CHAPTER V.

PRACTICAL RESPONSIBILITY III:

REASONABLE$_A$ NORMATIVE EXPECTATIONS

In Chapter IV, I have distinguished two senses in which normative expectations can be reasonable and answered some preliminary questions about the concept. In view of the fact that the concept of reasonable$_A$ normative expectations will be crucial to the account in Chapter VI, we need to dispense with the guiding metaphor of what is "within the agent's power" and offer a systematic account of reasonableness$_A$. Section 1 proposes an account of prima facie reasonableness. Sections 2 and 4 develop the concept of a defeating condition, taking care to avoid the fundamental problem.

1. When Are Normative Expectations Prima Facie Reasonable$_A$?

Thus far, the only restriction I have placed on normative expectations is that they be practical. This is to say, when we expect of someone that he bring it about that *p*, '*p*' must be logically and physically contingent. However, there are examples of expectations whose results are contingent and which we would judge intuitively unreasonable$_A$. For example, we would think that it is unreasonable$_A$ to expect of a person that he speak all the known languages fluently, but his speaking so many languages is not impossible. We would also think that it is unreasonable$_A$ to expect of a person that she win a lottery, but it is not impossible for her to win it. By a similar token, we would think it unreasonable$_A$ to expect of a person that she breathe, but it is certainly not necessary that she does. It would be unreasonable$_A$ to expect of a person that she bring it about that the seasons change, yet it is not necessary that they do. These are all

examples of performances that we would think are not "within the agent's power." We now need to dispense with the metaphor.

In this section, I want to begin by characterizing the notion of prima facie reasonableness$_A$. Prima facie reasonableness$_A$ is meant to capture the idea of what is "within our power" (as humans, say) to do. In particular, it abstracts from any special considerations the agent deserves in virtue of her particular circumstances. We will take the special circumstances into account when discussing the defeating conditions in sections 2 and 4. For example, the expectation to tell colors apart is prima facie reasonable$_A$, for it is something that is in general "in our power" to do. However, the prima facie reasonableness$_A$ of such an expectation is defeated, if the agent whom we hold to the expectation is color-blind.

I suggest that we ought to understand prima facie reasonableness$_A$ negatively, viz. in terms of what is *not* prima facie *un*reasonable$_A$ (section A). There are two kinds of situations in which an expectation is prima facie unreasonable$_A$: first, when it would be systematically frustrated by most agents in most circumstances; second, when it would be systematically fulfilled while its contrary is systematically frustrated by most agents in most circumstances. A concept that is crucial in this characterization is that of a systematic correlation. I will treat it as a theoretical place-holder and not give an account of it, but I will say a few words about it in section B.

I will speak of an expectation *to* φ (rather than an expectation *of* α *that* α φ) as being systematically fulfilled or frustrated or neither. Similarly, I will speak of certain conditions (defeating conditions) being systematically correlated with the fulfillment or frustration of an expectation *to* φ (rather than an expectation *of* α *that* α φ). I will use this manner of speaking in order to emphasize that the systematic correlations at stake hold irrespectively of the particular agent who is held to the expectation on a particular occasion.

A. Prima Facie Reasonableness$_A$

There are at least two ingredients in the metaphor of a performance being "within an agent's power." First, there is a sense in which the agent must be able to perform the action in question. If the agent could not succeed in performing the action, we would

intuitively think that the action was not "within the agent's power" at the time. An expectation of a two-year old child to win an Olympic swimming competition would surely be unreasonable$_A$. Second, there is a sense in which the agent must be able to make a difference. If what the agent is about to do would happen whether or not the agent did anything, we would be inclined not to think that what happened was in the agent's power.

Consider three types of cases where it would be intuitive to say that it would be unreasonable$_A$ to expect an agent to perform an action. It would be unreasonable$_A$ to expect of someone that he win an (unrigged) lottery. Winning the lottery is not something that is "up to him," that is "within his power" — it is almost certain that he will lose. It would also be unreasonable$_A$ to expect of a person that she breathe,[1] or that she make her heart beat. Breathing and having one's heart beat are not "within the agent's power" — it is something that happens no matter what the agent does. Finally, it would be unreasonable$_A$ to expect of the agent that he throw a coin so it comes up heads. Unlike the first case, the coin will not almost certainly come up tails; unlike the second case, the coin will not almost certainly come up heads. Yet, the coin's coming up heads is not something the agent controls. We can capture these three kinds of cases using the following test.

