

The logic of deterrence. A beginning¹

GEORG MEGGLE

Abstract

Research on deterrence phenomena has been coached in most cases in the frame of game theory, i.e., in numerical terms. This paper starts to spell out the logics of deterrence in qualitative terms, i.e., by using the action theoretical terms of Doing, Wanting and Believing as its basics. The focus is on (successful) consequence-based deterrence attempts being treated as a special kind of (successful) intentional action. Not-successful deterrence is to be distinguished from deterrence failure. The various reasons for not-successful deterrence give us a map of what one has to care about in order to maximize the chances of deterrence attempts being successful – and to maximize the chances of counteraction as well. But note, that this is just the first step in a bigger project.

Deterrence phenomena play an important part in the lives of all living creatures (even those lower down the scale of development) – and humans are no exception. Deterrence strategies are part of our behaviour repertoire in very different areas of life. They are a fundamental element within our upbringing. They dominate broad areas of the preventive components of criminal law. They define and uphold many personal relationships and ties – yet can also spell their collapse. Without all their supposedly carefully weighed-up deterrence calculations, business and politics would probably be unable to function. And if the superpowers' mutual nuclear deterrence didn't work, mankind's very survival would be at stake (as indeed has been the case on occasion). In fact it is probably due to this belief that ever since the Cold War (if not before), deterrence has for many been the quintessential means of preventing the outbreak of military conflict.

The classical instrument for analysing deterrence scenarios also goes back to the early days of the Cold War, namely mathematical *game theory* using metric concepts, a special type of applied *rational decision theory*. This product of the workshops of chiefly American intellectual arms research has numerous invaluable applications which now are also known to almost everyone in philosophical circles, especially those involved in analytic philosophy.

Don't get me wrong – by no means do I wish to belittle the merits of this analytical instrument. After all, it is the best one we have if we want to examine *certain* situations under a microscope. But a magnification of 100x or even more is not ideal for all purposes; sometimes 20x or 10x is simply better. Those seeking to get their bearings are best served by a large-scale map. Yet just such a *large-scale map* has so far been unavailable during the whole discussion on deterrence. Those involved have often not seen the wood for the trees described

¹ Paper given at the international congress on *War, Collective Responsibility and Reconciliation* in Belgrad, June 1998. A german version of this paper was already presented at the 19th Wittgenstein Symposion (1996) in Kirchberg – which in turn was an extract taken from extensive notes on this topic I have accumulated piecemeal since around 1985. One day I'd like to get round to ... but that'll be the day! I'd like to thank all those who participated in Kirchberg as well as in Belgrad in the subsequent lively discussions.

at least fictitiously in quantitative terms. My goal is to provide a map which shows us the way again both in and around the wood of deterrence (although at this stage of course we will have to content ourselves with a very rough sketch).

Instead of using metric or even just comparative utility, preference and subjective probability concepts, I would only like to work (at least at the beginning, and thus exclusively throughout this paper) with their qualitative relatives. I will try and make do with action theory's trinity of *believing*, *wanting* and *doing*. Of course, I shall also include what we also still need even in such a primitive action logic, namely an intensional semantics, a simple concept of cause, and a few postulates for connections between non-simultaneous believing and wanting.

Once again for simplicity's sake, I shall accept the possibility of *extremely strong idealisation* among these basic concepts of believing and wanting. The simplest belief logic available is that of *strong (rational) belief*, and so that is what I shall use. And as far as *wanting* is concerned, I shall make use of that *strong concept of wanting* which fortunately is governed by almost exactly the same principles as this strong belief.²

The *long-term aim* is to develop a *deterrence logic*, namely a *formal language DET* in which the main concepts of deterrence can be expressed, and to determine an *interpretation concept* with which an inference concept for sentences containing DET-expressions can be formulated. Although we will still be miles away from such a logic even after the explanations given below, we shall (I hope) be able to see more clearly what other steps must be taken to achieve this objective.

Various concepts of deterrence (referred to below as DET-concepts). I shall distinguish between various DET-concepts which are closely related to one another. I will start with the most primitive concept which corresponds to the situation that Y *is deterred* from doing h by an event E. Here we are mainly interested in the special case of consequence-based deterrence, i.e. the fact that Y is deterred from doing h by E owing to the prospect of its consequence C. In a second step I shall then declare *attempts at deterrence* to be special intentional actions with the aim of causing such (consequence-oriented) deterrence, where once again I am particularly interested in the special case in which the consequence of deterrence is allegedly brought about by the deterrer himself. In a third step I shall then declare successful *attempts at deterrence* to be attempts which have actually achieved their aim of deterring the target of deterrence.

Thus explained, these DET-concepts are still extremely general. And so they should and must be at this first level. Special cases will only be defined in the special part of DET-logic (but at a later juncture; they will not be dealt with here).

