Human Sphexishness

Why We Make the Same Mistakes Over and Over Again

Christopher von Bülow*

Abstract

Why is it that sometimes we do the same inappropriate things over and over again, i.e., why are we sphexish? It's because we have adopted detrimental, but self-reinforcing, routines and beliefs – memes – which we aren't conscious of any more or have a strong urge to adhere to.

1 Introduction

1.1 The basic idea: Self-reinforcing systems of routines and beliefs

Everybody acts irrationally sometimes. If, however, someone makes the same mistake over and over again, that is a very blatant kind of irrationality indeed. I will call this kind of behavior *sphexishness*, following Hofstadter (1986, p. 529).

I propose a general explanation for this sort of irrational behavior. In a nutshell: if Otto makes the same mistake again and again it is because, at some point in the past, he has learned things – has acquired beliefs and routines – which since then, though detrimental to him, have affected his behavior in such a way that he didn't unlearn them again.

In a little more detail: the beliefs and routines (*memes*, for short) directly responsible for Otto's unprofitable behavior form part of a system of memes Otto has, some of which he isn't fully aware of and some of which he is more or less compelled to adhere to, that disposes him to interpret, or seek out, or shun, certain situations in such a way that the memes in the system are reinforced.

Two examples to illuminate this abstract description. The first is taken from Watzlawick 1976 (p. 59) and is, though about horses, instructive with regard to people as well. A horse is standing in its stable, and there is an electrode in the floor below one of its hooves. By repeatedly giving an electric shock to the horse shortly after ringing a bell, the horse is conditioned to lift its hoof at the sound of the bell. Thus it acquires, as it were, the belief, 'every time the bell rings I get hurt if I don't lift my hoof'. After a while the procedure is changed: though the bell is still being rung from time to time, there are no more electric jolts. Of course the horse keeps on lifting its hoof, though this now is a wasted effort.

The horse keeps on wasting its energy because the routine it has acquired – 'when the bell rings, lift your hoof!' – prohibits it from noticing that the belief that underlies the routine is wrong: it will *not* get hurt every time the bell rings, even if it doesn't

*eMail: Christopher.von.Buelow@uni-konstanz.de; Homepage: www.uni-konstanz.de/FuF/Philo/Philosophie/Spohn/vonBuelow

lift its hoof. If it didn't lift its hoof it would notice that it doesn't get hurt, but by lifting its hoof it avoids the situation of having its hoof on the floor after the bell has rung and so doesn't notice the absence of an electric shock. Thus, the routine prevents the falsification of the belief which engendered it in the first place. So, the horse harbours two mutually reinforcing memes which are detrimental under the changed circumstances. (In a detailed analysis more memes will be seen to be involved.)

The second example is about a detrimental system of memes that perpetuates itself by inducing the holder to misinterpret certain situations. Imagine that Otto believes himself to be flawed in some way that is recognizable by others and inclines them to dislike him. The supposed defect might consist in, say, an ugly nose, excessive insecurity, or a lower-class provenance. His flaw and its consequences are troubling him very much; they are rarely far from his thoughts, and he is almost constantly on the lookout for signs of rejection. So, whenever he feels to be mistreated without an obvious reason, the explanation that springs to his mind is the one that is nearest to hand: the cause must be his defect.

Otto's belief in the dire consequences of his presumed flaw makes him preoccupied with it (an internal routine), and his preoccupation makes him prone to (mis-) interpretations confirming this belief.[a]

[a] kurze Vorausschau auf den Rest?

1.2 *Is* there a problem?

Is sphexishness a real problem? Would anybody in his or her right mind really repeat a substantial mistake again and again without realizing and putting a stop to it? I believe that sphexishness is not only real but indeed very common. It is not the aim of this paper to convince you of this claim, but I shall at least try to make it somewhat more plausible.

The easy way to obtain probable evidence of sphexishness is to think about people close to or frequently met by us: parents, partners, children, friends, colleagues, subordinates, superiors, and so on. For many of them we will be able to think of some kind of mistake they regularly commit. 'He's always late', 'she always forgets things', 'he always drives too fast', and so forth. If we think about habits or traits which they might be happier or more successful without we will find more examples: 'X shouldn't talk so much', '*Y* shouldn't be so uptight', 'Z shouldn't be so domineering', '*W* shouldn't dress so sloppily'. All these may constitute cases of sphexishness.

The more difficult but also more interesting way of looking for everyday sphexishness is to think about *ourselves*. Is there something in my life that always goes wrong, maybe always in a similar fashion? Something that usually doesn't work out the way I wanted it to? Something where I have the same complaint again and again? (Think about the ways your parents, children, friends, etc., annoy you or make you unhappy; think about the ways you are dissatisfied with yourself...)

I suggest that these are the places where our *own* sphexishness surfaces. Where is the mistake? where is the irrationality?, you may ask. Well, if we *have* the means to improve our situation then why don't we just do it, instead of letting ourselves be frustrated time after time? If, on the other hand, we *don't* have the means then it's no use complaining and we would be better off if we just accommodated. Either way something is fishy. (This matter deserves of more careful analysis, but this is not the place for it.)

Furthermore, compulsions, phobias and addictions may constitute (extreme) cases of sphexishness.[b]

[b] mehr konkrete Beispiele?

