör x utföra H mot y.

(□BGR) medför t.ex. följande satser: om du vill att din partner är ärlig mot dig, så bör du vara ärlig mot din partner; och om din partner vill att du behandlar henne rättvist, så bör din partner behandla dig rättvist. Vi antar i den här uppsatsen att den gyllene regeln medför den s.k. silverregeln, som säger att vi bör undvika att behandla andra på sätt som vi själva inte vill bli behandlade. Mer precist,

> **(□BSR)**. Det är nödvändigt att: Det gäller för alla individer x och y och alla handlingar H att: Om x vill att y inte utför handling H mot x, så bör x inte utföra H mot y.

(□BSR) medför t.ex. följande satser: om du vill att din granne inte stjäl från dig, så bör du inte stjäla från din granne; om du vill att din skolkamrat inte slår dig, så bör du inte slå din skolkamrat; och om Magdalena vill att Yvonne inte sprider lögner om henne, så bör Magdalena inte sprida lögner om Yvonne. Dessa konsekvenser av (□BGR) och (□BSR) tycks vara rimliga. Men vi skall nedan undersöka några problematiska implikationer av den gyllene regeln tolkad på detta sätt.

Notera att både □BGR och □BSR innehåller uttrycket ”Det är nödvändigt att”. Det finns flera olika typer av nödvändighet: logisk, metafysisk, naturlig, historisk m.m. Men för argumentens giltighet spelar det inte så stor roll hur vi väljer att precisera detta uttryck. Vi antar att nödvändigheten är en s.k. ”S5-nödvändighet”.

3. Argument från *intrapersonella* viljekonflikter

De tre första argumenten vi skall undersöka i den här uppsatsen utgår ifrån att det *finns* eller åtminstone tycks vara *möjligt* att det finns *intrapersonella* viljekonflikter. En *intrapersonell* viljekonflikt är en konflikt som rör en enskild persons vilja, till skillnad från en *interpersonell* viljekonflikt som involverar flera olika personer.

3.1. Argument 1: argumentet från en motsägelsefull vilja

Vårt första argument går ut på att det tycks vara möjligt att det finns personer som har en motsägelsefull vilja. Antag att det är så. Då kan vi – tillsammans med vissa andra rimliga antaganden – härleda en motsägelse från (□BGR). Låt oss betrakta ett konkret exempel.

Antag att vi har två personer, a och b, som lever i ett stormigt förhållande med varandra. Båda har varit otrogna och båda misstänker att den andre har varit otrogen. b frågar nu a om a har varit otrogen. Bör a berätta sanningen för b eller inte? Vad säger den gyllene regeln? Antag att a har en motsägelsefull vilja; a vill att b berättar sanningen för a och a vill att b inte berättar sanningen för a. a vill att b berättar sanningen, eftersom a plågas av sina tvivel och eftersom a tror att det är bättre att sanningen kommer fram än att leva i en lögn. Samtidigt vill a att b inte berättar sanningen, eftersom det kommer att vara så oerhört smärtsamt om det visar sig att b har varit otrogen och eftersom det troligen kommer att innebära det definitiva slutet för förhållandet. Vi skulle också kunna säga att en del av a vill att b berättar sanningen och att en annan del av a vill att b inte berättar sanningen. Detta tycks vara möjligt. Människor tycks ibland vara ambivalenta, de tycks ibland vilja att något är fallet samtidigt som de vill att det inte är fallet. Men om detta är möjligt, kan vi härleda en motsägelse från (□BGR) på följande sätt.

Argument 1

1. $\Diamond(VaTba \wedge Va\neg Tba)$ i w_1 [Antagande]
2. $VaTba \wedge Va\neg Tba$ i w_2 [1, ML]
3. $VaTba$ i w_2 [2, SL]
4. $Va\neg Tba$ i w_2 [2, SL]
5. $VaTba \rightarrow OTab$ i w_2 [□BGR, ML, PL]
6. $Va\neg Tba \rightarrow O\neg Tab$ i w_2 [□BGR, ML, PL]
7. $OTab$ i w_2 [3, 5, SL]
8. $O\neg Tab$ i w_2 [4, 6, SL]
9. $(OTab \wedge O\neg Tab) \rightarrow O(Tab \wedge \neg Tab)$ i w_2 [AG, ML]
10. $O(Tab \wedge \neg Tab)$ i w_2 [7, 8, 9, SL]
11. $O(Tab \wedge \neg Tab) \rightarrow \Diamond(Tab \wedge \neg Tab)$ i w_2 [BK, ML]
12. $\Diamond(Tab \wedge \neg Tab)$ i w_2 [10, 11, SL]
13. $\neg\Diamond(Tab \wedge \neg Tab)$ i w_2 [ML]
14. Falsum [12, 13, SL]

Txy i argumentet ovan står för ”x berättar sanningen för y”. Vx, O och ◇ är satsoperatorer som tar satser som argument och ger satser som värde. VxA läses ”x vill att A”, OA läses ”Det bör vara fallet att A” (”Det är obligatoriskt

att A"), och $\Diamond$A läses "Det är möjligt att A". "$\rightarrow$" är vanlig materiell implikation, "$\wedge$" vanlig konjunktion, och "$\neg$" vanlig negation. w_1 och w_2 står för möjliga världar. "SL" är en förkortning av "satslogik" och innebär att steget är satslogiskt giltigt. "PL" är en förkortning av "(högre ordningens) predikatlogik" och innebär att steget är predikatlogiskt giltigt. "ML" är en förkortning av "modallogik" och innebär att steget är giltigt i alla normala modallogiska S5-system. "BK" står för "Bör kan principen", dvs. följande maxim:

> (BK) Det är nödvändigt att: Det bör vara fallet (Det är obligatoriskt) att A endast om det är möjligt att A. OA $\rightarrow$ $\Diamond$A, för alla A.

"AG" är en förkortning av "Agglomeration", dvs. följande princip:

> (AG) Det är nödvändigt att: Om det bör vara fallet att A och det bör vara fallet att B, så bör det vara fallet att A och B. (OA $\wedge$ OB) $\rightarrow$ O(A $\wedge$ B), för alla A och B.

Argumentet är uppenbart giltigt och vi har endast använt oss av mycket väletablerade "slutledningsregler". Om vi vill förkasta slutsatsen, måste vi alltså ge upp minst en sats i härledningen. Följaktligen måste vi överge minst en av satserna 1, 5, 6, 9, 11. Steg 9 följer ur (AG) och steg 11 följer ur (BK). Om vi överger (AG) kan vi således ge upp steg 9, och om vi överger (BK) kan vi ge upp steg 11. Både (AG) och (BK) är emellertid mycket rimliga principer. (AG) kan bevisas i alla s.k. normala deontiska system och (BK) är en intuitivt mycket rimlig tes. Jag har i andra arbeten försökt försvara uppfattningen att moralen är "konsistent" (se Rönnedal (2012), ss. 73–96 för mer information). Om (AG) och (BK) är sanna, måste vi förkasta 5, 6 och/eller 1. 1 är helt enkelt en formalisering av vårt ursprungliga scenario, och detta tycks vara möjligt. Men kanske är denna intuition inte riktig, kanske är det inte möjligt att det finns personer med en motsägelsefull vilja. I så fall kan vi ge upp steg 1 och hålla fast vid ($\Box$BGR). Personligen är jag dock benägen att tro att det är möjligt att ha en motsägelsefull vilja. Notera att det är tillräckligt för att vårt argument skall gå igenom att det är *möjligt* med en motsägelsefull vilja; argumentet kräver inte att någon *faktisk* har en sådan vilja. Men om steg 1 är sant, måste vi antingen förkasta steg 5 eller 6. Och steg 5 och 6 är logiska implikationer av ($\Box$BGR). Det följer att vi i så fall måste förkasta ($\Box$BGR). Vi kan emellertid inte sluta oss till att vi måste överge den gyllene regeln. För det är inte uppenbart att ($\Box$BGR) är den enda eller bästa tolkningen av denna maxim. Vi skall nu undersöka några andra

möjliga tolkningar av denna princip som undviker argument 1. Jag skall nämna fem möjliga interpretationer.

(i) Vi kan försvaga den gyllene regeln och hävda att den inte är nödvändig. Istället för (□BGR) tolkas den gyllene regeln på följande sätt:

(BGR). Det gäller för alla individer x och y och alla handlingar H att: Om x vill att y utför handling H mot x, så bör x utföra H mot y.

(□BGR) är starkare än (BGR); (□BGR) medför (BGR), men (BGR) medför inte (□BGR). Ett argument som är problematiskt för (□BGR) innebär därför inte nödvändigtvis några problem för (BGR). (□BGR) medför 5 och 6. Men 5 och 6 kan inte härledas ur (BGR). Tolkar vi den gyllene regeln på detta sätt, undviker vi alltså den kontradiktoriska slutsatsen.

Låt mig nämna två möjliga kontraargument. För det första: allt annat lika, bör vi alltid formulera och tro på så starka teorier som möjligt. (□BGR) är starkare än (BGR). Alltså bör vi föredra (□BGR) framför (BGR). Detta är knappast något konklusivt argument för (□BGR), eftersom vi just har sett hur vi kan härleda en motsägelse ur denna princip (tillsammans med andra rimliga premisser). Men om det finns andra möjliga lösningar på argument 1, kan det vara rimligt att överväga dessa istället.

För det andra, det tycks som om vi kan formulera ett giltigt argument som påminner om argument 1 där vi utgår från premissen att det *faktiskt* är så att a vill att b utför handling H mot a och att a vill att b *inte* utför handling H mot a. Och i så fall kan vi på nytt härleda en motsägelse med hjälp av (BGR). Detta kontraargument bygger på antagandet att det inte bara är *möjligt* att det finns någon med en motsägelsefull vilja, utan att det *faktiskt* finns någon med en sådan vilja. Personligen är jag benägen att tro att även detta är fallet (åtminstone vid något tillfälle). I så fall finns det problem även med (BGR).

Det är dock inte uppenbart att detta kontraargument är korrekt. Om det *faktiskt* inte finns någon med en motsägelsefull vilja, så drabbas inte (BGR) av de problem som tycks drabba (□BGR). Det räcker inte med att visa att det är *tänkbart* att det finns någon med en motsägelsefull vilja eller att det är *möjligt* att det finns någon sådan individ för att vederlägga (BGR). Vi måste hitta ett verkligt exempel. Finns det faktiskt någon med en motsägelsefull vilja? Låt mig nämna hur en anhängare av (BGR) skulle kunna förneka detta och på så sätt undvika det aktuella problemet.

En anhängare av (BGR) skulle kunna göra en distinktion mellan en *allt taget i beaktande* vilja och vad vi skulle kunna kalla en *prima facie* vilja. Hon

kan vidare hävda att (BGR) handlar om en allt taget i beaktande vilja och inte en prima facie vilja. (BGR) skall med andra ord tolkas på följande sätt.

> **(BGR′)**. Det gäller för alla individer x och y och alla handlingar H att: Om x (allt taget i beaktande) vill att y utför handling H mot x, så bör x utföra H mot y.

Vi kan acceptera att det finns någon som prima facie vill att A och prima facie vill att inte-A, men förneka att det finns någon som allt taget i beaktande vill att A och allt taget i beaktande vill att inte-A. Att vilja något allt taget i beaktande är någonting annat än att drömma om att A eller att önska att A eller att anse att det finns vissa faktorer som talar för att A. Om man allt taget i beaktande vill att A, så är man fast besluten att A och har en intention att, om möjligt, se till att A. Och det är inte alls uppenbart att det finns någon som har en motsägelsefull vilja i denna mening. I vårt exempel ovan kan vi t.ex. hävda att a (prima facie) vill att b berättar sanningen och att a (prima facie) vill att b inte berättar sanningen. Men a har inte bestämt sig för vad a vill *allt taget* i beaktande. Det är inte så att a allt taget i beaktande vill att b berättar sanningen och att a allt taget i beaktande vill att b inte berättar sanningen; a velar fram och tillbaka. För att vederlägga (BGR′) räcker det inte med att visa att det finns någon, a, som *prima facie* vill att någon, b, utför handling H mot a och *prima facie* vill att det inte är fallet att b utför H mot a. Vi måste hitta någon, a, som *allt taget i beaktande* vill att någon, b, utför handling H mot a, och *allt taget i beaktande* vill att det inte är fallet att b utför H mot a. Det är därför inte säkert att (BGR) (tolkad som (BGR′)) drabbas av det aktuella problemet.