² To be more precise, the following principles given in my book *Grundbegriffe der Kommunikation*, Berlin/New York, 1997, will be assumed:
 B(X,A) for: Y believes that A
 K(X,A) for: X knows that A – with $K(X,A) := B(X,A) \& A$
 W(X,A) for: X wants (strongly prefers) A
 D(X,f) for: X does f

1. Deterrence events

1.1 The first *necessary conditions* for Y being deterred by the event E from performing the action h – in short: DET(Y,E,h):³

- (1) E >> $\neg D'(Y,h)$
 E causes Y (at a time t' following E) not to do h.

(1) is of course not sufficient. Somebody has an accident – which trivially means that there are many things he no longer does for the simple reason that he can no longer do them. For example, the victim will now no longer be able to travel to Majorca for his next holiday. But it would be absurd if we were to say that the accident deterred or will deter the victim from spending his next holiday on Majorca. The victim simply won't be going anywhere on holiday, and so doesn't even have the possibility of going on holiday to Majorca.

Only somebody who can actually travel to Majorca can be deterred by an event from not going to Majorca. Thus the following is also required for DET(Y,E,h):

- (2a) POSS(D'(Y,h))
 It is possible that Y will do h.

However, the following is also essential:

- (2b) $B^{0r}(Y,POSS(D'(Y,h)))$
 Y believes (including at a time t^{0r} before t') that he will (at t') be able to do h.

Another consideration: If Y didn't want to go to Majorca in the first place, there'd be little point in posing the question of what *caused* him not to go there – not to mention the question of what *deterred* him from travelling there. In a nutshell, deterring events are only such if their effects are *omissions* (in a narrow sense), i.e. something which is not done even though one originally wanted to do it.

- (3) $W^{00r}(Y,D'(Y,h))$
 X wanted (at a time t^{00r} prior to time t^{0r}) to do h (at t').

And here's another Majorca example. Just as he does every year, Mr Y would have flown to Majorca on holiday this year too (action h) – if devastating forest fires hadn't broken out there (event E). Although he originally wanted to go to Majorca (= 3) and knows he could still do so (= 2a & 2b), he doesn't go after all (= $\neg D'(Y,h)$) – because of these forest fires (= 1). Hence now the forest fires raging on Majorca have deterred him from going to Majorca. The prospect of spending his holiday there of all places no longer appeals.

Expressed in more general terms, deterrent events cause someone who *wanted to do something* before they were an issue to *no longer wish to do it*. They require not merely (1), but also (1) to result from (4). And this is exactly what (5) says.

³ "A causes B" will in the following be represented by $A \gg B$, for which I will here only apply the primitive causal principles which I also used in *Grundbegriffe der Kommunikation*. For more on causal logic see F. von Kutschera, *Bewirken*, *Erkenntnis* 24, 1986, pp. 253–81.

The definiens of D1 is equivalent to the above conditions (1) to (5).

2 Consequence-based deterrent events

2.1 Let's go back to our holidaymaker before he becomes obsessed with forest fires, who originally wanted to go to Majorca on holiday, but was then put off the idea by the forest fires raging there, or the (correct or incorrect) information about them, or his (correct or incorrect) assumption about them. What is the basis of the effect of these fires, this information or this conviction which is so deterrent for him? Although we can't be sure, if he's anything like the rest of us, it'll probably comprise ideas to the following effect: When I go on holiday I expect a few days of maximum calm and relaxation etc. So I can do without a holiday on Majorca amidst an unpredictable inferno. Expressed in more general terms, Y imagines what a "holiday on Majorca" means this year, sees (or at least thinks he can see) the consequences this would have – and this immediately brings us to the promised additional requirements for what distinguishes *consequence-based deterrence*:

- (6a) $E \gg B^{0r}(Y, D'(Y, h) \gg C)$
E makes Y believe that his doing h would have the consequence C.
- (6b) $W^{0r}(Y, \neg C)$
Y wants not-C.
- (7) (4) as a consequence of (6a) & (6b)
The fact that E makes Y no longer want to do h is due to (6a) & (6b).

2.2 *Old or new preference?* Does it hold that Y's wanting not-C is only caused by E – or is Y being put off doing h compatible with the view that this wanting not-C has nothing to do with being caused by E? Not only is the latter view compatible with it; it also represents the norm.

The same goes for the intended deterrence. X regards the (supposed) fact $W^{0r}(Y, \neg C)$ as sufficient *reason* for Y to change his preference in response to E without X having to do f. Although in the following I generally refer to this normal case, I shall not stipulate it as a definition.

2.3 *Minimal rationality and (KANT)*. The fact that, as demanded by (7), the change of preference caused by E is due to (6a) and (6b) assumes the following principle:

- (KANT) $W(a, A) \ \& \ B(a, A \supset B) \ \rightarrow \ W(a, B)$
"He who wants the purpose (given that reason has a decisive influence on his actions) also wants [according to the agent's belief] the means necessary for the purpose." Kant, *Grundlegung zur Metaphysik der Sitten*, B 44–45.