Let us begin with a deceptively simple scenario, which will suggest the gist of the test. Let us imagine that we want to test whether an agent can perform a certain type of action. To do so, we will give him a series of tasks, to which he will respond in the best possible way: we are assuming, in other words, that he is cooperative, that there are no other designs, intentions, expectations in play, the agent is at ease, under no pressure, etc.[2] The tasks are of two kinds, to φ and not to φ, and they are interspersed randomly in a series.

Four situations are of special interest. Suppose that an agent systematically frustrates the expectation to φ (situations (iii) and (iv) in Table 1). When he is expected

---

[1] It might be reasonable$_A$ to expect of a person to take a breath at a particular moment, or stop breathing for a couple of seconds, but not to stop breathing altogether or breathe at all.

to φ, he does not. In such a case, it would be unreasonable$_A$ to expect of the agent that he φ. The agent cannot succeed in fulfilling the expectation. Suppose that the agent regularly fulfills the expectation to φ but frustrates the expectation not to φ (ii). What this will mean is that the agent φs indiscriminately. In such a case, we would tend to think that the agent's φing is not up to him, that the agent cannot make a difference, and hence that it would be unreasonable$_A$ to expect of him that he φ. This configuration would obtain if we expected the agent to breathe, for example. Finally (i), when the agent fulfills all the expectations (when expected to φ, the agent responds by φing, when expected not to φ, the agent responds by not φing), we would tend to think that φing and not φing are "within the agent's power," that it is not unreasonable$_A$ to expect of the agent that he φ.

	Task: φ	Task: not-φ
(i)	fulfilled (φ)	fulfilled (not-φ)
(ii)	fulfilled (φ)	frustrated (φ)
(iii)	frustrated (not-φ)	fulfilled (not-φ)
(iv)	frustrated (not-φ)	frustrated (φ)

Table 1. Possible result patterns of a simplified test sequence.

It may be worthwhile noting that there is an interesting difference between situations (iii) and (iv). Situation (iii) is analogical to situation (ii). When the agent systematically frustrates the expectation to φ but fulfills the expectation not to φ, the agent simply does not φ. Once again, it would be unreasonable$_A$ to expect him to φ (or not to φ). We would judge that his not-φing was not up to him. This case corresponds to what would happen were the agent expected to win the lottery, for example. That expectations would be systematically frustrated, while its contrary would be systematically fulfilled. Situation (iv) is different, however. Here the agent is counter-

[2] This is an unrealistic assumption. I am making it in order to sharpen the intuitions at stake. The account I am proposing does not depend on it, however.

competent, as it were. When expected to φ, he does not φ. When expected not to φ, he does φ. The case is somewhat curious. The most natural way of thinking about it is that the agent acts contrary to the expectations. But this contradicts our simplifying assumption that the agent is cooperative, under no other pressures, etc. At the same time, it would seem an altogether implausible accident of nature that despite being in the best possible conditions, the agent always frustrates the expectations to which he is held. However curious the case is in its pure form (given the simplifying assumptions), there are cases that appear to approximate it. Take the pair of expectations to throw a coin so that it comes up heads and to throw it so that it comes up tails. It is certainly not true that anyone of us is counter-competent with respect to throwing a fair coin. It is true, however, that most of the expectations in a random series would be frustrated in the long run.

This simplified test scenario allows to make a little clearer some of our intuitions concerning especially the situations in which it would be unreasonable$_A$ to hold an agent to an expectation. It will be immediately objected, however, that the test scenario is unrealistic. It presupposes that the agent responds to the expectation in the very best conditions. But such conditions are almost never present. And even if they were, it is not clear that we could count on them in giving an account of unreasonableness$_A$. I do not believe that we have to rely on such strict test conditions. Rather, the way in which we gather the knowledge concerning unreasonable$_A$ expectations and conditions that make expectations unreasonable$_A$ (defeating conditions) takes account of our general knowledge (cutting across times, places, particular agents) concerning the way in which most agents behave. Rather than requiring that the agent fulfill or frustrate an expectation, we might require that most agents across a wide range of circumstances that approximate the ideal test conditions systematically fulfill or frustrate the expectation. (I will say a little more about the concept of a systematic correlation in the next section.) In this way, the special circumstances will tend to be evened out, as it were. For instance, when subjecting a particular agent to such a test, we might worry about what social scientists worry about when testing humans, viz. that the individual's responses will be changed by the very fact that they are taking part in an artificial test situation. In the simplest case, the individual might not be responding in the best possible way to the task,

but might be doing the opposite on purpose, say. These kinds of peculiarities will tend to disappear if we collect a large number of data cutting across a variety of settings.