(ii) Vi kan tolka den gyllene regeln som en tumregel. Vi går med på att (□BGR) inte är bokstavligt talat sann. Men den kan ändå vara användbar om de flesta instanser av den är sann. Denna position kan tyckas vara märklig. Om en sats inte är sann, hur kan den då vara användbar? En analogi kanske kan förklara tankegången. Antag att det finns en miljon svanar i världen. Alla dessa svanar är vita, utom en som är svart. Då är satsen: ”Alla svanar är vita” falsk, men den kan ändå vara användbar för att göra förutsägelser om färgen hos enskilda svanar. Från påståendet att alla svanar är vita följer det att den svan Kassandra äger är vit. Det är förstås möjligt att den svan Kassandra äger är svart, men detta är mycket osannolikt. Om vi därför använder tumregeln att alla svanar är vita för att sluta oss till att Kassandras svan är vit, så är det mycket troligt att vår trosföreställning att Kassandras svan är vit är sann. På

liknande sätt kan det förhålla sig med normer som inte är sanna i en strikt mening.

Vi vill förstås helst formulera teorier som är strikt sanna. Så, om det är möjligt att formulera en tolkning av den gyllene regeln som är sann i en strikt mening är en sådan tolkning av den gyllene regeln att föredra framför en tolkning som endast är approximativt sann, allt annat lika. Tumregler och approximativt sanna principer bör emellertid inte föraktas. Filosofer har ibland varit alltför upptagna av att hitta absolut nödvändiga, universella och undantagslösa normer. Men faktum är att de flesta normer som människan har formulerat tycks vara ett slags tumregler som endast är approximativt sanna. Och sådana principer kan likväl vara mycket användbara.

(iii) Vi kan hävda att den gyllene regeln handlar om prima facie plikter och inte allt taget i beaktande plikter. Vi tolkar alltså den gyllene regeln på följande sätt.

> **(□PFGR)**. Det är nödvändigt att: Det gäller för alla individer x och y och alla handlingar H att: Om x vill att y utför handling H mot x, så är det en prima facie plikt att x utför H mot y.

Hur hjälper detta oss att undvika en motsägelse? (AG) är en rimlig princip om vi talar om allt taget i beaktande plikter, men inte om vi refererar till prima facie plikter. Det är möjligt att det är prima facie obligatoriskt att A och att det är prima facie obligatoriskt att B, samtidigt som det är falskt att det är prima facie obligatoriskt att A och B. Steg 10 i argument 1 är då inte berättigat. Därmed går deduktionen inte igenom.

(iv) Vi kan begränsa den gyllene regeln så att den endast handlar om den rationella viljan.

> **(□RGR)**. Det är nödvändigt att: Det gäller för alla individer x och y och alla handlingar H att: Om x vill att y utför handling H mot x och x's vilja är rationell, så bör x utföra H mot y.

Med hjälp av denna princip kan vi inte längre härleda 5 och 6. Det är rimligt att anta att a's vilja inte är rationell om a vill att A och a vill att inte-A. Vi kan kräva att alla fullständigt rationella personer, x, uppfyller följande villkor: $(\forall xA \wedge \forall xB) \rightarrow \forall x(A \wedge B)$ och $\forall xA \rightarrow \Diamond A$, dvs. om x vill att A och x vill att B, så vill x att A och B; och x vill att A endast om det är möjligt att A. Antagandet att a är fullständigt rationell, att a vill att b berättar sanningen och att a vill att b inte berättar sanningen leder då till en kontradiktion, varför

det inte kan vara sant. Argument 1 går alltså inte igenom om vi använder denna tolkning av den gyllene regeln.

(v) Vi kan tolka den gyllene regeln som en s.k. *vid* villkorlig norm och inte en s.k. *snäv* villkorlig norm, dvs. vi antar att O-operatorn i den gyllene regeln har *vid* och inte *snäv* räckvidd. Vi kan med andra ord förstå den gyllene regeln på följande sätt:

> **(□VGR)**. Det är nödvändigt att: Det gäller för alla individer x och y och alla handlingar H att det bör vara fallet att: Om x vill att y utför handling H mot x, så utför x H mot y.

Tolkad på detta sätt är den gyllene regeln ett slags konsistens-princip som talar om för oss hur vi skall vara konsistenta i vilja och handling, men den ger inte någon direkt vägledning om hur vi bör handla. Principen hävdar att följande kombination är otillåten: att vi vill att x utför H mot oss samtidigt som vi inte utför H med x. Om vi tolkar den gyllene regeln på detta sätt, kan vi inte längre härleda steg 5 och 6 från denna maxim. Alltså går argumentet inte igenom givet denna tolkning.

Med hjälp av alla dessa strategier kan vi alltså undvika argument 1. Man kan också tänka sig kombinationer av dessa, t.ex. av (i) och (iv). Vi kan t.ex. försvaga den gyllene regeln så att den inte är en nödvändig sanning och samtidigt begränsa den så att den endast varierar över personer med en rationell vilja. Vilken strategi som är bäst beror förstås också på vilka övriga argument för och emot den gyllene regeln det finns, och är en fråga som vi inte skall gå in på närmare här.

3.2. Argument 2: argumentet från en *direkt* motsägelsefull vilja

Vårt andra argument liknar det första och vår utvärdering av detta argument kan därför vara mer kortfattad. Det tycks vara möjligt att det finns personer som har en *direkt* motsägelsefull vilja (och inte endast en motsägelsefull vilja). Och om det är möjligt, så kan vi – tillsammans med vissa övriga rimliga antaganden – härleda en motsägelse från (□BGR). Låt oss betrakta följande variant av det konkreta exemplet i Avsnitt 3.1.

Antag att vi har två personer, a och b, som befinner sig i exakt samma situation som personerna i Avsnitt 3.1. Den enda skillnaden är att vi antar att a har en *direkt* motsägelsefull vilja, dvs. a vill att b berättar sanningen för a och att b inte berättar sanningen för a. Detta tycks vara möjligt. Men om det är så, kan vi härleda en motsägelse från (□BGR) på följande sätt.

Argument 2
1. $\Diamond$Va(Tba $\wedge \neg$Tba) i w_1 [Antagande]
2. Va(Tba $\wedge \neg$Tba) i w_2 [1, ML]
3. Va(Tba $\wedge \neg$Tba) $\rightarrow$ O(Tab $\wedge \neg$Tab) i w_2 [$\Box$BGR, PL, ML]
4. O(Tab $\wedge \neg$Tab) i w_2 [2, 3, SL]
5. O(Tab $\wedge \neg$Tab) $\rightarrow \Diamond$(Tab $\wedge \neg$Tab) i w_2 [BK, ML]
6. $\Diamond$(Tab $\wedge \neg$Tab) i w_2 [4, 5, SL]
7. $\neg\Diamond$(Tab $\wedge \neg$Tab) i w_2 [ML]
8. Falsum [6, 7, SL]

Argument 2 liknar argument 1, men det är ändå värt att ta upp argument 2 separat, eftersom inte exakt samma lösningsförslag tycks kunna användas i de båda fallen. Lösningsförslag (iii) i Avsnitt 3.1 tycks inte kunna användas i detta fall. Argument 2 innehåller ingen motsvarighet till steg 7–10 i argument 1. Så även om det skulle vara sant att (AG) inte gäller för prima facie plikter och den gyllene regeln antas handla om prima facie plikter, så kan vi inte undvika argument 2.

Man skulle kunna hävda att (BK) endast gäller för allt taget i beaktande plikter och inte för prima facie plikter. Om detta är riktigt, så är det möjligt att det är en prima facie plikt att A trots att det inte är möjligt att A. Antag att den gyllene regeln gäller för prima facie plikter. Då följer inte längre steg 6 i argument 2 från 4 och 5. Och på så sätt skulle vi kunna undvika den problematiska slutsatsen i argument 2. Personligen är jag emellertid benägen att tro att (BK) gäller även för prima facie plikter. Det är därför tveksamt om detta är en tillfredsställande invändning mot argument 2.

Kanske är det något mer rimligt att ifrågasätta premiss 1 i argument 2 än premiss 1 i argument 1. Kanske finns det *faktisk* ingen som har en direkt motsägelsefull vilja av det slag som premiss 1 i argument 2 antar. Men det tycks ända vara *möjligt*. Och det räcker för att argumentet skall gå igenom.

Alla andra lösningsförslag som diskuterades i Avsnitt 3.1 kan emellertid även tillämpas för att undvika slutsatsen i argument 2. Argument 2 utgör därför inte ett problem för *alla* möjliga tolkningar av den gyllene regeln.

3.3. Argument 3: argumentet från en splittrad vilja

Vårt tredje argument är en version av ett argument som har diskuterats av Gensler. Så här skriver Gensler i en fotnot på s. 211 i Gensler (2013).

> Suppose you're a broccoli-hating waiter and Becky orders broccoli. By Pyrite 1 [If you want X to do A to you then do A to X], you shouldn't serve her broccoli (since you don't want others to serve you

broccoli) and you should serve her broccoli (since you want others to bring you what you order).

Låt oss uttrycka en mer precis variant av detta argument. Vi antar att följande scenario är möjligt. Du är en servitris på en restaurang. Becky beställer broccoli av dig. Du gillar inte broccoli. Du vill att det inte är fallet att Becky serverar dig broccoli (eftersom du inte gillar broccoli). Du vill att om det vore fallet att du skulle beställa broccoli av Becky, så skulle Becky servera dig broccoli (du vill att andra serverar dig vad du beställer). Becky beställer broccoli av dig. Om detta är sant, så tycks vi kunna härleda en motsägelse ur (□BGR) givet vissa andra rimliga antaganden. Låt oss uttrycka detta med hjälp av formella symboler. ”Sxy” läses ”x serverar y broccoli”, och ”Bxy” läses ”x beställer broccoli av y”, ”d” refererar till dig och ”b” refererar till Becky. ”A □→ B” läses ”Om A vore fallet, så skulle B vara fallet”.

Argument 3
1. Vd¬Sbd → O¬Sdb [(□BGR), PL, ML]
2. Vd¬Sbd [Antagande]
3. O¬Sdb [1, 2, SL]
4. Vd(Bdb □→ Sbd) → O(Bbd □→ Sdb) [(□BGR), PL, ML]
5. Vd(Bdb □→ Sbd) [Antagande]
6. O(Bbd □→ Sdb) [4, 5, SL]
7. O(Bbd □→ Sdb) → O(Bbd → Sdb) [KD]
8. O(Bbd → Sdb) [6, 7, SL]
9. Bbd [Antagande]
10. Bbd → □Bbd [HN]
11. □Bbd [9, 10, SL]
12. (O(Bbd → Sdb) ∧ □Bbd) → OSdb [AD]
13. OSdb [8, 11, 12, SL]
14. OSdb ∧ O¬Sdb [3, 13, SL]
15. (OSdb ∧ O¬Sdb) → O(Sdb ∧ ¬Sdb) [AG]
16. O(Sdb ∧ ¬Sdb) [14, 15, SL]
17. O(Sdb ∧ ¬Sdb) → ◇(Sdb ∧ ¬Sdb) [BK]
18. ◇(Sdb ∧ ¬Sdb) [16, 17, SL]
19. ¬◇(Sdb ∧ ¬Sdb) [ML]
20. Falsum [18, 19]

Den här deduktionen liknar tidigare härledningar. Men argument 3 innehåller vissa steg som inte har några motsvarigheter i argument 1 och 2. De viktigaste nya stegen är 7, 10 och 12. Övriga nya satser i härledningen följer

med hjälp av väletablerade härledningsregler. 7, 10 och 12 är logiska konsekvenser av vissa bakomliggande allmänna antaganden, nämligen (KD) (”Kontrafaktisk deontisk logik”), (HN) (”Historisk nödvändighet”) och (AD) (”Aletisk deontisk logik”).