2.4 Definition.

$$\begin{aligned}
 \text{D2:} \quad \text{DET}(Y,E,C,h) := & \quad (a) \quad \neg D'(Y,h) \\
 & \quad (b) \quad W^{00'}(Y,D'(Y,h)) \\
 & \quad (c) \quad E \gg B^{0'}(Y,D'(Y,h)) \gg C \\
 & \quad (d) \quad W^{0'}(Y,\neg C)
 \end{aligned}$$

Y is deterred from doing h in view of the (supposed) E-consequence C iff Y (a) doesn't do h, although (b) Y originally wanted to h, because (c) E made Y believe that doing h would bring about the consequence C – a consequence which (d) Y doesn't want to occur.

3 Attempts at deterrence

3.0 *Deterrence as the objective.* The previous DET-concepts exclusively stressed the genuinely deterring effect of certain events, actions or precautions, regardless of whether or not they actually were intended to have such an effect. Let us describe an action directed towards such effects occurring as an *attempt at deterrence* (DET-AT). It follows from the above distinction that once again we must first distinguish between two types of such attempts depending on whether the deterring effect is consequence-based.

DET-ATs are actions by which X (the *deterrence subject*) intends to cause Y (the *deterrence addressee*) to be deterred. They are thus special cases of '*instrumental actions*' – actions performed in an effort to achieve a certain *aim*.

3.1 Instrumental doing: the general case.

$$\begin{aligned}
 \text{D0:} \quad I(X,f,A') := & D(X,f) \ \& \ W^0(X,A') \ \& \ B^0(X,D(X,f)) \gg A' \\
 & \text{By doing } f \text{ X intends to bring about } A' \text{ (i.e. } A \text{ at } t') \text{ iff X does } f, \text{ X wants } A', \text{ and X} \\
 & \text{believes that by doing } f \text{ he makes it the case that } A'.
 \end{aligned}$$

Of course, X may have a number of aims he wishes to pursue with the same action f. The sum of these aims is then the *overall aim*.

3.2 *Instrumental action: special cases.* We must distinguish between three types of causing. X intends by doing f to *bring about* A' iff $I(X,f,A') \ \& \ B(X,\neg A)$. By doing f he intends to *maintain* A (including at t', i.e. to cause A (apart from at t) to still be the case at t') iff $I(X,f,A') \ \& \ B(X,A)$. And X intends by doing f to *guarantee* A' iff $I(X,f,A') \ \& \ \neg B(X,A) \ \& \ \neg B(X,\neg A)$.

Attempts at deterrence are attempts to cause Y *not* to do something, i.e. attempts of the general form $I(X,f,\neg A')$. Such attempts could also be termed *preventive attempts*.

3.3 Definition.

$$\begin{aligned}
 \text{D3:} \quad \text{DET-AT}_{\text{mg}}(X,Y,f,h) := & I(X,f,\text{DET}_g(Y,T(X,f),h)) \\
 & \text{Y tries by doing } f \text{ to (in a most general sense) deter}_{\text{mg}} \text{ Y from doing } h \text{ iff X intends by} \\
 & \text{doing } f \text{ that Y by X's doing } f \text{ is deterred}_g \text{ from doing } h.
 \end{aligned}$$

This definition leads us to the following theorem:

$$\begin{aligned}
\text{T.A1:} \quad \text{DET-AT}_{\text{mg}}(\text{X}, \text{Y}, \text{f}, \text{h}) \leftrightarrow & \quad (\text{i}) \text{ D}(\text{X}, \text{f}) \\
& \quad (\text{ii}) \text{ W}^0(\text{X}, \neg \text{D}'(\text{Y}, \text{h})) \\
& \quad (\text{iii}) \text{ B}^0(\text{X}, \text{W}^{00'}(\text{Y}, \text{D}'(\text{Y}, \text{h}))) \& \\
& \quad (\text{iv}) \text{ B}^0(\text{X}, \text{D}(\text{X}, \text{f}) \gg \text{W}^{00'}(\text{Y}, \neg \text{D}'(\text{Y}, \text{h})))
\end{aligned}$$

By doing f, X tries to deter Y from doing h (in a broader sense) iff

(i) X does f, (ii) X wants Y not to do h, even though (iii) Y (so X believes) originally wanted on his own initiative (i.e. without X doing f) to do h, and (iv) X believes that Y will only want to abandon doing h precisely in response to X doing f.

We must distinguish between *deterrent aims* and *deterrence aims*. The former are the aims which X must have in order for his action to be regarded as an attempt at deterrence (i.e. goals which are by definition essential for such a DET-attempt), while the latter are the other aims pursued by X by means of his attempt at deterrence (going beyond such deterrent aims).