We can dress these intuitions thus:

(Success Condition):

> It is prima facie unreasonable$_A$ to hold α to an expectation to φ if the expectation to φ is systematically pf-frustrated.[3]

(Difference Condition):

> It is prima facie unreasonable$_A$ to hold α to an expectation to φ if the expectation to φ is systematically pf-fulfilled while the expectation not to φ is systematically pf-frustrated.[4]

The conditions allow us to understand why non-practical expectations are prima facie unreasonable$_A$. The expectation to see to it that 2+2=3 would be systematically pf-frustrated and so unreasonable$_A$ in virtue of the success condition. Likewise, the expectation to see to it that 2+2=4 would be unreasonable$_A$ in virtue of the difference condition: it would be systematically pf-fulfilled while the contrary expectation (to see to it that 2+2≠4) would be systematically pf-frustrated. The expectation to see to it that the e-mail goes through faster than light would be systematically pf-frustrated, while the expectation that the Earth move around its axis would be systematically pf-fulfilled, while the contrary expectation systematically pf-frustrated.

But the concept of prima facie reasonableness$_A$ can discriminate further than cases of non-practical expectations. An expectation to win a fair lottery would be prima facie unreasonable$_A$. Such an expectation would surely be systematically pf-frustrated. An

---

[3] In Chapter III, I have distinguished between prima facie and agentive fulfillment (frustration, respectively) of expectations. An expectation to φ is agentively fulfilled only by *actions* of φing (raising the arm), while it is prima facie fulfilled by performances, whether they be actions or nonactions (raisings and risings of the arm). I use 'pf-' to mark the prima facie sense of fulfillment (frustration, respectively). I will discuss the importance of this constraint in section 2.

[4] The success and difference conditions roughly correspond to what Belnap calls the positive and the negative condition of agency ("Before Refraining: Concepts for Agency," *Erkenntnis* 34, 1991, 137-169; see also Belnap and Perloff, "Seeing to It that," *op.cit.*). Unlike the positive condition, the success condition does not require that the success be guaranteed. The difference condition excludes the situations where the agent cannot make a difference but without committing us to incompatibilism.

expectation to speak all the known languages fluently would be systematically pf-frustrated, and so is prima facie unreasonable$_A$. By contrast, an expectation to breathe would be systematically pf-fulfilled while its contrary would be systematically pf-frustrated, and thus prima facie unreasonable$_A$. It would also be prima facie unreasonable$_A$ to expect of a person that she bring it about that the seasons change, for such an expectation would be systematically pf-fulfilled, while its contrary would be systematically pf-fulfilled.

We will work under the hypothesis that no other conditions characterize prima facie unreasonableness$_A$. We can then define reasonableness$_A$ negatively:

(R) an expectation is *prima facie reasonable$_A$* iff it is not prima facie unreasonable$_A$.

As agents, we are guilty until proven innocent.[5] It is reasonable$_A$ to expect of us any performance unless there are special conditions that would make such an expectation unreasonable$_A$. The concept of reasonableness$_A$ is thus characterized negatively in terms of what it is not unreasonable$_A$ to expect of an agent.[6]

We should be clear that the concept of prima facie reasonableness$_A$ is not yet a concept that would be sufficient to capture our intuitions concerning what it would be (all-out) reasonable$_A$ to expect of a particular agent. It is surely not reasonable$_A$ to expect of a student who has been taken seriously ill that he turn in homework on time, yet such an expectation would be prima facie reasonable$_A$. It is unreasonable$_A$ to expect a blind person to read aloud, but such an expectation is prima facie reasonable$_A$. It would be unreasonable$_A$ to expect of a newly arrived foreigner that he speak like a native but such an expectation is prima facie reasonable$_A$. It is clear that we need to take the special circumstances in which the agent finds herself into account. This is the role of defeating conditions.

---

[5] A similar principle concerning moral responsibility is defended in Keith D. Wyma, "Moral Responsibility and Leeway for Action," *American Philosophical Quarterly* 34 (1997), 57-70.

[6] This is also the deep reason why the concept of reasonableness$_A$ does not admit an intermediate category of performances that are neither reasonable$_A$ nor unreasonable$_A$. We will remember that the concept of reasonableness$_N$ does admit of such performances (see Chapter IV).

B. Systematic Correlations

As we saw in the last section, one of the concerns is that the agent might be uncooperative, intent on acting contrary to the expectations, etc. It is for this reason that the simplified test discussed there relied on assuming that the agent finds himself in ideal conditions (that he is cooperative, under no pressures, that no other intentions or expectations are in play, etc.). These conditions are never or rarely actually satisfied but they can be approximated.