- immer wieder zu spät kommen (wenn mit Beschwerde verbunden),
- immer klein beigeben,
- weakness of will, self-deception (wenn wiederholt; d.h. als trait?),
- den falschen Leuten vertrauen und dann enttäuscht werden (?),
- sich zu viel vornehmen,
- abends zu viel essen,
- abends zu lange lesen

(Beschwerde: hinreichend, aber nicht notwendig)

2 Making the same mistake again and again

2.1 The paradigm of sphexishness

Sphexishness owes its name to the digger wasp *Sphex ichneumoneus*, which can be caught in a – to us – glaringly obvious rut without breaking out or otherwise reacting to it in any way.

> When the time comes for egg laying, the wasp *Sphex* builds a burrow for the purpose and seeks out a cricket which she stings in such a way as to paralyze but not kill it. She drags the cricket into the burrow, lays her eggs alongside, closes the burrow, then flies away, never to return. In due course, the eggs hatch and the wasp grubs feed off the paralyzed cricket, which has not decayed, having been kept in the wasp equivalent of a deepfreeze. To the human mind, such an elaborately organized and seemingly purposeful routine conveys a convincing flavor of logic and thoughtfulness—until more details are examined. For example, the wasp's routine is to bring the paralyzed cricket to the burrow, leave it on the threshold, go inside to see that all is well, emerge, and then drag the cricket in. If the cricket is moved a few inches away while the wasp is inside making her preliminary inspection, the wasp, on emerging from the burrow, will bring the cricket back to the threshold, but not inside, and will then repeat the preparatory procedure of entering the burrow to see that everything is all right. If again the cricket is removed a few inches while the wasp is inside, once again she will move the cricket up to the threshold and reenter the burrow for a final check. The wasp never thinks of pulling the cricket straight in. On one occasion this procedure was repeated forty times, always with the same result. (Wooldridge 1968, p. 70)

For Hofstadter (1986, p. 529) this is the paradigm of sphexishness. I will be coming back to this example and later on propose that human sphexishness is different from and more interesting than *Sphex*'s sphexishness.

2.2 Doing it again and again: Dispositions (instead of loops)

Before we look at how human sphexishness is possible we have to get a better idea of what sphexishness *is*. Hofstadter (1986) describes antisphexishness, the opposite of sphexishness, as 'a general *sensitivity to patterns*', '*an ability to see sameness*' (p. 531), an 'ability to *break out of loops of all sorts*' (p. 532) or to 'detect and terminate any and all patterned behavior' (p. 536 f.; all italics are his). But is being stuck in a loop or manifesting a behavioral pattern always bad? Only if the behavior in question is to the organism's disadvantage, if it is a "mistake". What in the behavior of an organism counts as a mistake? And under which circumstances would we expect an organism to be able to break out of a loop?

In a precursor (von Bülow 2001) of this paper, I have talked about sphexishness in terms of loops, but now I think it's better to use 'disposition' as the basic notion. We aren't interested in an organism repeating some behavior because of an unlikely series of accidents; the kind of loop we are interested in is caused by a certain *disposition* of the organism caught in it. Breaking out of a loop is not interesting if it happens purely by chance; we want the organism to break out as a result of some

kind of adaptation, of "learning", i.e., of losing the disposition responsible for the loop.

Furthermore, since we do not want to restrict our concept of loops to behaviors recurring periodically in fixed time-intervals, we need to refer to whatever *triggers* the behavior. This will be some type of (external and/or internal) situation (or condition, or stimulus), and the organism's behavior may or may not be adequate for a given situation of that type. So it will be natural to talk about sphexishness in terms of dispositions to perform certain *behaviors* in certain types of *situations*.

I use the word 'disposition' in the following way. If o is an organism (it might be you or me, or the wasp in the experiment, or the tulip on your window-sill), S is a type of situation (feeling tired, meeting people, encountering the cricket on the burrow's threshold after having checked the interior, being in sunshine), B is a behavior (drinking a cup of coffee, behaving friendly, pulling the cricket into the burrow, turning your leaves toward the sun), and t is a point of time, then I say:

o **has (at** t**) the disposition to do** B **in situations of type** S**:**

> at t, the overall makeup (or structure, or organization) of o is such that being in a situation of type S will cause o to perform the behavior B (so, if o should be in a situation of type S at t then o would do B).

I will denote dispositions by 'S/B'. Instead of 'situation of type S' I will usually write 'S-situation'.

[[1] I use upper-case letters for types and lower-case letters for tokens. So s might be a certain situation of type S, and b a certain manifestation of the behavior B. – The (type of) situation S and the behavior B may share some "parameters", e.g., the cricket and the burrow, so the definition above might also have said something like, 'for all x, y, being in a situation of type S (or "S(x, y)") with respect to x, y will cause o to perform the behavior B (or "B(x, y)") with respect to x, y'.]

I use 'situation' in a very broad sense that includes the "internal situation", or *state*, o is in; and by 'state' I do not only mean o's physical state (body temperature, energy resources, perceptive powers, skeletal frame, hormonal state, etc.) but his mental or psychological state as well (e.g., o's beliefs and desires, his skills and his memories, which topics he is interested in, the thoughts, emotions, wants, or actions he is currently occupied with, his mood, his focus of attention) – if o has such. Thus, S might be 'feeling bored' or 'believing oneself to be a good tennis player'.