The aim of the attempt at deterrence that Y does not do h is the *primary deterrent aim*. It is on account of this deterrent aim that X starts the whole attempt at deterrence in the first place, and from which all his other deterrent aims (for example Y himself not wanting to do h) connected (conceptually) with this attempt are derived.

3.4 This concept of an attempt at deterrence is formulated such that it is not yet certain whether these attempts are attempts of bringing about, maintaining or guaranteeing: $\text{DET-AT}_{\text{mg}}(\text{X}, \text{Y}, \text{f}, \text{h})$ is compatible with not only $\text{B}(\text{X}, \neg \text{D}(\text{Y}, \text{h}))$ but also $\text{B}(\text{X}, \text{D}(\text{Y}, \text{h}))$ and $\neg \text{B}(\text{X}, \neg \text{D}(\text{Y}, \text{h}))$. (Note, however, that: $\text{DET-AD}_{\text{mg}}(\text{X}, \text{Y}, \text{f}, \text{h}) \rightarrow \text{B}(\text{X}, \text{D}'(\text{Y}, \text{h}))$.) This generality is also present in the following concepts of attempts at deterrence.

However:

$$\text{T.2:} \quad \text{DET-AT}_{\text{mg}}(\text{X}, \text{Y}, \text{f}, \text{h}) \rightarrow \text{I}(\text{X}, \text{f}, \neg \text{W}^{00'}(\text{Y}, \text{D}'(\text{Y}, \text{h}))) \& \text{B}(\text{X}, \text{W}^{00'}(\text{Y}, \text{D}'(\text{Y}, \text{h})))$$

$$\text{T.3:} \quad \text{DET-AT}_{\text{mg}}(\text{X}, \text{Y}, \text{f}, \text{h}) \rightarrow \text{I}(\text{X}, \text{f}, \neg \text{D}'(\text{Y}, \text{h})) \& \text{B}(\text{X}, \text{W}^{00'}(\text{Y}, \text{D}'(\text{Y}, \text{h})))$$

Deterrence attempts are thus (according to T.2) attempts to make Y refrain from executing his project h (as assumed by X), and (according to T.3) attempts to prevent Y to from doing what (in X's opinion) he originally intended.

4 Consequence-based deterrence attempts

4.1 Definition.

$$\text{D4:} \quad \text{DET-AT}_{\text{g}}(\text{X}, \text{Y}, \text{f}, \text{C}, \text{h}) := \text{I}(\text{X}, \text{f}, \text{A}(\text{Y}, \text{D}(\text{X}, \text{f}), \text{C}, \text{h}))$$

$$\begin{aligned}
\text{T.4:} \quad \text{DET-AT}_{\text{g}}(\text{X}, \text{Y}, \text{f}, \text{C}, \text{h}) \leftrightarrow & \quad (\text{i}) \text{ D}(\text{X}, \text{f}) \\
& \quad (\text{ii}) \text{ W}^0(\text{X}, \neg \text{D}'(\text{Y}, \text{h})) \\
& \quad (\text{iii}) \text{ B}^0(\text{X}, \text{W}^{00'}(\text{Y}, \text{D}'(\text{Y}, \text{h}))) \\
& \quad (\text{iv}) \text{ I}(\text{X}, \text{f}, \text{B}'(\text{Y}, \text{D}'(\text{Y}, \text{h})) \gg \text{C}) \\
& \quad (\text{v}) \text{ B}^0(\text{X}, \text{W}^{00'}(\text{Y}, \neg \text{C}))
\end{aligned}$$

4.2 Other theorems.

T.5: $\text{DET-AT}_g(X, Y, f, C, h) \rightarrow \text{DET-AT}_{mg}(X, Y, f, h)$

T.6: $\text{DET-AT}_g(X, Y, f, C, h) \rightarrow I(X, f, \neg D'(Y, h))$

Special case: $C = D''(X, r)$.

5 Deterrence attempts in a narrow sense

5.0 The fact that Y doesn't want $D''(X, r)$ is usually a mediated preference. Y doesn't want it because he doesn't want the (supposed) consequence of X doing r. It is on precisely this premise that deterrence attempts in a narrow sense rely.