> (KD) Om det är obligatoriskt att om A vore fallet så skulle B vara fallet, så är det obligatoriskt att om A så B. $O(A \boxdot\!\!\rightarrow B) \rightarrow O(A \rightarrow B)$, för alla A och B.
>
> (HN) Om A är en atomär sats, så är det historiskt nödvändigt att A. Om A är atomär, så $\Box A$, för alla A.
>
> (AD) Om det är obligatoriskt att A implicerar B och det är (historiskt) nödvändigt att A, så är det obligatoriskt att B. $(O(A \rightarrow B) \wedge \Box A) \rightarrow OB$, för alla A och B.

Eftersom argument 3 förutsätter fler satser än argument 1 och 2, finns det ett antal nya strategier att kritisera denna deduktion. Låt oss se närmare på dessa.

(i) Vi kan överge (KD). Utan denna princip kan vi inte härleda steg 7, $O(Bbd \boxdot\!\!\rightarrow Sdb) \rightarrow O(Bbd \rightarrow Sdb)$, i argument 3. Och utan detta steg är argumentet inte giltigt.

(KD) tycks emellertid vara en rimlig princip. Enligt de mest populära tolkningarna av kontrafaktiska villkorssatser är det logiskt sant att $(A \boxdot\!\!\rightarrow B)$ implicerar $(A \rightarrow B)$.[6] Och i alla s.k. normala deontiska system gäller det att om ”$A \rightarrow B$” är ett teorem, så är ”$OA \rightarrow OB$” ett teorem. Kombinerar vi ”kontrafaktisk logik” med ”deontisk logik”, kan vi alltså härleda (KD). Vi tycks således ha goda skäl att hålla (KD) för sann. Strategi (i) kan därför ifrågasättas.

(ii) Det är möjligt att förkasta (AD) och om vi gör det kan vi inte härleda steg 12, $(O(Bbd \rightarrow Sdb) \wedge \Box Bbd) \rightarrow OSdb$, från denna princip. Är det verkligen sant att om det är obligatoriskt att A implicerar B och det är historiskt nödvändigt att A, så är det obligatoriskt att B? Om steg 12 inte är sant, så går argument 3 inte igenom.

(AD) kan emellertid bevisas i alla s.k. normala aletisk-deontiska system som innehåller den s.k. mål medel principen[7], och vi tycks därmed ha mycket

[6] Se t.ex. Stalnaker (1968) och Lewis (1973).

[7] Enligt mål medel principen gäller det att om det bör vara fallet att A och det är nödvändigt att A implicerar B, så bör det vara fallet att B.

goda skäl att acceptera (AD). Personligen tycker jag att (AD) är en intuitivt tilltalande maxim. Denna strategi kan alltså ifrågasättas.

(iii) Vi kan överge (HN). (HN) är onekligen en mer kontroversiell princip än många andra som används i argumentet och den är inte uppenbart korrekt. Om vi förkastar (HN), kan vi inte längre härleda steg 10, $Bbd \rightarrow \Box Bbd$, från denna princip. Och utan steg 10 faller argumentet. Är det sant att allt som faktiskt är fallet är historiskt bestämt? Om det är det, så misslyckas denna invändning mot argument 3. Det är rimligt att acceptera (HN) om det är rimligt att anta att nuet är historiskt bestämt. Nuet kan vara historiskt bestämt även om framtiden är öppen.

(iv) Ett annat sätt att undvika den kontradiktoriska slutsatsen i argument 3 är att begränsa den gyllene regeln så att den endast gäller för enkla handlingar och en kategorisk vilja, och inte för komplexa handlingar och en villkorlig vilja. Vi kan, med andra ord, formulera den gyllene regeln på följande sätt:

> **(□KGR)**. Det är nödvändigt att: Det gäller för alla individer x och y och alla *enkla* handlingar H att: Om x vill att y utför handling H mot x, så bör x utföra H mot y.

Tanken här är att vi *inte* kan instansera H med vilka handlingspredikat som helst. Om vi använder denna tolkning av den gyllene regeln, kan vi inte längre härleda steg 4 ifrån denna princip. Och utan steg 4 så faller argument 3.

Vi behöver alltså inte överge den gyllene regeln. Men detta svar är likväl problematiskt, även om vi inte längre kan härleda en motsägelse från (□KGR) på samma sätt som ovan. För om vi använder denna tolkning av den gyllene regeln tycks den medföra vissa kontraintuitiva påståenden i vårt exempel. Vi kan inte bevisa steg 4 från (□KGR), men vi kan fortfarande sluta oss till steg 1. Den gyllene regeln medför då att du *inte* bör servera Becky broccoli i vårt exempel. Men rent intuitivt så är det sant att du *bör* servera Becky broccoli. Om du arbetar som servitris, är det väsentliga inte vad *du* gillar och inte gillar utan vad *kunden* beställer.

(v) Ytterligare ett lösningsförslag är att kräva en ”samma situation-klausul” i den gyllene regeln. Från det faktum att du vill att Becky *inte* skall servera dig broccoli följer det då *inte* att du *inte* bör servera Becky broccoli. Det enda som är väsentligt är vad du skulle vilja om du befann dig i Beckys situation (och Becky befann sig i din situation) (eller vad du nu vill skulle

vara fallet om du befann dig i Beckys situation (och Becky befann sig i din situation)). Denna idé kan preciseras på minst två olika sätt:

> **(□PGR)**. Det är nödvändigt att: Det gäller för alla individer x och y och alla handlingar H att: Om det är sant att om x befann sig i y's situation (och y i x's situation) så skulle x vilja att y utför handling H mot x, så bör x utföra H mot y.
>
> **(□SGR)**. Det är nödvändigt att: Det gäller för alla individer x och y och alla handlingar H att: Om x vill att om x skulle befinna sig i y's situation (och y i x's situation) så skulle y utföra handling H mot x, så bör x utföra H mot y.

Även om alla nödvändiga sanningar är ekvivalenta, så är innehållen i dessa preciseringar inte ekvivalenta. Skillnaden mellan principerna är subtil, men betydelsefull. I (□SGR) har vilje-operatorn vid räckvidd, i (□PGR) har den snäv räckvidd. Jag kallar (□PGR) för "(□PGR)" eftersom innehållet i denna maxim tycks vara ekvivalent med den s.k. platinaregeln, som säger att om y vill att x utför H mot y, så bör x utföra H mot y. Om vi använder någon av dessa tolkningar av den gyllene regeln, kan vi inte längre härleda steg 1 i argument 3 från denna princip. Och utan steg 1 faller argumentet. Det kan finnas andra oberoende skäl att anta att den gyllene regeln bör begränsas på detta sätt. Lösningsförslaget förefaller därför vara mycket intressant.

För att undvika argument 3 kan man alltså använda någon av strategierna (i)–(v) ovan. Förslag (i)–(iv) förefaller vara problematiska av olika anledningar, men (v) är ett förslag som är prima facie tilltalande. Dessutom tycks det som om man kan använda åtminstone några av strategierna från Avsnitt 3.1. Om vi t.ex. antar att den gyllene regeln handlar om prima facie plikter, kan vi förkasta steg 15 i argument 3. På så sätt kan vi undvika den kontradiktoriska slutsatsen. Olika kombinationer av de enskilda lösningarna är också möjliga. Däremot är det inte säkert att lösningsförslag (iv) i Avsnitt 3.1 kan användas för att undvika slutsatsen i argument 3, åtminstone inte om inte detta förslag modifieras. Låt mig säga lite mer om detta.

Det är rimligt att anta att någon som är fullständigt förnuftig inte både vill att A och vill att inte-A. Men det aktuella exemplet tycks inte innehålla någon sådan "uppenbar" motsägelse. Vi har endast antagit att du vill att Becky inte serverar dig broccoli och att du vill att om du skulle beställa broccoli så skulle Becky servera dig broccoli, i symboler: Vd¬Sbd och Vd(Bdb □→

Sbd). Men detta tycks inte vara irrationellt, åtminstone inte på ett uppenbart sätt; {Vd¬Sbd, Vd(Bdb □→ Sbd)} tycks vara konsistent. Är det inte möjligt att din vilja är rationell trots att du vill att det inte är fallet att Becky serverar dig broccoli samtidigt som du vill att om du hade beställt broccoli från Becky, så hade Becky serverat broccoli till dig? Denna vilja tycks inte vara oförnuftig. I så fall kan argument 3 inte undvikas genom att begränsa viljan på det sätt som vi gjorde i lösningsförlag (iv) i Avsnitt 3.1.

Eftersom {Vd¬Sbd, Vd(Bdb □→ Sbd)} åtminstone inte på ett uppenbart sätt är inkonsistent, kallar jag det här argumentet för ”argumentet från en *splittrad* vilja” och inte ”argumentet från en *motsägelsefull* vilja”. Viljan är splittrad i den meningen att din kategoriska vilja pekar i en annan riktning än din hypotetiska vilja. Du vill att Becky *inte* serverar dig broccoli, samtidigt som du vill att *om* du hade beställt broccoli så skulle Becky ha serverat dig broccoli. Dessutom vill du *inte* bli behandlad på samma sätt av Becky som Becky vill bli behandlad av dig. Du vill att Becky *inte* serverar dig broccoli, men Becky vill att du serverar henne broccoli. Det tycks som om vi kan härleda en motsägelse ur {Vd¬Sbd, Vd(Bdb □→ Sbd)} endast tillsammans med den gyllene regeln. Det är inte viljan i sig som är problematisk. I så fall tycks det som om det inte hjälper att anta att den gyllene regeln handlar om en rationell vilja för att undvika argument 3.[8]

Detta problem påverkar emellertid inte lösningsförlag (v). Utgår vi ifrån (□SGR) eller (□PGR) kan vi förkasta steg 1 i argument 3. Och i så fall går argument 3 inte igenom.

4. Argument från *interpersonella* viljekonflikter

Våra fyra sista argument utgår ifrån att det tycks finnas *interpersonella* viljekonflikter, eller att sådana konflikter åtminstone tycks vara *möjliga*. Om det är riktigt, så kan vi – tillsammans med vissa andra rimliga antaganden – härleda en motsägelse ur vissa former av den gyllene regeln. Det grundläggande problemet ifråga finns omnämnt i litteraturen. Vi skall börja med att titta på Richard Whatelys, Don Lockes, och Harry Genslers formuleringar. Sedan skall vi undersöka fyra olika möjliga preciseringar av resonemanget.

[8] Det finns sätt att modifiera denna typ av lösning. Om vi antar att en förnuftig vilja också är opartisk i en viss mening, kan vi undvika argumentet. Ett alternativ är att vi explicit anger att den gyllene regeln endast kan tillämpas på en opartisk vilja. Jag skall inte i detalj gå in på hur denna lösning kan utarbetas eftersom det skulle ta oss allt för långt bort från den här uppsatsens huvudtema.