I use 'behavior' in a broad sense, too, comprising anything the person or body does, intentionally or involuntarily, externally or internally, physically or mentally: reflexes, digestion, heartbeat, thinking, remembering, imagining, intentional acts, interpreting external or internal events, and so on.

[A behavior may have a complicated structure, much like a computer program: the behavior B might consist in first doing B_1, then B_2, then repeating B_1 until some condition C is met, then doing B_3, and then either B_4 or B_5, depending on whether C′ is the case or not. So, a behavior may be "assembled" from simpler behaviors by way of devices like concatenation ('do this, then do that'), loops (in the computer-science sense: 'do this eleven times' or 'while C, do this' or 'do that until C′') and conditional branching ('if C, then do this, else do that'). – If o has a disposition S/B, and B is of the form 'if C, do B_C, else do $B_{\neg C}$', we might as well speak of o having *two* dispositions, namely, $(S \wedge C)/B_C$ and $(S \wedge \neg C)/B_{\neg C}$, where $S \wedge C$ and $S \wedge \neg C$ are the two possible ways of S obtaining, in regard to C. Neither of the two descriptions of o's

[1] This is the first of many technical remarks not essential for understanding the paper. Readers should feel free to skip them.

dispositional status is more correct than the other, though one may be more fruitful than the other.

On the other hand, a behavior B may also be very unspecific, like 'being friendly' or 'moving'. Then, doing B will amount to being friendly, or moving, in whichever available way "seems best" to o under the circumstances. The behavior B might even be 'living', but dispositions to *live* in certain types of situation are bound to be uninteresting, since we are dealing only with living organisms anyway.

Some may object to my talking about behavior being *caused* when o is a person. It may seem to them that either I exclude, e.g., intentional acts from the range of behaviors B which may figure in a disposition S/B, or else I deny o's freedom of the will. I don't want to do either. I consider causation (or determination) of behavior to be compatible with freedom; in the case of intentional acts, the most salient causal factors may best be described as o's desire for some state of affairs, X, and o's belief that B is the best way to achieve X. But if you are uncomfortable with the notion of causation in this context you are welcome to replace my formulation with 'will reliably result in o's deciding to do B', or something to that effect, for the appropriate cases. I cannot delve into this matter; please read Dennett 1984.]

Every "behavior" in this sense constitutes a type of "situation", as I use the term (but not vice versa). Therefore we can have chains of interlocking dispositions of o, and corresponding cascades of behaviors, once the first disposition is manifested: a situation might trigger an interpretation, which brings about certain thoughts and emotions which in turn are accompanied by a subliminal muscular tension, and as a reaction to the situation so interpreted there arises an urge to act in a certain way.

2.3 What we can do: Practical capability

Sphex has the disposition to make an inspection of her burrow after putting the cricket on its threshold (and before dragging it inside). Obviously, this has worked well enough in the past for digger wasps not to have died out. But in the experimental situation described by Wooldridge, where a biologist interferes with the normal run of events, this disposition works to the wasp's disadvantage. She has the disposition to make an inspection of the burrow *without* regard to whether she has already made one, and so time after time the biologist gets the opportunity to move the cricket away from the threshold. Thus the wasp can be caught in a futile loop where she spends lots of time and energy without making any headway.

Sphex in her rut looks stupid to us; the loop reveals that *Sphex* doesn't have an inkling of what she is doing. But do we really *expect* her to know what she is doing? It would be "unfair" to expect her, say, to lurk just inside the burrow to find out what's going on and then sting the saboteur's finger. We might as well praise her that things aren't worse: *Sphex* at least notices that the cricket isn't where it should be; she does not, upon finishing her examination of the burrow, just make dragging motions from the threshold to the interior.

In calling an organism o sphexish we want to point out a certain deficiency of o, the inability to stop repeating the same mistake and do something more useful instead. But we aren't interested in just any kind of inability whatsoever. We don't blame *Sphex* for not tying the cricket to a stake in front of the burrow, nor for not having X-ray eyes which would render her inspection tours unnecessary. We recognize that these aren't options for *Sphex*; nobody would expect her to do, or be like, *that*.

What we are interested in is *Sphex*'s inability to do something that we would consider very *easy* for her. It seems "fair" to expect that after a few repetitions, *Sphex* should notice the futility of further inspections of the burrow and drag the cricket

right in. What, if anything, justifies this expectation? Well, dragging the cricket in without further ado is clearly the right thing for her to do in that situation, and she seems perfectly *capable* of doing so; after all, she does pull crickets into burrows all the time.

But then again, her getting caught in that loop is evidence that, in some sense, she is *incapable* of dragging the cricket straight in after getting it up to the threshold, incapable of shaking off her disposition to examine the burrow after putting the cricket on the threshold. Now, does this evidence make us shed our prior belief that *Sphex* should in some sense be able to do the right thing? I think not. Though the experiment is evidence of what *Sphex* is, or is not, capable of, it is at the same time evidence for us of her "stupidity" for not being able to do something that should be so easy for her.

This suggests that there are two senses of 'capable' relevant here. I would like to say that *Sphex* is, though *technically* capable, *practically* incapable of doing the right thing. – One more example, before I try to make these senses of 'capable' more precise.