We might imagine an expector, who chooses agents, times, occasions at random, and subjects the agents to the expectation that they φ. She will exclude those who are clearly uncooperative, who are under special pressures, or where she suspects other expectations to be involved. Given this large set of data, she can then decide that an expectation to φ is systematically frustrated if most agents, most of the time, have frustrated the expectation in question; or that the expectation to φ is systematically fulfilled if most agents, most of the time, have fulfilled the expectation in question.[7] Similarly, given large amounts of data, such an expector can tell whether a particular type of event, *C*, is systematically correlated with the fulfillment or frustration of an expectation to φ. A *C*-type event will be systematically correlated with the fulfillment (frustration) of an expectation to φ if, given the occurrence of events of type *C*, the expectation to φ is systematically fulfilled (frustrated) ceteris paribus.

Although we do not have access to such a large set of data, we do have access to hypotheses and theories concerning the mechanisms involved in the fulfillment and frustration of expectations. Thus, we would consider it quite intuitive to think that being in a coma is systematically correlated with the frustration of the expectation to talk, to smile, etc., and with the fulfillment of the expectation to lie motionless. Our judgment is affirmed not only by the preponderance of data but also by our understanding of the causal processes involved in a coma.

---

[7] By not insisting that the quantifiers be universal, we can further take into account cases where the ideal conditions have been violated.

Many of the systematic correlations (especially those involving defeating conditions) will be causal in nature. In fact, when such correlations are causal, and when we understand the causal mechanisms behind them, we are most confident of the correlation. However, it would not be wise to exclude the possibility of "merely statistical" correlations. This aggravates the issue of the vagueness of the concept of systematic correlation. The notion appeals to a vague quantifier 'most'. It is not clear furthermore that it could be made any more precise without introducing ad hoc arbitrariness. While this is certainly the case, it is not clear that we should seek any more precision for the purposes at hand. For one thing, this issue is not peculiar to the domain of agency. It is a more general problem confronted on a daily basis in scientific testing, for example. For another, it would be unwise to exclude the possibility that the vagueness is a part of the concept itself and that there will be gray cases where it will be simply unclear whether a systematic correlation is in place or not.

Another problem with cashing out the concept of a "systematic correlation" consists in the fact that any such attempt will involve an appeal to ceteris paribus conditions. Drinking great quantities of coffee may be systematically correlated with extreme agitation. But not if one is a "caffeine addict." If one consumes large amounts of caffeine in the first place, one's reaction will be very different. The original correlation holds only ceteris paribus. Once again, however, this fact does not present any special problem for action theory. It is a general problem not only for scientific[8] but also for most ordinary claims we make.[9]

2. Defeating Conditions

Defeating conditions make an otherwise reasonable$_A$ expectation unreasonable$_A$, or an otherwise unreasonable$_A$ expectation reasonable$_A$. We can speak of *defeating*

---

[8] Nancy Cartwright, *How the Laws of Physics Lie* (Oxford: Clarendon Press, 1983); Carl G. Hempel, "Provisos," in (eds.) Adolf Grunbaum, Wesley C. Salmon, *The Limitations of Deductivism* (Berkeley: University of California Press, 1988), pp. 3-22; Marc B. Lange, *The Design of Scientific Practice. A Study of Physical Laws and Inductive Reasoning* (Ph.D. Dissertation: University of Pittsburgh, 1990); Leszek Nowak, *The Structure of Idealization* (Dordrecht/Boston: Reidel, 1980).

[9] Robert Brandom, *Making It Explicit* (Cambridge: Harvard University Press, 1994). Nicholas Rescher, *Standardism,* forthcoming.

conditions proper in the former case (section A) and of *counterdefeating conditions* in the latter (section B). In section C, I consider the way in which the fundamental problem bears on the account of reasonableness$_A$. Finally, in section D I ask whether we should think of defeating conditions as causes.

A. Defeating Conditions Proper

The expectation that a student turn in homework on time is prima facie reasonable$_A$. The expectation is not systematically frustrated, nor is its contrary. But when the student falls seriously ill, its reasonableness$_A$ is defeated. The expectation that a person run in a race is prima facie reasonable$_A$. But it is no longer reasonable$_A$ if she has broken a leg. The expectation that a person walk straight is reasonable$_A$ but not when he has been pushed by another. These are examples of what I will call defeating conditions of the first kind, or hindering conditions.[10] They can be understood on the lines suggested above:

(1) An event of type *C* is a *defeating condition of the first kind (hindering condition)* with respect to an expectation to φ iff the occurrence of an event of type *C* is systematically correlated with the pf-frustration of the expectation to φ and with the pf-fulfillment of the expectation not to φ.