Så här uttrycker Whately problemet:

> … if you had to decide between two parties who were pleading their cause before you, you might consider that *each* of them wished for a decision in his *own* favor. And how, then, you might ask, would it be possible to apply the rule ? since in deciding *for* the one party you could not but decide *against* the other. A literal compliance with the rule, therefore, would be, in such a case, *impossible*. Whately (1856, s. 26)

Locke nämner ett par problem med den gyllene regeln i Locke (1981). Det är Lockes andra argument som intresserar oss i den här uppsatsen. Så här formulerar Locke detta problem:

> The second difficulty is that adhering to the Golden Rule in respect of one person may require us to break the Golden Rule in respect of another: doing unto one person as we would have him do unto us may involve our doing unto some other person as we would not have that other person do unto us. The Golden Rule inevitably breaks down in those situations, and there will be many, where the preferences of the different parties conflict with each other. In fact, it will break down in just those situations where moral problems most commonly arise. Locke (1981, s. 536)

Gensler ger oss lite mer kött på benen eftersom han diskuterar ett konkret exempel på en interpersonell viljekonflikt. Den grundläggande poängen är emellertid densamma.

> GR can issue contradictory commands when your action affects X and Y – since you may want A done when you put yourself in X's place but want A not done when you put yourself in Y's place.
> Suppose I own a store and need to hire just one worker. Alice and Betty apply, and I must choose between them. Both are qualified, but Alice more so. When I consider myself in Alice's place, I desire that I be hired instead of Betty; so GR tells me to hire Alice and not Betty. But when I consider myself in Betty's place, I desire that I be hired instead of Alice; so GR tells me to hire Betty and not Alice. So GR gives contradictory commands… When our actions affect several

> people, who gain or lose depending on our choice, GR may give contradictory commands. Gensler (2013, s. 211–212)

4.1. Argument 4:
argumentet (emot (□SGR)) från en *indirekt* interpersonell viljekonflikt

Vi skall nu undersöka fyra möjliga preciseringar av detta grundläggande problem. Enligt de två första argumenten finns det en *indirekt* konflikt mellan olika personers viljor, och enligt de två andra argumenten finns det en *direkt* konflikt mellan olika personers viljor.

Argument 4
1. Om jag vill att om jag befann mig i Alices situation och Alice befann sig i min situation så skulle Alice anställa mig, så bör jag anställa Alice. [(□SGR)]
2. Om jag vill att om jag befann mig i Bettys situation och Betty befann sig i min situation så skulle Betty anställa mig, så bör jag anställa Betty. [(□SGR)]
3. Jag vill att om jag befann mig i Alices situation och Alice befann sig i min situation så skulle Alice anställa mig. [Antagande]
4. Jag bör anställa Alice. [1, 3, SL]
5. Jag vill att om jag befann mig i Bettys situation och Betty befann sig i min situation så skulle Betty anställa mig. [Antagande]
6. Jag bör anställa Betty. [2, 5, SL]
7. Om jag bör anställa Alice och jag bör anställa Betty, så bör jag anställa Alice och Betty. [AG]
8. Jag bör anställa Alice och Betty. [4, 6, 7, SL]
9. Jag bör anställa Alice och Betty endast om det är möjligt att jag anställer Alice och Betty. [BK]
10. Det är möjligt att jag anställer Alice och Betty. [8, 9, SL]
11. Det är inte möjligt att jag anställer Alice och Betty. [Antagande]
12. Falsum. [10, 11]

Detta argument innebär problem för (□SGR). Steg 1 och steg 2 följer ur (□SGR). Steg 3 och steg 5 är ”konsekvenser” av beskrivningen av vårt exempel. Steg 11 har ingen direkt motsvarighet i tidigare argument. I Genslers formulering av argumentet förekommer påståendet att jag måste välja mellan Alice och Betty. Detta påstående kan tolkas på lite olika sätt. Enligt en möjlig interpretation är det ekvivalent med steg 11. Övriga satser i argumentet bevisas i princip på samma sätt som tidigare. Det är relativt enkelt att formulera ett liknande argument emot (□PGR) (se argument 5 nedan). Eftersom argument 4 innehåller flera nya satser, som inte har någon

motsvarighet i tidigare argument, kan detta argument också potentiellt kritiseras på ett antal nya punkter.

Flera av de lösningsförslag som tidigare har presenterats kan även användas för att undvika argument 4, t.ex. det förslag som går ut på att den gyllene regeln handlar om prima facie plikter. Tolkar vi den gyllene regeln på detta sätt, kan vi endast dra slutsatsen att det är prima facie obligatoriskt att jag anställer Alice och prima facie obligatoriskt att jag anställer Betty. Från detta följer det inte att det är prima facie obligatoriskt att jag anställer Alice och Betty. Inte heller följer det någonting om vad jag har för plikt allt taget i beaktande.

I det här skedet kan det vara värt att ta upp ett potentiellt problem med denna lösning. Tolkas den gyllene regeln på detta sätt, kan den inte ge någon vägledning i många fall. Bör jag t.ex. anställa Alice eller Betty? Om den gyllene regeln endast handlar om prima facie plikter ger den ingen entydig vägledning. Och många situationer tycks påminna om denna. Finns det inte nästan alltid interpersonella viljekonflikter som man bör ta hänsyn till?

Den här invändningen är emellertid inte konklusiv. I många fall kan det finnas andra faktorer i omgivningen som tillsammans med den gyllene regeln ger vägledning. Man kan hävda att man måste väga in alla faktorer som är relevanta i en viss situation för att komma fram till vad som är allt taget i beaktande obligatoriskt. I Genslers exempel är t.ex. Alice mer kvalificerad än Betty. Detta talar för att jag allt taget i beaktande bör anställa Alice och inte Betty. Liknande faktorer kan finnas i andra situationer.

(i) Låt oss nu undersöka en möjlig lösning som är ”unik” för argument 4. Vi har antagit att det inte är möjligt att ge jobbet både till Alice och till Betty. Men det är inte uppenbart att detta är sant. Om vi förkastar denna premiss, faller argumentet. Kanske *bör* jag inte anställa båda, men att hävda detta är inte detsamma som att hävda att det är *omöjligt*. Kanske måste jag inte välja mellan Alice och Betty. Kanske kan Alice och Betty dela på tjänsten; kanske kan de arbeta halva tiden var.

Vi kan emellertid kringgå detta kontraargument om vi antar att det *inte bör* vara fallet att jag anställer både Alice och Betty. Då kan vi med hjälp av steg 8 i argument 4 direkt härleda en kontradiktion. Även om det inte är *omöjligt* att jag anställer både Alice och Betty, så *bör* jag kanske inte göra det, eftersom jag inte har råd (låt oss stipulera att det förhåller sig så). Det skulle driva min firma i konkurs, det skulle ruinera mig och göra både Alice och Betty arbetslösa.

Vi kan också formulera en variant av argument 4 som innehåller en direkt interpersonell konflikt. Detta för oss över till argument 6 (se nedan). Först skall vi emellertid se hur vi kan formulera ett liknande argument emot (□PGR) snarare än emot (□SGR).

4.2. Argument 5:
argumentet (emot (□PGR)) från en *indirekt* interpersonell viljekonflikt

Argument 5
Vi byter ut steg 1, 2, 3, och 5 i argument 4 ovan mot 1', 2', 3', och 5' nedan. Övriga steg är desamma som i argument 4. Resultatet är argument 5.
1'. Om det är sant att om jag befann mig i Alices situation och Alice befann sig i min situation så skulle jag vilja att Alice anställer mig, så bör jag anställa Alice. [(□PGR)]
2'. Om det är sant att om jag befann mig i Bettys situation och Betty befann sig i min situation så skulle jag vilja att Betty anställer mig, så bör jag anställa Betty. [(□PGR)]
3'. Om jag befann mig i Alices situation och Alice befann sig i min situation så skulle jag vilja att Alice anställer mig. [Antagande]
5'. Om jag befann mig i Bettys situation och Betty befann sig i min situation så skulle jag vilja att Betty anställer mig. [Antagande]

Detta argument påminner om argument 4, och i stort sett samma invändningar som kan riktas mot argument 4 kan också riktas mot argument 5. Vi behöver därför inte säga mer om argument 5 i den här uppsatsen.

4.3. Argument 6:
argumentet (emot (□SGR)) från en *direkt* interpersonell viljekonflikt
Jag kallade argument 4 för ett argument från en *indirekt interpersonell* viljekonflikt. Anledningen till detta är att vi tänker oss att Alice vill att jag anställer Alice och Betty vill att jag anställer Betty samtidigt som det inte är möjligt att anställa båda. Att detta argument handlar om en *interpersonell* konflikt är lätt att se, eftersom det inte nödvändigtvis tycks vara fallet att Alices vilja *i sig* är motsägelsefull eller Bettys vilja *i sig* är motsägelsefull. Att det är en *indirekt* konflikt beror på att vi har antagit att det är *omöjligt* att jag anställer både Alice och Betty. Om det inte vore omöjligt för mig att anställa båda, skulle det inte finnas någon konflikt mellan Alices vilja och Bettys vilja. Att det rör sig om en *konflikt* beror på att det inte är möjligt att både Alices vilja och Bettys vilja tillfredsställs.

Argument 6 och 7 påminner om argument 4 och 5. Men i argument 6 och 7 antar vi en *direkt* viljekonflikt. Vi antar att Alice inte bara vill att jag skall anställa Alice utan även att jag *inte* skall anställa Betty, och att Betty inte bara vill att jag skall anställa Betty utan även att jag *inte* skall anställa Alice. Då kan vi formulera argument 6 på följande sätt.

Argument 6
1. Om jag vill att om jag befann mig i Alices situation och Alice befann sig i min situation så skulle Alice anställa mig och inte Betty, så bör jag anställa Alice och inte Betty. [(□SGR)]
2. Om jag vill att om jag befann mig i Bettys situation och Betty befann sig i min situation så skulle Betty anställa mig och inte Alice, så bör jag anställa Betty och inte Alice. [(□SGR)]
3. Jag vill att om jag befann mig i Alices situation och Alice befann sig i min situation så skulle Alice anställa mig och inte Betty. [Antagande]
4. Jag bör anställa Alice och inte Betty. [1, 3, SL]
5. Jag vill att om jag befann mig i Bettys situation och Betty befann sig i min situation så skulle Betty anställa mig och inte Alice. [Antagande]
6. Jag bör anställa Betty och inte Alice. [2, 5, SL]
7. Om jag bör anställa Alice och inte Betty, så bör jag inte anställa Betty. [OD]
8. Jag bör inte anställa Betty. [4, 7, SL]
9. Om jag bör anställa Betty och inte Alice, så bör jag anställa Betty. [OD]
10. Jag bör anställa Betty. [6, 9, SL]
11. Om jag bör anställa Betty och det bör vara fallet att jag inte anställer Betty, så bör det vara fallet att jag anställer Betty och att jag inte anställer Betty. [AG]
12. Det bör vara fallet att jag anställer Betty och att jag inte anställer Betty. [8, 10, 11, SL]
13. Det bör vara fallet att jag anställer Betty och att jag inte anställer Betty endast om det är möjligt att jag anställer Betty och att jag inte anställer Betty. [BK]
14. Det är möjligt att jag anställer Betty och att jag inte anställer Betty. [12, 13, SL]
15. Det är inte möjligt att jag anställer Betty och att jag inte anställer Betty. [ML]
16. Falsum [14, 15]

(i) Detta argument innehåller ett antal nya steg som kan ifrågasättas. Argumentets hållbarhet beror framför allt på två nya steg, steg 7 och 9. Steg 7 och 9 följer ur principen (OD) (O-Distribution), dvs. från följande sats:

> (OD) Det är nödvändigt att: Om det bör vara fallet att A och B, så bör det vara fallet att A och det bör vara fallet att B. $O(A \wedge B) \rightarrow (OA \wedge OB)$, för alla A och B.

Så, om vi förkastar (OD) kan vi inte längre härleda steg 7 och steg 9 i argument 6 och utan dessa steg går argumentet inte igenom. (OD) är emellertid en intuitivt mycket rimlig princip och den kan också bevisas i alla s.k. normala deontiska system. Detta lösningsförslag kan därför ifrågasättas.