5.1 Definition.

D5: $\text{DET-AT}(X, Y, f, r, C, h) :=$

- (1) $I(X, f, \text{DET}(Y, T(X, f), C, h))$
- (2) $I(X, f, B^{0'}(Y, D'(Y, h)) \gg D''(X, r))$
- (3) $B^0(X, B^{0'}(Y, D''(X, r)) \gg C)$

T.7: $\text{DET-AT}(X, Y, f, r, C, h) \leftrightarrow$

- (i) $D(X, f)$
- (ii) $W^0(X, \neg D'(Y, h))$
- (iii) $B^0(X, W^{00'}(Y, D'(Y, h)))$
- (iv.a)** $I(X, f, B^{0'}(Y, D'(Y, h)) \gg D''(X, r))$
- (iv.b)** $B^0(X, B^{0'}(Y, D''(X, r)) \gg C)$
- (v) $B^0(X, W^{0'}(Y, \neg C))$

Let's have a look at an example of DET-AT_g , but not DET-AT : In order to deter shoplifters, the shop XYZ has installed an automatic balance which weighs customers when they enter and leave the shop. Customers who are heavier when they leave the shop (which does not contain a café etc.) than they were on entering it are questioned. Customers are informed about the balance by a sign when they come in, as well as of the fact (which is especially important for the shop itself!) that following the inevitable discovery of theft *no* further action will be taken. (This was in fact an almost exact quotation from an item recently heard on the news on the radio.)

The classic example of an DET-AT : h is a nuclear first strike; r is nuclear retaliation; C is a nuclear holocaust.

5.2 *Strict subjectivity*. Situation A is called (with respect to the subject X) *strictly subjective* iff A is true iff A is believed by X to be true. $B(X, A)$ and $W(X, A)$ are the paradigm cases of such situations (as $B(X, A) \leftrightarrow B(X, B(X, A))$ and $W(X, A) \leftrightarrow B(X, W(X, A))$.) As T.8 shows, the existence of an attempt at deterrence (except for the execution of the relevant action f) also constitutes a strictly subjective situation:

$$T.8: \quad \text{DET-AT}(X,Y,f,r,C,h) \leftrightarrow \quad T(X,f) \ \& \ B(X,\text{DET-AT}(X,Y,f,C,h))$$

6 Successful deterrence

6.0 Successful instrumental doing.

$$D0.1 \quad \text{IS}(X,f,A') := \quad I(X,f,A') \ \& \ (D(X,f) \gg A')$$

X's attempt by doing f to cause A' is successful iff by doing f, X intends to achieve (attempts) A', and his doing f actually brings it about that A'.

An instrumental action of X' thus is successful iff X's conviction (with respect to bringing about his aim) is proven correct. Using Max Weber's terminology: iff his action is not only "subjectively means-end-rational", but also "objectively means-end-rational". The simple occurrence of A' (perhaps brought about by something totally different) would not suffice for the success of such an action.

6.1 Successful attempts at deterrence in general.

$$D3.1 \quad \text{DET-S}_{\text{mg}}(X,Y,f,h) \quad := \quad \text{DET-AT}_{\text{mg}}(X,Y,f,h) \ \& \ \text{DET}_g(Y,T(X,f),h)$$

$$T.9: \quad \text{DET-S}_{\text{mg}}(X,Y,f,h) \quad \leftrightarrow \quad \begin{array}{l} (1) \ \text{DET-AT}_{\text{mg}}(X,Y,f,h) \\ (2) \ \neg D'(Y,h) \\ (3) \ W^{00'}(Y,D'(Y,h)) \\ (4) \ D(X,f) \gg W^{0'}(Y,\neg C) \end{array}$$

6.2 Successful C-based attempts at deterrence.

$$D4.1 \quad \text{DET-S}_g(X,Y,f,C,h) \quad := \quad \text{DET-AT}_g(X,Y,f,C,h) \ \& \ \text{DET}(Y,T(X,f),C,h)$$

$$T.10: \quad \text{DET-S}_g(X,Y,f,C,h) \quad \leftrightarrow \quad \begin{array}{l} (1) \ \text{DET-AT}_g(X,Y,f,C,h) \\ (2) \ \neg D'(Y,h) \\ (3) \ W^{00'}(Y,D'(Y,h)) \\ (4) \ D(X,f) \gg W^{0'}(Y,D'(Y,h) \gg C) \\ (5) \ W^{0'}(Y, \neg C) \end{array}$$

6.3 Successful attempts at deterrence in the narrow sense.

$$D5.1 \quad \text{DET-S}(X,Y,f,r,C,h) \quad := \quad \begin{array}{l} (1) \ \text{DET-AT}(X,Y,f,r,C,h) \\ (2) \ \neg D'(Y,h) \\ (3) \ W^{00'}(Y,D'(Y,h)) \\ (4a) \ D(X,f) \gg B^0(Y,D'(Y,h) \gg D''(X,r)) \\ (4b) \ B^0(Y,D''(X,r) \gg C) \\ (5) \ W^{0'}(Y, \neg C) \end{array}$$

$$T.11 \quad \text{DET-S}(X,Y,f,r,C,h) \quad \leftrightarrow \quad \text{DET-AT}(X,Y,f,r,C,h) \ \& \ \text{DET}(Y,T(X,f),C,h) \ \& \ (4a) \ \& \ (4b)$$

the deterrer's attempts at deterrence, in his view (at least at the time of his deterrence attempt), deterrence success coincides with the non-occurrence of deterrence failure:

T.A12: $DET-AT(X,Y,f,h) \rightarrow B(X, \neg DET-AT-F(X,Y,f,h) \equiv DET-S(X,Y,f,h))$

In the deterrer's view, the absence of deterrence failure and deterrence success boil down to the same thing.