Imagine that Otto is in love with Anna, but is very afraid to call her and tell her so. (Also imagine circumstances under which the latter would be the best thing for him to do.) He *wants* to call her and thinks about doing it; he tries to *make* himself do it – but fails, time after time. So, in some sense it is very difficult for him to call Anna and tell her that he has fallen in love with her.

But in another sense it is very easy for him to do: he just has to walk to the phone, pick up the receiver, dial Anna's number and, when she answers his call, speak the words, 'Anna, I'm in love with you'. Surely there is nothing difficult about making phone calls or speaking, whatever the words? It's not like he wanted to run a hundred meters in five seconds!

Now, is this phone call easy for him, or difficult? I would say it is *practically* difficult for him, although *technically* easy. But what precise meaning is this supposed to have?

I intend the concept of *practical* capability to capture more or less what we observe in practice. If *Sphex* again and again fails to drag the cricket straight into the burrow after bringing it up to the threshold then *Sphex* is practically incapable of dragging it straight in. If Otto again and again tries to make himself call Anna, without success, then Otto is practically incapable of doing it, or at least it is practically very difficult for him. On the other hand, if Michael Jordan succeeds 99 times out of a hundred in throwing a basketball through the basket from the free throw line then he is not only practically *capable* of doing this, it is practically *easy* for him.

In judging about o's practical capability to do B in situations of type S, it will not do to observe just a few attempts at, or opportunities for, doing B, because we might be misled by chance failures or successes. Even looking at very many attempts/opportunities may not give us a correct picture if the sample set isn't *representative*, e.g., if, out of an even larger set of attempts, we look at only the failed ones. I cannot elaborate on the notion of a representative sample; I just assume that it is sufficiently clear and unproblematic for my purposes here.

It will in general not be possible to make a series of experiments to test for practical capability, as in the case of *Sphex*. In particular where failure at doing B in an S-situation means death, the organism can only fail once. Therefore a general definition of practical capability cannot refer to representative sets of *factual* attempts/opportunities, but instead has to take recourse to possible or counterfactual attempts.

Thus we arrive at the following definition:

o is (at t) practically capable of doing B in S-situations:

for any representative set of S-situations, Σ, there are at least *some* $s \in \Sigma$ such that if o should be in s at t then o would do B.

Instead of 'o is practically capable of doing B in S-situations' we may also say, 'doing B in S-situations is practically *possible* for o'.

[Readers may feel that this definition lacks some phrase like, 'if o tried to do B', or, 'if o wanted to do B'. But *with* some such phrase the definition would yield unwanted results when applied to, e.g., *Sphex* and omission of the inspection tour: *Sphex* cannot even *try* to refrain from inspecting the burrow; therefore the supplemented if–then clause ('if o should be in s at t, and tried to do B, then o *would* do B') will always be trivially true; therefore *Sphex* would be practically *capable* of refraining, contrary to what we observe in practice. That is why, earlier on, I have been talking about "attempts *or opportunities*" for doing B. There may be some behaviors of *Sphex* we *can* reasonably call attempts, e.g., if she were to move a cricket towards the interior of the burrow, but failed to drag it inside because it was too big; but I don't want to restrict the definition to those behaviors. – This has the seemingly counter-intuitive consequence that if you abstain from doing B simply because you do not want to do B then you are *ipso facto* practically incapable of doing B (under those circumstances). Nevertheless you will maintain that you are perfectly capable of doing B even if you don't want to. And indeed you are, only this capability is an instance of *technical* capability. – For humans I will in general tacitly assume as part of the situation S that they *want* to execute the (useful) behavior B in question.]

Practical ease is defined analogously:

Doing B in S-situations is practically easy for o (at t):

for any representative set of S-situations, Σ, for *most* or *all* $s \in \Sigma$, if o should be in s at t then o would do B.

Obviously, if doing B in S-situations is practically *easy* for o then it is practically *possible* for o, i.e., o is practically *capable* of doing B in S-situations.

Doing B in S-situations is practically difficult for o (at t):

doing B in S-situations *isn't* practically easy for o at t.

[Alternatively, practical difficulty might be defined following the same schema as before: doing B in S-situations is practically difficult for o at t iff, for any representative set of S-situations, Σ, there are *many* $s \in \Sigma$ such that if o should be in s at t then o would *not* do B. Hopefully, this alternative definition would be equivalent to the one given, assuming that there *are* representative sets of S-situations and that all representative sets yield similar results.

All three notions defined are *fuzzy*: instead of a sharp boundary separating the cases where the respective condition is satisfied from those where it isn't, there will in general be many cases with an unclear status. But since there are also many cases with a *clear* status this shouldn't bother us, as long as we keep to the latter ones. – Furthermore, practical ease and difficulty are *gradual* notions: doing B can be practically easier, or less easy, and accordingly less or more *difficult* practically, than doing B′.

That o is practically incapable of doing B in S-situations implies that doing B in S-situations is practically difficult for o. It may seem somewhat unnatural (except to mathematicians) to say that doing B is *difficult* for o if o is *incapable* of doing B. Nevertheless I think it worthwhile to consider incapability/impossibility as an extreme case of difficulty. I will buffer the inconvenience by writing, redundantly, 'doing B is difficult *or impossible* for o'.