Argument 6 innehåller en *direkt interpersonell* viljekonflikt eftersom följande mängd satser är inkonsistent: {Jag anställer Alice och inte Betty, Jag anställer Betty och inte Alice}. Innehållet i Alices vilja är därför inkonsistent med innehållet i Bettys vilja. Det är inte logiskt möjligt att båda deras viljor tillfredsställs. Steg 15 kan bevisas i alla s.k. normala modallogiska system.

4.4. Argument 7: argumentet (emot (□PGR)) från en *direkt* interpersonell viljekonflikt

Argument 7

Vi byter ut steg 1, 2, 3, och 5 i argument 6 ovan mot 1', 2', 3', och 5' nedan. Övriga steg är desamma som i argument 6. Resultatet är argument 7.

1'. Om det är sant att om jag befann mig i Alices situation och Alice befann sig i min situation så skulle jag vilja att Alice anställer mig och inte Betty, så bör jag anställa Alice och inte Betty. [(□PGR)]

2'. Om det är sant att om jag befann mig i Bettys situation och Betty befann sig i min situation så skulle jag vilja att Betty anställer mig och inte Alice, så bör jag anställa Betty och inte Alice. [(□PGR)]

3'. Om jag befann mig i Alices situation och Alice befann sig i min situation så skulle jag vilja att Alice anställer mig och inte Betty. [Antagande]

5'. Om jag befann mig i Bettys situation och Betty befann sig i min situation så skulle jag vilja att Betty anställer mig och inte Alice. [Antagande]

Detta argument påminner om argument 6. I princip kan samma invändningar som kan riktas mot argument 6 också riktas mot argument 7. Ett antal lösningsförslag som redan har nämnts kan användas för att undvika slutsatsen

i argument 6 och 7. Vi kan t.ex. anta att den gyllene regeln handlar om prima facie och inte allt taget i beaktande plikter.

Ett annat möjligt svar på argumenten i Avsnitt 4 utgår ifrån förslag (v) i Avsnitt 3.1. Enligt detta förslag är den gyllene regeln en s.k. *vid* plikt, inte en *snäv* plikt. Det här förslaget kan kombineras med andra förslag; vi kan t.ex. även kräva att den gyllene regeln innehåller en samma situation-klausul. Då kan denna princip formuleras på följande sätt:

> **(□SVGR)**. Det är nödvändigt att: Det gäller för alla individer x och y och alla handlingar H att det bör vara fallet att: Om x vill att om x befann sig i y's situation (och y i x's situation) så skulle y utföra handling H mot x, så utför x H mot y.

Om vi vidare antar att (HN) inte gäller för satser som uttalar sig om vad någon vill, så kan vi använda denna version av den gyllene regeln istället för (□PGR) och (□SGR) för att undvika alla argument i Avsnitt 4. Utan (□PGR) och (□SGR) kan vi inte härleda vissa väsentliga premisser i de olika argumenten. Från (□SVGR) följer det att det bör vara fallet att om jag vill att om jag befann mig i Alices situation (och Alice befann sig i min situation) så skulle Alice anställa mig, så anställer jag Alice. Det följer också att det bör vara fallet att om jag vill att om jag befann mig i Bettys situation (och Betty befann sig i min situation) så skulle Betty anställa mig, så anställer jag Betty. Antag att det bör vara fallet att jag inte anställer både Alice och Betty (eller att det är omöjligt att jag anställer båda och att den s.k. mål medel principen är sann). Då följer det att det bör vara fallet att jag inte både vill att om jag befann mig i Alices situation (och Alice befann sig i min situation) så skulle Alice anställa mig och vill att om jag befann mig i Bettys situation (och Betty befann sig i min situation) så skulle Betty anställa mig. Om jag vill att om jag befann mig i Alices situation (och Alice befann sig i min situation) så skulle Alice anställa mig samtidigt som jag vill att om jag befann mig i Bettys situation (och Betty befann sig i min situation) så skulle Betty anställa mig, så kommer jag då att bryta mot en norm. Men detta är förstås inte detsamma som att vi kan härleda en motsägelse ur våra antaganden. Det är inte logiskt omöjligt att bryta mot normer. För att undvika att bryta mot denna härledda norm måste den gyllene regeln s.a.s. satisfieras i relation till alla individer.

Ett potentiellt problem med denna lösning är att (□SVGR) inte omedelbart är handlingsvägledande. (□SVGR) är endast ett slags konsistens-princip som handlar om hur vi måste uppnå konsistens mellan vår vilja och

våra handlingar. Detta problem är emellertid inte konklusivt. Det är inte uppenbart att den gyllene regeln måste vara direkt handlingsguidande. Den gyllene regeln är intressant även som ett slags konsistens-princip.

(i) Låt mig slutligen nämna ett möjligt ”unikt” lösningsförslag på argumenten i Avsnitt 4, utan att i detalj gå in på hur detta skulle kunna utvecklas. Man skulle kunna hävda att den gyllene regeln kan tillämpas endast på personer som *inte* är inbegripna i en interpersonell viljekonflikt. Om denna idé kan göras begriplig och försvaras, så tycks det som om argumenten i Avsnitt 4 skulle kunna bemötas. Uppenbarligen skulle mycket mer behöva sägas för att försvara en sådan lösning. Men jag skall nöja mig med att omnämna detta förslag i den här uppsatsen.

5. Slutsats

Enligt den s.k. gyllene regeln bör vi behandla andra så som vi själva vill bli behandlade. Detta är en mycket populär princip, men regeln har också kritiserats. I den här uppsatsen har jag diskuterat en rad argument emot den gyllene regeln som alla i någon mening går ut på att principen tillsammans med vissa andra rimliga antaganden är inkonsistent. Sammanlagt undersökte jag sju olika argument. Därefter försökte jag visa hur dessa argument kan besvaras och hur den gyllene regeln kan försvaras. I ljuset av vår diskussion tycks vi kunna dra följande slutsatser. (1) Det finns tolkningar av den gyllene regeln som har problematiska konsekvenser och som vi därför troligtvis bör undvika. (2) Inte alla tolkningar av denna princip drabbas av de argument som diskuteras i den här uppsatsen. (3) Då man försöker avgöra värdet hos denna norm bör man försöka fokusera på de bästa möjliga tolkningarna.

Låt mig slutligen säga någonting mera generellt om de olika lösningsförslagen i den här uppsatsen.

Det försvar som går ut på att tolka den gyllene regeln som en tumregel kan användas för att ”undvika” alla argument i den här uppsatsen. Denna ”lösning” innebär dock att vi antar att den gyllene regeln inte är bokstavligt talat sann. Allt annat lika är det förstås intressantare om vi kan hitta någon tolkning av den gyllene regeln som är bokstavligt talat sann.

Det förslag som hävdar att den gyllene regeln handlar om prima facie plikter och inte allt taget i beaktande plikter ”löser” alla argument förutom argumentet från en direkt motsägelsefull vilja (argument 2). Den tolkning av den gyllene regeln som förslaget bygger på är därför helt klart intressant.

Det försvar som går ut på att ”begränsa” den gyllene regeln till att handla om en rationell vilja kan användas för att ”undvika” argument 1 och 2. Om

man antar att rationalitetsbegreppet innefattar opartiskhet, kan man eventuellt även lösa vissa andra argument. Detsamma gäller om man ”utökar” rationalitetsbegreppet så att det innefattar interpersonell konsistens. Jag har dock inte i detalj försökt utarbeta dessa förslag i den här uppsatsen.

Det förslag som bygger på att vi inför en samma situation-klausul förefaller i sig endast lösa argument 3. Det kan därför tyckas som om detta förslag är relativt ointressant. Det kan emellertid finnas andra goda skäl att införa en sådan klausul. Tillsammans med andra modifieringar av den gyllene regeln är förslaget mycket intressant.

Om vi kombinerar det förslag som går ut på att den gyllene regeln handlar om en rationell vilja med det förslag som hävdar att den gyllene regeln handlar om prima facie plikter och inte allt taget i beaktande plikter, kan vi formulera en version av den gyllene regeln som undviker alla argument i den här uppsatsen. Denna form av den gyllene regeln ser ut på följande sätt:

> **(□RPFGR)**. Det är nödvändigt att: Det gäller för alla individer x och y och alla handlingar H att: Om x vill att y utför handling H mot x och x’s vilja är rationell, så är det en prima facie plikt att x utför H mot y.

Den lösning som går ut på att tolka den gyllene regeln som en *vid* och inte en *snäv* plikt tycks kunna användas för att undvika alla argument i den här uppsatsen. (□SVGR) är en mycket intressant tolkning av den gyllene regeln som också innehåller en samma situation-klausul.

Andra möjliga kombinationer kan också vara intressanta.

Vilken tolkning av den gyllene regeln är den bästa? Jag kommer inte att försöka besvara denna fråga i den här uppsatsen. Det finns en mängd andra argument för och emot olika versioner av den gyllene regeln som vi inte har diskuterat i den här uppsatsen. Personligen är jag benägen att tro att det finns flera olika tolkningar av den gyllene regeln som är rimliga.

Referenser

Blackstone, W. T. (1965). The Golden Rule: A Defense. *Southern Journal of Philosophy*, ss. 172–177.

Bruton, S. V. (2004). Teaching the Golden Rule. *Journal of Business Ethics*, Vol. 49, No. 2, ss. 179–187.

Cadoux, A. T. (1912). The Implications of the Golden Rule. *International Journal of Ethics*, Vol. 22, Nr. 3, ss. 272–287.

Carson, T. L. (2010). *Lying and Deception: Theory and Practice*. Oxford: Oxford University Press.

Carson, T. L. (2013). Golden Rule. i Hugh LaFollette (red.) *The International Encyclopedia of Ethics*, ss. 2186–2192.

Duxbury, N. (2009). Golden Rule Reasoning, Moral Judgement and Law. *Notre Dame Law Review* 84, ss. 1529–1605.

Gensler, H. J. (1986). Ethics is Based on Rationality. *The Journal of Value Inquiry* 20, ss. 251–264.

Gensler, H. J. (1996). *Formal Ethics*. London and New York: Routledge.

Gensler, H. J. (2013). *Ethics and the Golden Rule*. New York and London: Routledge.

Gewirth, A. (1978). The golden rule rationalized. *Midwest Studies in Philosophy*, I11, ss. 133–147.

Gould, J. A. (1980). Blackstone's Meta-Not-So-Golden-Rule. *The Southern Journal of Philosophy*, Vol. 18, Issue 4, ss. 509–513.

Hare, R. M. (1963). *Freedom and Reason*. Oxford: Oxford University Press.

Hertzler, J. O. (1934). On Golden Rules. *International Journal of Ethics*, Vol. 44, Nr. 4, ss. 418–436.

Hirst, E. W. (1934). The Categorical Imperative and the Golden Rule. *Philosophy*, Vol. 9, Nr. 35, ss. 328–335.

Hobbes, T. (1985). *Leviathan*. Penguin Books. (red. C. B. Macpherson). (Ursprungligen publicerad 1651.)

Hoche, H.-U. (1978). Die Goldene Regel. Neue Aspekte eines alten Moralprinzips. *Zeitschrift für philosophische Forschung*, Bd. 32, H. 3, ss. 355–375.

Huang, Y. (2005). A Copper Rule versus the Golden Rule: A Daoist-Confucian Proposal for Global Ethics. *Philosophy East and West*, Vol. 55, Nr. 3, ss. 394–425.

Lewis, D. (1973). *Counterfactuals*. Oxford: Basil Blackwell.

Locke, D. (1981). The Principle of Equal Interests. *The Philosophical Review*, Vol. 90, Nr. 4, ss. 531–559.

Mill, J. S. (1987). *Utilitarianism*. Buffalo. New York: Prometheus Books. (Ursprungligen publicerad 1863).

Neusner, J. och Chilton, B. (red.) (2008). *The Golden Rule: The Ethics of Reciprocity in the World Religions*. Continuum.