Treachurous linguistic usage. This point is also important for our sometimes lackadaisical talk of the alleged success of one type of deterrence or another. For example, we are often told that the strategy of mutual nuclear deterrence has so far worked, and all in all has thus been successful. Do we actually realise what this assumes? The weaker (and yet probably more accurate) claim would be that the strategy of deterrence has so far not obviously miscarried – and is thus not a failure. Yet there is still a rather large gap between the absence of deterrence failure and success, as we shall see below.

8 Not-successful deterrence – and its reasons

8.1 *Practical relevance.* Sufficient reasons for not-successful deterrence are extremely relevant for both the deterrence subject and for their (potentially regarding themselves as such) deterrence addressee. Before launching an attempt at deterrence, a rational (in a narrow sense) deterrer will try to find out whether any of these reasons (which would preclude the success of his attempt) exist – and if necessary precede his attempt by doing everything to make sure that none of these reasons prevails. And those who don't want to be deterred by somebody else know what action must be taken: they merely have to make the potential deterrer believe them to be someone for whom at least one of the reasons deemed sufficient to ensure not-success (therefore above all not the right preference) applies – which in turn of course will motivate the rational deterrer to do everything to ensure that this in fact is not the case, etc.

8.2 *Non-(2): Failure.* In 7.3f above I mentioned the most trivial case of not-successful deterrence: Y does what the attempt at deterrence was actually supposed to prevent. The attempt has failed. And thus_R (R = in the rationality case assumed here!) the following holds: The attempt at deterrence has not brought about the intended change of preference in Y. Instead, Y has done what he wanted to do.

In the following I assume (at least in the beginning) that no such failure has occurred. The negation of the DET-S-conditions (3) to (5) – and their combinations (ignored here) – are thus the only reasons sufficient for the absence of success of deterrence.

8.3 *Non-(3):* The assumption made by X that Y *wanted to perform* the action whose omission constituted the primary deterrence aim of his attempt at deterrence before this attempt was launched is incorrect. It is simply not the case that Y wanted to do h. And so_R it is simply not the case either that Y would have done h if the attempt at deterrence hadn't been made. Y wouldn't have done h in the first place! Hence the attempt at deterrence was quite simply unnecessary.

More detailed examination of non-(3) entails distinguishing between two of its special cases:

- non-(3): $\neg W^{00'}(Y, D'(Y, h))$
 (I) $W^{00'}(Y, \neg D'(Y, h))$
 (II) $\neg W^{00'}(Y, D'(Y, h)) \& \neg W^{00'}(Y, \neg D'(Y, h))$

In (I) Y wanted from the outset precisely what X wanted of him, namely not to do h. He could_R not therefore also at the same time want to do h. (Non-(3) follows from (I).) In (II) Y was indifferent to doing or not doing h (at t').

It is clear what a rational party who doesn't want to be deterred by X with respect to h will ceteris paribus do: he will try to make X believe that (I) is already the case and that no deterrence or other coercion is required in order to make him not do f – because he has no intention of doing f anyway (at any rate, this is what X is supposed to believe).

8.4 *Non-(4)*: It is not correct that, to start with the general case of *simple* consequence-based deterrence, only and precisely X's doing f *caused* Y to believe that his doing h would have the consequence C. And as by definition the case (i) $\neg D(X, f)$ is precluded in this attempt at deterrence, this could again mean the following – where for Y's relevant consequence belief which X intends to bring about by doing f (i.e. for $B^{0'}(Y, D'(Y, h) \gg C$)), I will write in short $B^{0'}C$.

- (ii) $D(X, f) \& \neg B^{0'}C$
 (iii.1.1) $D(X, f) \& \neg B^{00'}C \& B^{0'}C \& (\neg D(X, f) \supset B^{0'}C)$ ⁴
 (iii.2.1) $D(X, f) \& B^{00'}C \& B^{0'}C \& (\neg D(X, f) \supset B^{0'}C)$

(ii) is again the *obvious* case of not having effected what X had desired to effect. However, this is also true of the other two cases, where Y's C-belief would have existed without the influence of $D(X, f)$. Once again, X's attempt to bring something about was unnecessary.

Non-(4a) and/or non-(4b) are special cases which I shall not go into here.