Practical capability/possibility is a modal notion. We could define the corresponding notion of necessity by saying: o is at t *practically compelled* to do B in S-situations iff o is at t practically *incapable* of *not* doing B in S-situations; or, alternatively: ... iff, for any representative

set of S-situations, Σ, for *all* $s \in \Sigma$, if o should be in s at t then o would do B. This obviously is a special case of practical ease. – That o has the disposition S/B entails that o is practically compelled to do B in S-situations, and I think also vice versa; therefore we will not need this new term.]

2.4 What we *could* do: Technical capability

To explicate *technical* capability, we investigate how practical difficulty might be turned into practical ease. If we want to claim that doing B in S-situations is technically easy for o then we have to point out a way of changing o, B, or S such that the *new* "task" is practically easy, the existence of which justifies considering the *old* task easy as well, even if it is practically difficult. A task is technically easy if no "big" change is required to turn it into a practically easy one. The question now is, which kind of changes are "small" enough?

Riding his bicycle instead of calling Anna would be practically easy for Otto, but this doesn't tell us anything about why it is reasonable to say that calling Anna is in some sense very easy for Otto. Putting the paralyzed cricket into the burrow in spite of the biologist's interference would be practically very easy for a human, but this doesn't give us a clue about in which sense it should be easy for a digger wasp.

So, taking a completely different behavior, or taking a completely different (kind of) organism instead of o, are changes too far-reaching to give us grounds for judging the old, unchanged, task easy. – Now, if there are behaviors B′ approximately as demanding as B, and most of *them* are practically easy for o in S-situations, maybe then we might feel justified in expecting B, too, to be easy for o?

This approach won't work either. If we take recourse to behaviors B′ "approximately as demanding" as B to justify calling B itself technically easy we are begging the question. 'As demanding' surely cannot mean 'as demanding *practically*' here because B′ is by assumption practically not very demanding (i.e., difficult) at all, whereas B *is* practically difficult (if its technical ease is to be interesting). Rather, it will mean '*technically* as demanding/difficult', and so we are presupposing the notion we want to explain.

Maybe doing B is technically easy if doing any proper *part* of B is practically easy in an appropriate kind of situation? Doing all those easy parts in the right combination should then be easy, too, one might think.

For example, while it is practically difficult for Otto to call Anna and tell her he's in love with her, it might at the same time, and in the same type of situation, be practically *easy* for him to call Anna and ask her for the topic of next week's philosophy colloquium, never once hinting at his infatuation with her. It would take very special circumstances to make him *fail* to do so if he wanted to; say, that his telephone were defect or that he had lost Anna's number. It would also be practically easy for him (if he wanted it) to say out loud, 'Anna, I'm in love with you', in a place where no one can hear him.

But although, e.g., drinking and breathing are both practically easy for us, drinking and breathing at the same time is practically easy only for infants. So, we would not be justified in expecting even simple kinds of combinations of easy tasks always to be easy, too.

Anyway, *Sphex*, for example, doesn't have the behavioral flexibility presupposed by this sort of argument. This seems to be her problem: that she isn't even flexible enough to *omit* her inspection tour when it is inappropriate.

We will get a better grasp of the notion of technical capability by thinking about the *causes* why a (supposedly) technically easy task may be practically difficult.

It is practically difficult for Otto to call Anna and tell her he's in love with her because he is afraid, say, of risking rejection. There might be situations of the same type where he succeeds in calling Anna, e.g., if a friend had just eloquently encouraged him and thereby diminished his fear, or if he had just accomplished some great feat and still was in high spirits. Situations of this very special kind may be included in a representative set, too, but they would form a minority.

[More generally, that doing B in S-situations is practically *difficult* for o is fully compatible with there being a proper subtype S^* of S such that doing B in S^*-situations is practically *easy* for o; and vice versa.]

Otto's difficulty consists in his fear of rejection. For a claustrophobic, entering an elevator will be practically difficult because of his fear of narrow spaces; for someone suffering from an obsessive-compulsive disorder, not washing his hands when he believes them dirty will be practically difficult, presumably because of his fear of disease. For a nicotine addict, refraining from smoking over a substantial stretch of time will be practically difficult, though not because of fear but because of his craving a smoke; for a sailor on a ship whose supply of drinking water has run out, refraining from drinking sea-water will be practically difficult, again because of his craving.

Not only negative feelings or deficiency-induced craving can make "easy" tasks practically difficult. One might also *enjoy* doing something so much as to be unable to bring it to its intended conclusion, as in the case of premature ejaculation.

These examples suggest that feelings and urges or something of their kind are what sometimes makes it (practically) hard for people to perform seemingly simple behaviors. There may be persons, e.g., someone severely retarded, for whom it is practically difficult or impossible to make *any* kind of phone call all by themselves, i.e., making phone calls is *technically* difficult for them. But Otto doesn't need to be smarter or more dextrous to make his phone call; he has the knowledge and skills and the "physical" ability necessary to do what he wants. He could do it right away – if he weren't so afraid, if he cared less about Anna's reaction, if he *felt* differently. The sailor dying of thirst could easily refrain from drinking sea-water – if he didn't feel his thirst and the accompanying urge to imbibe any available liquid. Technically, there is nothing difficult about the behaviors mentioned in the examples, especially where the behavior consists in letting some act be; the difficulty lies in what the behaviors and their consequences "mean" to those people (or their bodies).