Pufendorf, S. (1964). *On the Law of Nature and Nations*. New York: Wildy and Sons. (Ursprungligen publicerad 1672).

Reinikainen, J. (2005). The Golden Rule and the Requirement of Universalizability. *The Journal of Value Inquiry* 39, ss. 155–168.

Rönnedal, D. (2012). *Extensions of Deontic Logic: An Investigation into some Multi-Modal Systems*. Department of Philosophy, Stockholm University.

Rönnedal, D. (2015). The Golden Rule and The Platinum Rule. *The Journal of Value Inquiry*, Volume 49, Issue 1, ss. 221–236.

Rönnedal, D. (2016). Den Gyllene Regeln och Substitutionsfunktioner. *Filosofiska Notiser*, Årgång 3, Nr 2, Augusti, ss. 53–80.

Singer, M. G. (1963). The Golden Rule. *Philosophy*, Vol. 38, Nr. 146, ss. 293–314.

Stalnaker, R. C. (1968). A Theory of Conditionals. I N. Rescher (red.), *Studies in Logical Theory*. Oxford: Blackwell, 1968.

Wattles, J. (1996). *The Golden Rule*. New York, Oxford: Oxford University Press.

Weiss, P. (1941). The Golden Rule. *The Journal of Philosophy*, Vol. 38, Nr. 16, ss. 421–430.

Whately, R. (1856). *Introductory Lessons on Morals*. Cambridge: John Bartlett.

Daniel Rönnedal
Filosofiska institutionen
Stockholms universitet
daniel.ronnedal@philosophy.su.se

Den Gyllene Regeln och Egoismen

Daniel Rönnedal

Abstrakt
Enligt den gyllene regeln bör vi behandla andra så som vi själva vill bli behandlade. Det här är en mycket gammal moralisk princip som är känd sedan tusentals år och som i någon form tycks vara en del av alla världsreligioner. Många filosofer från olika kulturer, traditioner och historiska perioder har också accepterat denna princip. I den här uppsatsen argumenterar jag för att det i normala fall ligger i vårt eget intresse att följa den gyllene regeln, att det finns goda egoistiska skäl att leva i enlighet med denna norm. Mitt argument inkluderar två essentiella premisser: att andra tenderar att behandla oss så som vi behandlar dem, och att det ligger i vårt eget intresse att vi blir behandlade så som vi vill bli behandlade. Slutsatsen är att även om det inte alltid är fallet att andra behandlar oss så som vi behandlar dem och det inte alltid ligger i vårt eget långsiktiga intresse att vi blir behandlade så som vi vill bli behandlade, så är premisserna sanna i normala fall.

1. Introduktion

Enligt den s.k. gyllene regeln bör vi behandla andra så som vi själva vill bli behandlade. I den här uppsatsen argumenterar jag för att vi har goda egoistiska skäl att följa denna regel i normala fall. Jag försöker också bemöta två potentiella invändningar mot mitt argument.

Det råder delade meningar om exakt hur den gyllene regeln bäst uttrycks och hur den bör tolkas eller preciseras.[1] Här följer några andra mer eller mindre synonyma formuleringar: Du bör behandla andra så som du själv vill bli behandlad. Allt du vill att andra gör för dig bör du också göra för dem. Om du vill att någon, x, utför en viss handling, H, mot dig, så bör du utföra H mot x. Ibland formuleras regeln som ett imperativ: Behandla andra så som du

[1] Rönnedal (2015) tar upp ett antal frågor som är relevanta då man försöker förstå den gyllene regeln. Se också Rönnedal (2016a) och (2016b).

själv vill bli behandlad! Gör mot andra det du vill att de skall göra mot dig! När jag talar om den gyllene regeln i den här uppsatsen skall jag utgå ifrån följande formulering.

(GR): Du bör behandla andra så som du själv vill bli behandlad.[2]

Denna regel eller princip medför t.ex. följande satser. Om du vill att andra skall behandla dig med vänlighet och respekt, så bör du behandla dem med vänlighet och respekt. Om du vill att andra skall behandla dig rättvist, så bör du behandla dem rättvist. Om du vill att andra skall hålla sina löften till dig, så bör du hålla dina löften till dem. Om du vill att andra skall hjälpa dig när du behöver hjälp, så bör du hjälpa dem när de behöver hjälp.

Ibland har man gjort en distinktion mellan den gyllene regeln och den s.k. silverregeln. Enligt silverregeln bör vi undvika att behandla andra på sätt som vi själva inte vill bli behandlade. Också denna regel har olika möjliga tolkningar. Här följer några mer eller mindre synonyma varianter. Du bör inte behandla andra på sätt som du själv inte vill bli behandlad. Ingenting av det du vill att andra skall undvika att göra mot dig bör du göra mot dem. Om du vill att någon, x, inte utför en viss handling, H, mot dig, så bör du inte utföra H mot x. Denna regel medför t.ex. följande satser. Om du vill att andra inte ljuger för dig, så bör du inte ljuga för dem. Om du vill att andra inte stjäl från dig, så bör du inte stjäla från dem. Om du vill att andra inte misshandlar dig, så bör du inte misshandla dem.

Det råder delade meningar om hur den gyllene regeln och silverregeln förhåller sig till varandra, om de är logiskt oberoende av varandra eller inte och om vilken regel som är primär eller viktigast.[3] Jag antar i den här uppsatsen att silverregeln följer ur den gyllene regeln, även om jag inte tror att detta antagande är nödvändigt för uppsatsens huvudargument.

[2] Om vi antar att (GR) alltid är sann för alla personer i alla situationer, kan vi härleda vissa kontraintuitiva slutsatser. Se t.ex. Gensler (1996) (särskilt Kapitel 5) och Gensler (2013) (särskilt Kapitel 14) för en genomgång av några svårigheter. För att undvika vissa problematiska konklusioner måste vi i praktiken ofta ta hänsyn till att olika personer kan befinna sig i olika situationer och att olika individer kan ha olika preferenser och önskningar eller personliga egenskaper när vi tillämpar regeln. Men även (GR) förefaller vara rimlig i normala fall. Och formuleringen är tillräckligt precis för våra syften i den här uppsatsen.

[3] Se Gensler (2013), Kapitel 10, för mer om detta.

När jag säger att någon handlar i enlighet med den gyllene regeln (GR) menar jag att hon faktiskt behandlar andra så som hon själv vill bli behandlad.[4]

Alla världsreligioner, och många andra religioner, tycks innehålla någon variant av den gyllene regeln.[5] Även många filosofer har accepterat den i en eller annan form; låt mig nämna fyra exempel på anhängare som tillhör olika moralfilosofiska traditioner och tidsepoker: Thomas Hobbes, Samuel Pufendorf, John Stuart Mill och Harry Gensler. Hobbes anser att moralen är baserad på ett slags socialt kontrakt som människor, för att gynna sina intressen och undvika ett krigstillstånd, ingår. Den gyllene regeln är summan av de moraliska Naturlagarna. I samband med några anmärkningar Hobbes gör om en av dessa Naturlagar säger han t.ex.: ”This is that Law of the Gospell; *Whatsoever you require that others should do to you, that do ye to them*.”[6] Pufendorf, som tillhör en rättighetsetisk tradition, menar att vi kan få kunskap om den gyllene regeln genom förnuftet, och att den är en naturrättslig princip.[7] Så här skriver Mill, som är en s.k. utilitarist som anser att handlingar är riktiga i den mån de tenderar att leda till lycka och fel i den mån de tenderar att producera motsatsen till lycka: ”In the golden rule…, we read the complete spirit of the ethics of utility. To do as you would be done by, and to love your neighbor as yourself, constitute the ideal perfection of utilitarian morality.”[8] Gensler, som betraktar sin egen formella etik som en form av kantianism, menar att den gyllene regeln är ett slags konsistens-princip som kan härledas från vissa axiom som kräver att vi är opartiska och samvetsgranna (ärliga i våra moraliska omdömen). Så här säger Gensler om den gyllene regeln:

[4] Om antagandet att silverregeln följer ur den gyllene regeln är korrekt, så följer det att någon som handlar i enlighet med (GR) också undviker att behandla andra på sätt hon själv inte vill bli behandlad.
[5] Se Neusner och Chilton (red.) (2008).
[6] Se Kapitel XIV, s. 190 i Hobbes (1985).
[7] Pufendorf (1964), bok 2, 3:13.
[8] Mill (1987), Kapitel 2, s. 28.

> The golden rule (GR – "Treat others as you want to be treated") is the most important principle of formal ethics, the central jewel of the theory. GR is important in all the great world religions; Jesus, Hillel, and Confucius used it to summarize their teachings. And GR is influential among conscientious people in our own time.[9]

> The golden rule requires that we treat others as we want to be treated. GR is the most important principle of formal ethics – and perhaps the most important rule of life.[10]

Många intressanta frågor aktualiseras då man funderar på den gyllene regeln, t.ex. Hur skall principen förstås? Vilken tolkning av den är den bästa? Vad har den för logisk form? Vad finns det för argument för och emot denna regel? Är den gyllene regeln en sann, berättigad eller förnuftig regel? Hur förhåller den sig till olika moralfilosofiska teorier och metaetiska ståndpunkter? Vilka konsekvenser har principen för hur vi bör leva och förhålla oss till varandra?

I den här uppsatsen skall jag emellertid inte säga någonting om dessa frågor, utan jag skall koncentrera mig på en annan fråga, nämligen: Vad finns det för *skäl* att handla i enlighet med den gyllene regeln? Och närmare bestämt skall jag fokusera på om det ligger i vårt eget intresse, om det kan finnas egoistiska skäl att följa denna princip.[11]

2. Varför skall vi handla i enlighet med den gyllene regeln? Det egoistiska argumentet

Det kan finnas många olika skäl att handla i enlighet med den gyllene regeln: egoistiska, altruistiska, moralfilosofiska, religiösa m.m. En altruistisk person som bryr sig om andra och vill hjälpa dem och tror att vi hjälper andra om vi handlar i enlighet med (GR), kommer att ha ett subjektivt *altruistiskt skäl* att handla i enlighet med (GR). En person som tror att det är moraliskt riktigt att handla i enlighet med (GR) och vill handla moraliskt riktigt, kommer att ha

[9] Gensler (1996), ss. 12–13.
[10] Gensler (1996), s. 93.
[11] För mer historisk information om den gyllene regeln, se t.ex. Wattles (1996) och Gensler (2013), Kapitel 5. För en diskussion om den gyllene regelns förhållande till olika religioner, se Neusner och Chilton (red.) (2008). Filosofiska introduktioner till den gyllene regeln finner man bl.a. i Carson (2010), Kapitel 6, Carson (2013), Gensler (1996), särskilt Kapitel 5, och Gensler (2013). För mer information om den gyllene regeln se t.ex. Blackstone (1965), Bruton (2004), Cadoux (1912), Duxbury (2009), Gensler (1986), (2013), Gewirth (1978), Gould (1980), Hare (1963), Hertzler (1934), Hirst (1934), Hoche (1978), Huang (2005), Reinikainen (2005), Rönnedal (2015), Singer (1963), Wattles (1996) och Weiss (1941).

ett subjektivt *moraliskt skäl* att handla i enlighet med (GR). Och en religiös person som tror att det är Guds vilja att vi handlar i enlighet med (GR) och vill utföra Guds vilja, kommer att ha ett subjektivt *religiöst skäl* att handla i enlighet med (GR).[12] Jag skall emellertid bortse ifrån dessa möjliga skäl i denna uppsats. Det som intresserar mig här är om det finns några goda *egoistiska skäl* att handla i enlighet med (GR). Ligger det i vårt eget intresse att handla i enlighet med den gyllene regeln? Jag skall nu undersöka ett argument som talar för att svaret på denna fråga är ja. Låt oss kalla detta argument för ”det egoistiska argumentet”.