8.5 *Non-(5)*: It is not the case that Y (at t^{0'}) wants not-C. Once again we must distinguish here between the following:

- non-(5): $\neg W^{0'}(Y, \neg C)$
 (I) $W^{0'}(Y, C)$
 (II) $\neg W^{0'}(Y, \neg C) \& \neg W^{0'}(Y, C)$

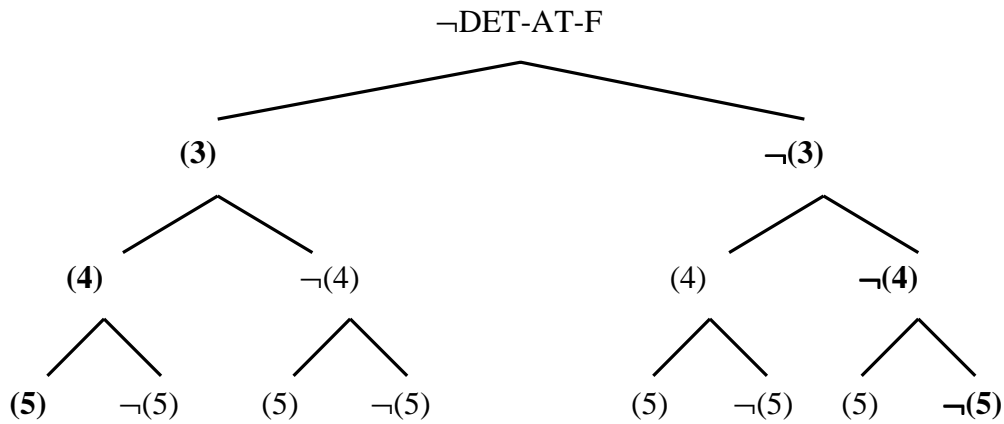
Here too, it is evident how one can best be forearmed against being the addressee of an attempt at deterrence undertaken by reference to (or threatening) C: let the potential deterrence subject know (or at least make him believe) that C is not at all a frightening prospect – quite simply because it's what the addressee wants! In other words: (I). The indifference case (II) has the same, albeit weaker, drift.

8.6 *Taken separately – and then together*. Of course, the reasons sufficient for not-successful deterrence previously regarded separately could also occur together. What's more, owing to their mutual logical independence, they could occur in (almost) any combination. The description and discussion of such combinations verified with examples would be quite an

⁴ Of course instead of the simple implications in this context conditional concepts should be used.

exciting affair requiring not only logic but a large helping of imagination. Here is just a rough map giving us the required survey of the various possible combinations.

Reasons for: Not failure, but not success either



The map illustrates the case of *simple* consequence-based attempts at deterrence. As the dichotomic differences occur on three levels, $2^3 = 8$ cases would need to be considered. The same map should now be placed below the failure case, for this too is ultimately a (and indeed *the* most obvious) reason for an attempt at deterrence not being successful. This doubles the number of possibilities, bringing us to 16. And the above map doesn't yet take into account the complications caused by the sub-routes (4a) and (4b), which in each of the two $D'(Y,h)$ vs. non- $D'(Y,h)$ cases brings about $2^4 = 16$ case distinctions, or 32 in total.

Finally, the maps used so far do not yet take into account either the other sufficient reasons distinguished above in 8.2–8.5, from which the various necessary definition conditions for deterrence success may for their part be incorrect.⁵

8.7 *Two logically 'impossible' extreme cases.* Discussing all these special cases resulting from the various combinations would be impossible here. But let us at least briefly deal with two cases, namely the two **extreme cases** (shown on the map in **bold type**) in which (Case 1 = branch on the far left) with the exception of condition (2) (= no failure) *all* other necessary conditions *for deterrence success are met* or (Case 2 = branch on the far right) with the exception of condition (1) – i.e. $DET-AT(X,Y,f,C,h)$ – and (2) *all* such conditions *are not* met.

Both these borderline cases raise a problem which is both fundamental and once again fraught with consequences for deterrence practice. These two cases are *not* – despite the impression given above – *freely combinable* with both $D'(Y,h)$ and $\neg D'(Y,h)$. For both:

(Case 1*) (3) **$W^{00r}(Y,D'(Y,h))$** &

⁵ For the relevant logical relations in this respect see my overview contained in §3 of *Gemeinsamer Glaube und Gemeinsames Wissen* (in: *Neue Realitäten – Herausforderung der Philosophie*, published by the AGPD, Berlin, 1993, pp. 761–767). The principles listed there for interpersonal belief also generally apply to interpersonal attitudes.

(4) $DX,f \gg B'(Y,D'(Y,h) \gg C) \&$

(5) $W'(Y,\neg C) \&$

But nevertheless: $\neg(2)$, i.e. $D'(Y,h)$ (failure)

and

(Case 2*) $\neg(3) \quad \neg W^{0'}(Y,T'(Y,h)) \&$

$\neg(4) \quad \neg(D(X,f) \gg B'(Y,D'(Y,h) \gg C)) \&$

$\neg(5) \quad \neg W^{0'}(Y,\neg C) \&$

But nevertheless: (2), i.e. $\neg D(Y,h)$ (absence of failure)

are 'logically impossible': The *nevertheless* condition in each case contradicts the rest of the respective (case*).