Nevertheless we cannot define o's being technically capable of doing B (in S-situations) by saying that B would be practically possible or easy for o if o *felt* about B in some appropriate way. This is because a definition of that sort would yield incorrect results for simple organisms like *Sphex*, who presumably aren't capable of feeling anything. But if we talk about *representation* instead, we get a definition that works even for some kinds of automata.

Sphex may be described like an automaton: When she has deposited the cricket at the burrow's threshold she is in a state σ_{insp} that makes her execute an inspection of the burrow. After returning from her inspection tour, *Sphex* is in a state σ_{pull} such that she proceeds to pull the cricket into the burrow if it is still at the threshold. If the cricket has been moved away σ_{pull} makes way for some state σ_{find} such that *Sphex* searches for the cricket and, if successful, brings it back to the threshold, thereupon entering state σ_{insp} again. But σ_{insp} has no variant $\sigma_{\text{insp}}^{\text{done}}$ which would make her drag

the cricket right in, "remembering" that she has already disposed of her inspection tour. When *Sphex* puts a cricket at the threshold she goes into state σ_{insp}, and σ_{insp} "means" that she now has to inspect the burrow. And so it goes on and on.

These states represent the wasp's situation: what she has already done and what she must still do. If, after dragging the cricket back to the threshold for the umpteenth time, she wouldn't be in state σ_{insp} ('cricket ready, now inspect the burrow!') but in σ_{pull} instead ('inspection accomplished, now pull the cricket in!') she would have broken out of her loop. Dragging the cricket inside would then be practically easy. If only *Sphex* had a different *representation* of her momentary situation, doing the right thing would be a piece of cake (practically).

[My characterization of these states is of course far from complete. After inspecting the burrow, *Sphex* is in a state σ_{pull} which works in such a way (which "means") that, in the *normal* run of events, she then pulls the cricket inside; which is why we may interpret σ_{pull} as "saying": 'inspection accomplished, now pull the cricket in!' But if circumstances require it this aspect of σ_{pull} may be overridden by more immediate tasks, as, for example, when the cricket has been moved. So, a better interpretation of σ_{pull} would be: 'inspection accomplished; now pull the cricket in, if it is at the threshold, else enter state σ_{find}!' Supposedly *Sphex* is (practically) capable in σ_{pull} of reacting to further contingencies, e.g., that a *Sphex*-eating predator appears on the scene or that rain heavy enough to endanger her starts to fall. A complete characterization of the "meaning" of σ_{pull}, of what dispositions *Sphex* has in that state, would thus be very complicated, but those complications are irrelevant to the questions at hand.

You may ask, 'if *Sphex* is in a state that "means" she has just accomplished her inspection, but actually she *hasn't* just accomplished her inspection (she did it a while ago), why should I consider that state a *representation* of her situation?' – This question would be the symptom of a misconception. To know what *Sphex*'s representation of her situation is you don't have to look at her situation, decide what would be appropriate representations of it (for *Sphex*), and then look at *Sphex* to see whether she "has" one of these representations or not. Rather, you just have to look at *Sphex* to see what kind of situation she "believes" herself to be in. (Afterwards you can still inquire how good is the "fit" between this representation and her actual situation.) In other words, 'o's representation of his momentary situation' is not a function of o's situation in general, but only of o's *internal* situation or state. Of course, o's representation of his situation will normally track his actual situation as it changes, and thus o's representation depends *causally*, while not conceptually, on his actual situation (more precisely, on which situations he has actually been in in the past).]

This talk about representation seems to me sufficiently general to capture the causes for the respective practical difficulties/incapabilities of *Sphex* as well as of the different persons in my examples. To overcome their practical difficulties, to render their tasks practically easy, none of them need greater practical or intellectual skill or power, or better minds, brains or bodies, than they already have *except insofar* as they would have to harbour representations of the goings-on different from those they actually have.

Now, it isn't quite sufficient to say:

> o is *technically capable* of doing B in S-situations iff there is a way, R, for o of (mis-)representing his momentary situation such that for any representative set of S-situations, Σ, there are at least *some* $s \in \Sigma$ such that if o should be in s, and "did" R, then o would do B.

Maybe *Sphex would* drag the cricket right into the burrow if she believed herself to be entering the White House upon an invitation for a candlelight dinner with Bill Clinton, carrying a complimentary cigar; but she will never, under no circumstances whatsoever, *entertain* such a representation, so this mere logical possibility doesn't matter when we assess what she is technically capable of.

For R to matter, o must be (practically) capable of "doing" R under *some* circumstances S_R. It would be too much to ask for o's being practically capable of "doing" R in situations of the very type S under consideration, because I suppose most organisms don't have much latitude in representing their momentary situation. For this same reason it will not make much difference whether we require R to be practically *possible* or practically *easy* for o in S_R-situations: the former presumably entails the latter. In order to render the definition slightly (maybe only superficially) more restrictive, I will use 'easy'. The definition then goes like this:

o **is (at t) technically capable of doing** B **in** S**-situations:**

> there is a way, R, for o of (mis-)representing his momentary situation and there is a type of situation, S_R, such that "doing" R in S_R-situations is practically easy for o at t, and for any representative set of S-situations, Σ, there are at least *some* $s \in \Sigma$ such that if o should be in s at t, and "did" R, then o would do B.