1. Om du handlar i enlighet med den gyllene regeln, så behandlar du andra så som du själv vill bli behandlad.
2. Andra behandlar dig så som du behandlar dem.
3. Om du behandlar andra så som du själv vill bli behandlad och andra behandlar dig så som du behandlar dem, så kommer andra att behandla dig så som du själv vill bli behandlad.

Alltså.

4. Om du handlar i enlighet med den gyllene regeln, så kommer andra att behandla dig så som du själv vill bli behandlad. [Från 1–3, Satslogik]
5. Om det är sant att om du handlar i enlighet med den gyllene regeln så kommer andra att behandla dig så som du själv vill bli behandlad, så ligger det i ditt eget intresse att handla i enlighet med den gyllene regeln (du har goda egoistiska skäl att handla i enlighet med denna princip).

Alltså.

6. Det ligger i ditt eget intresse att handla i enlighet med den gyllene regeln (du har goda egoistiska skäl att handla i enlighet med denna princip). [Från 4 och 5, Satslogik]

Det som gäller för dig gäller också för alla andra, vi har inte antagit att du har några unika egenskaper. Alltså kan vi sluta oss till att det för varje person ligger i hennes eget intresse att följa den gyllene regeln, att varje person har goda egoistiska skäl att handla i enlighet med denna princip.

Detta argument är logiskt giltigt, slutsatsen följer med nödvändighet ur premisserna. Slutsatsen kan därför endast vara falsk om någon av premisserna är falsk. Sats 4 följer ur 1–3 och sats 6 ur sats 4 och 5 med hjälp

[12] Om dessa *subjektiva* skäl också är *objektiva* beror väl på om det faktiskt är sant att vi hjälper andra om vi handlar i enlighet med (GR), om det faktiskt är moraliskt riktigt att handla i enlighet med (GR) och om Gud finns och faktiskt vill att vi handlar i enlighet med (GR).

av satslogik. Generaliseringen till alla personer följer med hjälp av klassisk predikatlogik. Så om vi vill ifrågasätta slutsatsen måste vi förkasta sats 1, 2, 3 eller 5. Premiss 1 och premiss 3 förefaller vara begreppsligt sanna och därmed i stort sett otvivelaktiga, åtminstone om man utgår ifrån en klassisk formulering av den gyllene regeln, vilket vi gör i denna uppsats. Det kan emellertid finnas skäl att ifrågasätta sats 2 eller sats 5. Låt oss därför se närmare på dessa.

Enligt premiss 2 behandlar andra dig så som du behandlar dem. Om du behandlar andra vänligt, behandlar de dig vänligt. Om du behandlar andra rättvist, behandlar de dig rättvist. Om du däremot behandlar andra ovänligt, behandlar de dig ovänligt osv. Det här är ett empiriskt och psykologiskt antagande som knappast är sant för alla personer i alla situationer vid alla tidpunkter och för alla (typer av) handlingar. Det är lätt att tänka sig fall där satsen är falsk. Vi kan t.ex. föreställa oss samhällen som övervägande består av utpräglade egoister eller utpräglade altruister. I den förra typen av samhälle kan det hända att de flesta kommer att behandla dig illa oavsett hur väl du behandlar dem. Och i den senare typen av samhälle kan det hända att andra kommer att behandla dig väl oavsett hur illa du behandlar dem. I sådana samhällen kan det finnas skäl för en egoist att inte följa den gyllene regeln, åtminstone inte alltid. Det är emellertid tvivelaktigt om det *finns* eller någonsin *har funnits* några sådana samhällen. Det går dock att hitta mer realistiska fall då det kanske inte är egoistiskt lönsamt att följa den gyllene regeln, t.ex. i vissa situationer i krig eller i vissa situationer då man interagerar med någon empatilös eller egoistisk person. Även om du t.ex. inte skadar andra i sådana situationer, är det inte säkert att de inte kommer att skada dig. Premiss 2 kan också vara falsk i situationer då individer med olika personliga egenskaper – inklusive önskningar och preferenser – samspelar, eller då personer som befinner sig i olika situationer möts. Låt oss betrakta ett trivialt exempel som involverar en ganska specifik handling. Om du t.ex. gillar att överraskas och du ordnar en överraskningsfest till din vän när hon fyller år, men din vän inte gillar att bli överraskad, så är det inte säkert att hon ordnar en överraskningsfest till dig när du fyller år. Därför kan det egoistiska argumentet inte användas för alla personer i alla situationer vid alla tidpunkter och för alla typer av handlingar. Men jag tror att premissen är riktig i normala fall för de flesta personer i relativt välfungerande samhällen. Om du beter dig vänligt mot andra, så är t.ex. sannolikheten stor att de kommer att bete sig vänligt mot dig. Om du beter dig ovänligt mot andra, är emellertid sannolikheten stor att de beter sig ovänligt mot dig. Om du stjäl

från någon annan är det inte säkert att denna person kommer att stjäla från dig. Men det är troligt att han eller hon kommer att bete sig på ett sätt som inte gynnar dina egna intressen, t.ex. genom att anmäla dig till polisen. Poängen med det egoistiska argumentet är inte att alla i alla situationer vid alla tidpunkter har goda egoistiska skäl att handla i enlighet med den gyllene regeln. Slutsatsen är sann i de fall premissen är sann (givet att alla andra premisser också är sanna). Och jag tror att den i regel är sann för personer i normala fall som lever i relativt välordnade samhällen.

Det här argumentet för premiss 2 bygger på mina egna observationer av hur människor tenderar att bete sig mot varandra. Den induktiva basen är därför begränsad. Man kan fråga sig om det finns andra skäl att tro på denna premiss. Låt mig nämna fyra överväganden.

För det första, jag tycks inte vara den enda som har gjort observationen att andra tenderar att behandla dig så som du behandlar dem, otaliga andra vittnar om samma sak. Denna föreställning tycks också vara en del av mänsklighetens samlade visdom, förkroppsligad i en mängd ordspråk, t.ex. ”Som man sår får man skörda”, ”Som man bäddar får man ligga”, ”Den som gräver en grop åt andra…”, ”what goes around comes around” osv.

För det andra finns det vissa mer teoretiska resonemang som stödjer premiss 2. Robert Axelrod har visat att beslutsprincipen *tit for tat* är framgångsrik i många olika miljöer.[13] Detta talar för att livsformer som följer en sådan princip har goda chanser att överleva. Det är därför mycket möjligt att vi pga. våra gener tenderar att använda eller handla i enlighet med en princip av detta slag, det förefaller vara naturligt för oss att behandla andra så som vi blir behandlade av dem. Det tycks t.ex. vara en spontan reaktion hos människor att de ofta känner hämndbegär om de har utsatts för någon oförrätt och att de känner tacksamhet om de har fått hjälp, begär och känslor som i sin tur kan leda till handling. Det är rimligt att anta att en medfödd benägenhet att hjälpa de som hjälper dig kan ha haft ett överlevnadsvärde i tidiga mänskliga sammanslutningar. I samhällen där ”retributionsprincipen” (”behandla andra så som de behandlar dig”) är en accepterade norm, är det sannolikt att andra kommer att behandla dig så som du behandlar dem av

[13] Se Axelrod (1984). *Tit for tat* är en beslutsmetod som innebär att vi börjar med att samarbeta med nya individer vi möter och därefter behandlar andra så som de behandlade oss vid det senaste mötet. Det vill säga, första gången du träffar x, så väljer du att samarbeta med x. Om x samarbetade med dig förra gången ni möttes, så samarbetar du med x denna gång; och om x inte samarbetade med dig förra gången ni möttes, så samarbetar du inte med x denna gång. Enkelt uttryckt innebär detta att en person som följer *tit for tat* kommer att behandla dig så som du behandlar henne.

kulturella skäl. I ett sammanhang där många använder *tit for tat* är den gyllene regeln en framgångsrik princip. Det kan alltså finnas både genetiska och sociologiska förklaringar till att andra tenderar att behandla dig så som du behandlar dem. Notera att premiss 2 är ett *deskriptivt* påstående om hur människor *faktiskt* tenderar att bete sig, inte ett *normativt* påstående om hur de *bör* handla.

För det tredje har det på senare tid börjat dyka upp en del empirisk forskning som talar för att premiss 2 är riktig i många fall.[14]

För det fjärde kan vi formulera ett historiskt-abduktivt argument för premiss 2 på följande sätt. (1) Den gyllene regeln är en populär norm som har en stor spridning. Det här är ett empiriskt påstående som det finns mycket goda skäl att hålla för sant. Som vi såg i introduktionen förekommer den gyllene regeln i någon form i alla världsreligioner och många filosofer har också accepterat den. Det är vanligt att föräldrar hänvisar till denna princip då de försöker uppfostra sina barn. (2) Förklaringen till att den gyllene regeln är en populär norm som har en stor spridning är att grupper där individerna handlar i enlighet med denna norm är framgångsrika. (3) (Den historiska) förklaringen till att grupper där individerna handlar i enlighet med den gyllene regeln är framgångsrika är att människor tenderar att behandla andra så som de själva blir behandlade. (4) Alltså har vi goda (abduktiva skäl) att tro att människor tenderar att behandla andra så som de själva blir behandlade. Den gyllene regeln tycks först ofta ha formulerats av personer som lever i samhällen där någon form av retributionsprincip är allmänt accepterad.[15] I sådana samhällen är den gyllene regeln i normala fall en framgångsrik strategi. I sociala sammanhang där många behandlar andra så som de själva blir behandlade, har grupper där individerna handlar i enlighet med den gyllene regeln goda möjligheter att frodas. Och det är precis det vi tycks se då vi betraktar historiska fakta. Det är oklart om den gyllene regeln hade lycktas få en sådan spridning om den inte hade varit framgångsrik.[16]

[14] Se t.ex. Vogel (2004) för en sammanfattning av delar av denna forskning.

[15] Dihle (1962).

[16] Premisserna i detta argument kan diskuteras. En annan möjlig förklaring till den gyllene regelns spridning är helt enkelt att principen är sann, berättigad, förnuftig eller rationell och att många människor har insett det. Men om den gyllene regeln är sann, berättigad, förnuftig eller rationell, så har vi goda skäl att tro att vi ”bör” handla i enlighet med denna regel. Om den gyllene regeln tolkas som en klokhetsprincip, så följer det omedelbart att vi har goda skäl att tro att det ligger i vårt eget intresse att handla i enlighet med denna maxim. Om den gyllene regeln tolkas som en moralisk princip, så följer inte detta utan övriga premisser. Det är inte säkert att det alltid ligger i vårt eget intresse att handla moraliskt. Men om det i normala fall ligger i vårt intresse att handla moraliskt och den gyllene regeln är en moralisk princip, så kan vi också sluta

Dessa skäl är inte konklusiva. Det ligger i sakens natur. Eftersom premiss 2 är en empirisk premiss kan vi knappast bevisa att den är sann. Det är inte nödvändigt sant att andra behandlar dig så som du behandlar dem. Men de ger ändå ett visst, inte obetydligt, stöd för påståendet att (många) andra *i normala fall* behandlar dig så som du behandlar dem.

Någon kanske skulle vilja ifrågasätta premiss 5. Kanske ligger det inte alltid i vårt eget intresse att vi blir behandlade så som vi vill bli behandlade. Detta skulle t.ex. kunna vara fallet om vi inte vet vårt eget bästa eller om våra önskningar är irrationella eller självdestruktiva. Kanske ligger det inte heller alltid i vårt eget *långsiktiga* intresse att vi blir behandlade så som vi vill bli behandlade. Våra önskningar tycks inte sällan vara kortsiktiga och jakten på omedelbar behovstillfredsställelse kan leda till problem på sikt. Det är möjligt att det kan förhålla sig på det sättet och troligt att inte alla alltid vet vad som är bäst för dem själva i det långa loppet. Sats 5 tycks därför inte vara nödvändigt sann. Men även om det kanske inte är begreppsligt sant att det ligger i vårt eget intresse att vi blir behandlade så som vi vill bli behandlade, åtminstone inte i vårt långsiktiga intresse, så förefaller det vara sant åtminstone i de flesta fall. Och därför tror jag att premissen åtminstone i de

---

oss till att det i normala fall ligger i vårt intresse att handla i enlighet med den gyllene regeln. Så, om den bästa förklaringen till den gyllene regelns spridning är att principen är sann, berättigad, förnuftig eller rationell och att många människor har insett det, så tycks vi kunna formulera ett antal alternativa argument för att det ligger i vårt intresse att handla i enlighet med denna princip.