*Case 1**: The whole deterrence concept (explained here) is based on the fact that a certain act can be prevented by means of a change of preference brought about. The expectation of success of this attempt necessarily linked to every attempt at deterrence is based on the expectation that the deterrence addressee will firstly be dissuaded from his plan to do h (or his indifference in this respect) – in other words that he will *no longer want to do h*, and secondly will *also behave accordingly*, i.e. he really *won't do h*.

However, precisely this is contradicted by Case 1*. Why? Because:

(R1.a) (3) to (5) $\rightarrow \neg W^{01}(Y,D'(Y,h))$

(R1.b) $\neg W^{01}(Y,D'(Y,h)) \rightarrow \neg D'(Y,h)$

And thus also:

(R1) ((3) to (5)) $\rightarrow \neg D'(Y,h)$

– which clearly **contradicts the "nevertheless $D'(Y,h)$ "** from *Case 1**.

Case 2* looks like this:

(R2.a) $\neg(3)\&\neg(4)\&\neg(5) \rightarrow P^{0'}(Y,D'(Y,h))$

(R2.b) $P^{0'}(Y,D'(Y,h)) \rightarrow D'(Y,h)$

And thus also:

(R2) $\neg(3)\&\neg(4)\&\neg(5) \rightarrow D'(Y,h)$

– which clearly **contradicts the "nevertheless $D'(Y,h)$ "** from *Case 2**.

8.8 *Logically impossible – but possible*. This contradiction harbours a problem which is easy to solve. Although deductions (R1) and (R2) and their premises are indeed logically valid sentences, they only apply for *rational players*, i.e. for players whose believing, wanting and doing are rational in the sense of the rationality postulates always presupposed here (for B, W and D). But the thing is, we *are not* always players of this type. And there is no logic that demands we should be such. In short: what's impossible for rational players may be possible for irrational ones.

8.9 *No practical conclusion.* Let's go back to 8.6. The extent to which the extreme cases described there are 'possible' is now evident – namely in so far as not the correct practical rationality conclusion (with the suitable action or omission conclusion) is drawn from the routes (premises (3) to (5) and their negations), in other words the relevant (in rational terms incorrect and thus irrational) "doing" is not acting, but instead merely a behaviour. This possibility was therefore assumed during the above passage on the complete combinability of all map points.

8.10 *Rationality assumption.* The expectation involved in every attempt at deterrence of being able to cause the omission of a certain action on the part of the deterrence addressee by inducing a change of preference assumes that such a change of preference will actually lead to a corresponding consequence (in this case an omission consequence), and that the addressee is rational in the sense of compliance with the principle of rationality:

(RP*) $W^0(a, D(a, h)) \rightarrow D(a, h)$, if POSS($D(a, h)$)
If a wants to do h and it is possible for him to do h, he will do it.

Otherwise (including in X's view) the whole attempt at deterrence would be pointless. If the addressee is irrational (in the sense specified), no matter how many other conditions for deterrence success prevail, he will still not display behaviour which judged accordingly is deemed rational. His irrational conduct turns the attempt at deterrence into a failure.

(DET-RAT) Attempts at deterrence assume the rationality of the party to be deterred. Such attempts can only be successful, if this assumption applies.

8.11 *Practical relevance.* (DET-RAT) has far-reaching consequences for deterrence practice. For the perfect way of defusing deterrence attempts becomes obvious – make sure the potential deterrence subject thinks you're *irrational*, at least in the relevant respects.

Attempts at deterrence assume the rationality of the party to be deterred in several respects. Hence there are also several ways to attempt irrationality immunisation. To mention the most important one here, the potential addressee makes sure that the potential deterrer X believes him to be someone for whom at the relevant time the rationality principle (RP*) doesn't even apply (assuming the possibility of performing the action h at the relevant time).

It's not difficult to guess the best way of achieving this. Just make sure that X believes you to be somebody who will violate this principle not only at the time concerned. An even better approach (because it's worse): Make sure that you are in general considered *mad* – not only by X, and with no time constraint. And the most convincing way to achieving this is unfortunately only too well-known – namely actually *become* mad. (It's a trick which has already been done a few times in history.)

In this paper I have only scratched the tip of the iceberg. It could well be a long time before the entire iceberg is in view – not to mention neatly mapped out and packed up in the correct ideas on rationality and moral. Yet analytical thinkers shouldn't be deterred by the sheer size of this research project, especially as in view of their other claims to clarity they've probably got a lot of catching-up to do in this momentous topic.