Practical capability entails technical capability: just take for R whatever representation of his situation o would actually have in situations of type S, and take S for S_R.

Technical ease and difficulty are defined analogously:

Doing B **in** S**-situations is technically easy for** o **(at** t**):**

> there is a way, R, for o of (mis-)representing his momentary situation and there is a type of situation, S_R, such that "doing" R in S_R-situations is practically easy for o at t, and for any representative set of S-situations, Σ, for *most* or *all* $s \in \Sigma$, if o should be in s at t, and "did" R, then o would do B.

As before, ease entails capability.

Doing B **in** S**-situations is technically difficult for** o **(at** t**):**

> doing B in S-situations *isn't* technically easy for o at t.

Take Otto, for example, sitting in front of the telephone trying to work up the nerve to call Anna and confess his love for her. The way he represents his situation to himself, one likely outcome of this phone call is that Anna would despise him for his presumption because she is such a high-class girl and he is such an unremarkable guy.

(Calling her will be a daunting task if Otto believes this outright. It would be easier if he were very certain intellectually, "in his head", that the phone call couldn't do much harm, and quite possibly much good. But even then it would not be *much* easier if, "in his guts", he still felt that it *would* harm him a lot. – Later on I hope to shed some light on the possibility of having at the same time, but in different ways, two contrary beliefs.)

Now imagine that Otto wrongly believes himself to be in a very different situation: he is convinced beyond all doubt that Anna loves him too, but he also believes that she thinks he doesn't know this and furthermore expects him to make the first move. Under the sway of this illusion the phone call would certainly be practically easy for him.

Or he might wrongly believe he isn't really talking to Anna but instead is talking to an actress while taking part in a role-playing game. Or, closer to reality, he might *disbelieve* in his own unworthiness and in Anna's adorability (and arrogance). In these cases, the task of calling Anna would presumably be practically easy, too.

That is why calling Anna is *technically* easy under the original circumstances: because there are ways for Otto of (mis-)representing the situation which make the task practically easy.

This is of course not to say that Otto can voluntarily *bring about* these belief states in himself. On the contrary, it is highly improbable that he acquires any of them, at least in the short run. (He might voluntarily engage in some therapeutic activity which, in the long run, results in him losing his feelings of inferiority and his anxiousness.) For a task to be technically easy for o it only matters that these representations or belief states *exist* and that the task *would* be practically easy for o if he were in such a state – no matter how, and whether, this might actually come about.

It isn't hard similarly to imagine possible (mis-)representations of their "tasks" and situations for the claustrophobic and the compulsive: The claustrophobic might see, and believe himself in, a ballroom instead of an elevator, and might fail to register the elevator's up- or downward acceleration. (This would of course have to be a hallucination, but would nevertheless be a (*mis*-)representation of his situation.) The compulsive might believe his hands to be perfectly clean. In both cases the task would become practically easy.

But the sense of '(mis-)representation' I need is broader than the foregoing examples reveal. Not only perceptions, beliefs, and the resulting feelings and behavioral inclinations count as (mis-)representations. Consider the smoker and the thirsty sailor. Both crave for a certain substance (nicotine and sea-water, respectively), which is their bodies' way of "telling" them (and "believing"), 'I – that is: you – absolutely immediately need this stuff! Get it, or else we'll incur severe damage!' At the same time, they both know that consuming the substances they crave wouldn't really be such a good idea. Their actual (mis-)representations of their respective situations thus include opposing motivating forces, one more rational, one less so. Now, if they didn't have the craving-parts of their (mis-)representations it would be practically easy for them to refrain from ingesting the deleterious substances.

Craving for something, or feeling, like the compulsive does, an obsessive urge to do something, are both (mis-)representations of the obtaining situations, namely, of what those situations require. Correspondingly, *not* craving or *not* feeling some urge are further possible (mis-)representations. I consider as belonging to a person's (mis-)representation of their situation not only what they think and believe about it and what they see, hear or otherwise perceive of it, but also what they want or desire, what they are more or less strongly *compelled* to do (or not to do) in it. Thus even reflex acts like eye blinks or knee jerks (more accurately: the internal states producing them) form part of (mis-)representations in this broad sense: they are, as it were, hard-wired "beliefs" of the body that it must do such and such in situations of type so-and-so to avoid harm. Reflexes are those behaviors where we are on a par with *Sphex*.

[I believe my concept of representation more or less fits in with Dennett's (1981, 1987) – though I'm not sure what *exactly* his concept is.]

If we accept this liberal understanding of '(mis-)representation' then even acts like cutting off one's own finger aren't so difficult technically, provided one has a good knife. Though Japanese *yakuza* may do this without much hesitation when they have failed their bosses, for me it would be next to impossible practically: the thought alone of cutting off my finger is for me highly repellent, and the pain resulting from the process will normally be insurmountable. But both the thought and

the pain are "only" representations of what I am doing and so we can neglect them when we think about technical capability. If I thought this was a twig instead of my finger, and if I didn't feel the pain, then it would be practically easy for me to cut off my finger; therefore it is technically easy.