En annan möjlig förklaring till den gyllene regelns spridning är att någon variant av denna princip tycks ingå i de flesta stora religioner. Regeln har spridits tillsammans med och som en del av dessa religioner. Om detta är riktigt, så kan vi inte omedelbart sluta oss till att grupper som lever i enlighet med den gyllene regeln är framgångsrika. Men man kan förstås fråga sig varför de olika världsreligionerna har fått en så stor spridning. En möjlig förklaring (eller delförklaring) är att grupper som lever i enlighet med de normer (eller åtminstone de ”viktigaste” normer) religionerna förespråkar är framgångsrika. Bland dessa normer ingår den gyllene regeln. I den utsträckning de olika anhängarna av de olika religionerna lever i enlighet med den gyllene regeln, har de goda chanser att vara framgångsrika. Om detta är riktigt, tycks vi kunna formulera ett alternativt argument för våra utgångspunkter.

Poängen med premiss 4 är inte att den gyllene regeln endast är framgångsrik i miljöer där många tenderar att behandla andra så som de själva blir behandlade. Men det tycks vara ett *historiskt faktum* att principen först explicit formuleras i samhällen där retributionsprincipen är populär och där många behandlar andra så som de själva blir behandlade. Och i sådana miljöer har grupper som följer den gyllene regeln goda chanser att frodas. I en strikt mening tycks det abduktiva argumentet endast ge stöd åt uppfattningen att det *har varit* fallet att många tenderar att behandla andra så som de själva blir behandlade. Det egoistiska argumentet bygger på premissen att det fortfarande är sant att många tenderar att behandla dig så som du behandlar dem. Men det verkar vara rimligt att tro att människan inte har förändrats så mycket sedan den tid då den gyllene regeln först explicit formulerades.

Även om premisserna i det abduktiva argumentet kan diskuteras, så tycks de ha en viss plausibilitet.

flesta fall kommer att vara sann. Notera också att premissen varken hävdar att det är *bra* att vi blir behandlade så som vi vill bli behandlade eller att vi *bör* bli behandlade så som vi vill bli behandlade. Dessa teser förefaller också vara korrekta i normala fall, men vi behöver inte anta detta för att det egoistiska argumentet skall gå igenom.

Sammanfattningsvis tror jag att argumentet är sunt för de flesta personer i normala situationer i relativt välfungerande samhällen. Domänen i den generaliserade slutsatsen måste alltså begränsas. Gör man detta och förstår argumentet med de undantag som nämnts ovan, talar det emellertid för att vi har goda egoistiska skäl att handla i enlighet med den gyllene regeln. Om detta är sant, tror jag också att vi har goda egoistiska skäl att internalisera den gyllene regeln och att det ligger i vårt eget intresse att använda denna princip som en tumregel då vi fattar olika beslut. Jag tror, med andra ord, att vi har goda egoistiska skäl att försöka göra den gyllene regeln till en del av vår natur, så att vi automatiskt handlar i enlighet med den i normala fall, även om det kan finnas situationer då det inte ligger i vårt eget intresse.

3. Slutsats

Jag har i den här uppsatsen argumenterat för att vi har goda egoistiska skäl att handla i enlighet med den gyllene regeln, principen att vi bör behandla andra så som vi själva vill bli behandlade, åtminstone i normala fall. Om det egoistiska argumentet är sunt, ligger det i vårt eget intresse att följa denna princip. Troligtvis finns det situationer i de flesta personers liv då de inte *själva* tjänar på att följa den gyllene regeln. Men i relativt välfungerande samhällen kommer sådana situationer sannolikt att utgöra undantag. Vidare tycks det följa att vi har goda egoistiska skäl att internalisera den gyllene regeln och att det ligger i vårt eget intresse att använda denna princip som en tumregel då vi fattar olika beslut. Ingenting av det jag har sagt i den här uppsatsen utesluter att det *också* kan finnas *andra* skäl att handla i enlighet med den gyllene regeln, t.ex. altruistiska, moralfilosofiska eller religiösa. Personligen tror jag att det finns sådana och att många som handlar i enlighet med denna princip gör det av icke-egoistiska anledningar.[17] Men det är också

[17] I Rönnedal (201X) argumenterar jag t.ex. för att vi i normala fall inte endast har goda *egoistiska* skäl att handla i enlighet med den gyllene regeln, utan också goda *altruistiska* skäl. I denna uppsats utgår jag ifrån ett antal spelteoretiska exempel. Argumenten i Rönnedal (201X) bygger därför på andra premisser än de premisser jag använder i den här uppsatsen, men slutsatsen pekar i samma riktning.

intressant att notera att t.ex. även en ateistisk, nihilistisk superegoist kan ha goda skäl att handla i enlighet med den gyllene regeln.[18]

Appendix

Man kan kanske tycka att slutsatsen i den här uppsatsen är trivial. Poängen är dock att vi har producerat ett deduktivt giltigt argument, med premisser som förefaller vara sanna i normala fall, som visar att det *måste* vara fallet att slutsatsen är sann om premisserna är sanna. Och att det finns ett sådant argument är inte alls trivialt.

Detta appendix innehåller en kvasiformalisering av det egoistiska argumentet. H står för ”Du handlar i enlighet med GR”; V_dA för ”Du vill att A”, där A kan bytas ut mot vilken sats som helst; Hdx står för ”Du utför handling H mot x”, där ”x” betecknar en godtycklig individ; Hxd står för ”x utför handling H mot dig”; I för ”Det ligger i ditt eget intresse att handla i enlighet med GR”.

Formalisering av det egoistiska argumentet

1. $H \rightarrow (V_dHxd \rightarrow Hdx)$
2. $Hdx \rightarrow Hxd$
3. $((V_dHxd \rightarrow Hdx) \wedge (Hdx \rightarrow Hxd)) \rightarrow (V_dHxd \rightarrow Hxd)$
4. $H \rightarrow (V_dHxd \rightarrow Hxd)$ [Från 1 och 3]
5. $(H \rightarrow (V_dHxd \rightarrow Hxd)) \rightarrow I$
6. I [Från 4 och 5]

Referenser

Axelrod, R. (1984). *The Evolution of Co-operation*. Penguin Books.

Blackstone, W. T. (1965). The Golden Rule: A Defense. *Southern Journal of Philosophy*, ss. 172–177.

[18] Med en ateist menar jag i detta sammanhang en person som inte tror på någon gud eller på någon religion och därför inte anser sig ha några religiösa skäl att följa den gyllene regeln. Med en nihilist menar jag en person som inte tror att det finns några moraliska sanningar och därför inte anser sig ha några moraliska skäl att följa den gyllene regeln. Med en superegoist menar jag en person som *endast* bryr sig om sig själv och därför inte anser sig ha några altruistiska skäl att handla i enlighet med den gyllene regeln. En ateistisk, nihilistisk, superegoist är därför en person som inte anser sig ha några religiösa, moraliska eller altruistiska skäl att följa den gyllene regeln. Även en sådan person kan alltså ha goda skäl att handla i enlighet med denna princip, dvs. att behandla andra så som hon själv vill bli behandlad.

Bruton, S. V. (2004). Teaching the Golden Rule. *Journal of Business Ethics*, Vol. 49, Nr. 2, ss. 179–187.

Cadoux, A. T. (1912). The Implications of the Golden Rule. *International Journal of Ethics*, Vol. 22, Nr. 3, ss. 272–287.

Carson, T. L. (2010). *Lying and Deception: Theory and Practice*. Oxford: Oxford University Press.

Carson, T. L. (2013). Golden Rule. I Hugh LaFollette (red.) *The International Encyclopedia of Ethics*, ss. 2186–2192.

Dihle, A. (1962). *Die Goldene Regel*. Göttingen: Vandenhoeck und Ruprecht.

Duxbury, N. (2009). Golden Rule Reasoning, Moral Judgement and Law. *Notre Dame Law Review* 84, ss. 1529–1605.

Gensler, H. J. (1986). Ethics is Based on Rationality. *The Journal of Value Inquiry* 20, ss. 251–264.

Gensler, H. J. (1996). *Formal Ethics*. London and New York: Routledge.

Gensler, H. J. (2013). *Ethics and the Golden Rule*. New York and London: Routledge.

Gewirth, A. (1978). The golden rule rationalized. *Midwest Studies in Philosophy*, I11, ss. 133–147.

Gould, J. A. (1980). Blackstone's Meta-Not-So-Golden-Rule. *The Southern Journal of Philosophy*, Vol. 18, Issue 4, ss. 509–513.

Hare, R. M. (1963). *Freedom and Reason*. Oxford: Oxford University Press.

Hertzler, J. O. (1934). On Golden Rules. *International Journal of Ethics*, Vol. 44, Nr. 4, ss. 418–436.

Hirst, E. W. (1934). The Categorical Imperative and the Golden Rule. *Philosophy*, Vol. 9, Nr. 35, ss. 328–335.

Hoche, H.-U. (1978). Die Goldene Regel. Neue Aspekte eines alten Moralprinzips. *Zeitschrift für philosophische Forschung*, Bd. 32, H. 3, ss. 355–375.

Hobbes, T. (1985). *Leviathan*. Penguin Books. (red. C. B. Macpherson). (Ursprungligen publicerad 1651.)

Huang, Y. (2005). A Copper Rule versus the Golden Rule: A Daoist-Confucian Proposal for Global Ethics. *Philosophy East and West*, Vol. 55, Nr. 3, ss. 394–425.

Mill, J. S. (1987). *Utilitarianism*. Buffalo, New York: Prometheus Books. (Ursprungligen publicerad 1863).

Neusner, J. och Chilton, B. (red.) (2008). *The Golden Rule: The Ethics of Reciprocity in the World Religions*. Continuum.

Pufendorf, S. (1964). *On the Law of Nature and Nations*. New York: Wildy and Sons. (Ursprungligen publicerad 1672).

Reinikainen, J. (2005). The Golden Rule and the Requirement of Universalizability. *The Journal of Value Inquiry* 39, ss. 155–168.

Rönnedal, D. (2015). The Golden Rule and The Platinum Rule. *The Journal of Value Inquiry*, Volume 49, Issue 1, ss. 221–236.

Rönnedal, D. (2016a). Den Gyllene Regeln och Substitutionsfunktioner. *Filosofiska Notiser*, Årgång 3, Nr 2, Augusti, ss. 53–80.

Rönnedal, D. (2016b). Den Gyllene Regeln och Intra- och Interpersonella Viljekonflikter. *Filosofiska Notiser*, Årgång 3, Nr 2, Augusti, ss. 81–106.

Rönnedal, D. (201X). Den Gyllene Regeln och Farmarnas Dilemma. *Tidskrift för Politisk Filosofi*. Antagen.

Singer, M. G. (1963). The Golden Rule. *Philosophy*, Vol. 38, Nr. 146, ss. 293–314.

Vogel, G. (2004). The Evolution of the Golden Rule. *Science, New Series*, Vol. 303, Nr. 5661, ss. 1128–1129+1131.

Wattles, J. (1996). *The Golden Rule*. New York, Oxford: Oxford University Press.

Weiss, P. (1941). The Golden Rule. *The Journal of Philosophy*, Vol. 38, Nr. 16, ss. 421–430.

Daniel Rönnedal
Filosofiska institutionen
Stockholms universitet
daniel.ronnedal@philosophy.su.se