[We might obtain various concepts of technical capability by excluding from variation this or that "mode" of representation. We might, for example, get a more restrictive concept if we allowed only for different *beliefs* about his situation from those o actually has; then o would be "b-technically capable" of doing B iff o would be practically capable of doing B in an appropriate belief state. But it might prove difficult to keep those different "modes" of representation cleanly apart, conceptually.]

2.5 Mistakes: Suboptimality

Equipped with these two notions of 'capability', we can now start to think about what counts as a mistake and what exactly it is that seems "stupid" to us in *Sphex*'s behavior. The first thing that makes a behavior a mistake is that it is *bad* for you, that it is against your *interests*.

[I assume that every organism has interests: survival and reproduction above all, but also many more "refined" interests, which can be quite idiosyncratic in humans. For an account of the origin of interests, see Dennett 1984, Section 2.1, 'Where Do Reasons Come From?', esp. pp. 21 ff.]

When is a behavior against your interests? This depends on circumstances. One and the same behavior can be bad for you in one type of situation, and good, in another. Depending on the situation, there are different sets of options open to you, and one option may have different degrees of utility. A behavior that seems "intrinsically" good for you is, all things considered, *bad* for you in a certain type of situation, S, if (in S) there are options still better for you (in S). A behavior that under most circumstances would be bad for you (e.g., cutting off your finger) may be good for you under certain circumstances S because it is the best you can do in S (say, if your finger is caught in the wall at the bottom of a well and the well is filling up and you would otherwise drown).

So, a behavior B is against your interests in situations of type S if you can do some B^+ in S-situations that would be better for you. Do you have to be *practically* capable of doing B^+, or would *technical* capability to do B^+ suffice to justify calling B a mistake? Intuitions here are clearer if we look at the case of *Sphex*, which cannot choose between different options; 'practical capability' is for her almost co-extensional[c] with 'practical compulsion'. *Sphex* is practically incapable of dragging the cricket right into the burrow after bringing it up to the threshold, but still her intervening inspection tour is a "mistake", against her interests (in the experimental setting) – because technically she is capable of dragging the cricket straight in. [c]?

Similarly in the case of persons: even when you do B under the influence of a strong compulsion, so that it is practically difficult or impossible for you to refrain from doing B, you still act against your interests in doing B if technically you are capable of doing some more useful B^+.

Is there an even broader concept of capability such that being capable of doing B^+ in such a weak sense would already be enough ground to consider doing B a "mistake"? – I cannot see any candidate.

S/B is suboptimal for o (at t) with respect to B^+:

(SO_1) o has at t the disposition S/B;

(SO_2) in general, B^+ is more useful for o in S-situations than B;

(SO_3) o is at t technically capable of doing B^+ in S-situations.

[d] Beispiele für B^+!

[d] Here, 'in general' means, averaging over a representative sample of S-situations.

S/B **is suboptimal for** o **(at** t**):**

there is a behavior B^+ such that S/B is suboptimal for o at t w.r.t. B^+.

S/B **is remediably suboptimal for** o **(at** t**) w.r.t.** B^+ **and** M**:**

(RSO_1) S/B is suboptimal for o at t w.r.t. B^+;

(RSO_2) the behavior M is a method for o of switching from S/B to S/B^+;

(RSO_3) all things considered, it would be advantageous for o at t to do M;

(RSO_4) o is at t technically capable of performing M under the impression of (repeatedly) doing B in S-situations.

In (RSO_3), among "all things" to consider are chiefly the following:

- the future frequency of S-situations,
- the relative usefulness of B^+ for o in S-situations, compared to B,
- the costs and side-effects of doing M.

S/B **is remediably suboptimal for** o **(at** t**):**

there are behaviors B^+ and M such that S/B is remediably suboptimal for o at t w.r.t. B^+ and M.

o **is sphexish (at** t**) w.r.t.** S/B**,** B^+**,** M**:**

(Sph_1) S/B is remediably suboptimal for o at t w.r.t. B^+ and M,

(Sph_2) o is – though *technically* capable, cf. (RSO_4) – practically incapable at t of doing M even after (repeatedly) doing B in S-situations.

[e] [f]

[e] System aus *einem* self-reinforcing meme?? 'This is true'? Self-fulfilling prophecies?! Oder zu durchschaubar, um zu fkt.en?

[f] 'Psychologists won't be surprised by these ideas, but I hope to convince analytic philosophers of points they are not normally convinced of.'

References

Dennett, Daniel C. 1981. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, Mass./London: MIT Press. Original edition 1978; sixth printing 1993.

———. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, Mass.: MIT Press.

———. 1987. *The Intentional Stance*. Cambridge, Mass./London: MIT Press.

Hofstadter, Douglas R. 1986. *Metamagical Themas: Questing for the Essence of Mind and Pattern*. Toronto/New York/London/Sydney/Auckland: Bantam.

von Bülow, Christopher. 2001, March. Menschliche *sphexishness*: Warum wir immer wieder dieselben Fehler machen. `www.inf.uni-konstanz.de/~buelow/sphex.pdf`.

Watzlawick, Paul. 1976. *Wie wirklich ist die Wirklichkeit? Wahn, Täuschung, Verstehen*. München/Zürich: Piper.

Wooldridge, Dean E. 1968. *Mechanical Man: The Physical Basis of Intelligent Life*. New York/San Francisco/Toronto/London/Sydney: McGraw-Hill.