Breaking a leg is systematically correlated with the pf-frustration to run a race and with the pf-fulfillment of the expectation not to run a race. It is a defeating condition with respect to the expectation to run the race. Thus, while it may be prima facie reasonable$_A$ to expect of a person that she run the race, in view of the fact that she has broken a leg, it would be unreasonable$_A$ to hold her to the expectation. Being seriously ill is systematically correlated with the pf-frustration of the expectation to turn in homework on time. In view of the fact that a student has fallen seriously ill, it would be unreasonable$_A$ to expect him to turn in the homework. Being pushed is systematically correlated with the pf-frustration of an expectation to walk straight, hence it would be

unreasonable$_A$ to expect of an agent who has been just pushed by another that he walk in a balanced way. These and others are examples of defeating conditions of the first kind. Suffering a spasm in one's arm is systematically correlated with the frustration of various kinds of expectations having to do with the control over one's arm. Not knowing that one is to be present at a certain meeting is systematically correlated with the frustration of the expectation to be at the meeting. Being in a coma is systematically correlated with the frustration of a great many expectations. Not having access to the right equipment is systematically correlated with the pf-frustration of the expectation to build a bridge. And so on.

It needs to be emphasized that the concept of a defeating condition is relativized to an expectation. An event-type that may be a defeating condition with respect to one expectation need not be a defeating condition with respect to another. Breaking a leg makes the expectation to run a race unreasonable$_A$, but it does not defeat the reasonableness$_A$ of the expectation to remember your friend's birthday. When an agent suffers from a tic it is unreasonable$_A$ to expect of him that he wink three times, but it may still be reasonable$_A$ to expect him to do the fox-trott.

The second kind of defeating condition corresponds to the difference condition rather than the success condition.

(2) An event of type *C* is a *defeating condition of the second kind (compelling or forcing condition)* with respect to an expectation to φ iff the occurrence of an event of type *C* is systematically correlated with the pf-fulfillment of the expectation to φ and with the pf-frustration of the expectation not to φ.

It is prima facie reasonable$_A$ to expect of a person that he walk. But this expectation ceases to be reasonable$_A$ if the person is in fact physically forced to walk by another. The application of appropriate physical force is systematically correlated with the pf-fulfillment of the expectation to walk and with the pf-frustration of the expectation not to

---

[10] The terminology is based on von Wright's nice distinction between hindering (preventing) and compelling (forcing) acts (*Norm and Action* [London: Routledge & Kegan Paul, 1963], pp. 54-55).

walk. Breaking a leg is systematically correlated with the pf-fulfillment of an expectation not to take part in a race, and with the pf-frustration of the expectation to take part in a race, so breaking a leg counts as a defeating condition (of the second kind) for the expectation not to take part in race.

Finally, I want to mention defeating conditions of a third kind, to which I have already alluded in discussing the concept of systematic correlation. Suppose that an agent's hands tremble erratically, severely impairing his job manufacturing electronic chips, say. However, despite the tremble his chances of succeeding are about 50%. In other words, given an intuitive grasp of 'systematic correlation', it would be wrong to say that the condition is systematically correlated either with the frustration or with the fulfillment of the expectation to connect the chip. Yet, it is true that neither the manufacturer of the chips (expecting the agent to connect the chips correctly) nor the rival manufacturer (expecting the agent to sabotage the chip production) can count on the agent. In such a case, our intuitive judgment that what is within the agent's power is limited can be manifested if we subjected the agent to a test. Suppose that the agent was to fulfill a series of expectations to connect the chips correctly and to connect the chips incorrectly, in a random order. In the long run, it would become evident that although the agent does occasionally fulfill the expectations, he systematically frustrates most of them.

(3) An event of type *C* is a *defeating condition of the third kind* with respect to a pair of expectations to φ and not to φ iff the occurrence of an event of type *C* is systematically correlated with the pf-frustration of expectations to φ and not to φ (in a random series).

We can offer a preliminary characterization of reasonableness$_A$.

It is reasonable$_A$ to expect of α that α φ if no defeating condition with respect to the expectation to φ occurred.[11]

---

[11] Note that the characterization appears to miss cases where no defeating conditions occur but the expectation is prima facie unreasonable$_A$ (the expectation to make sure that 2+2=3, e.g.). For simplicity, I will treat the case of prima facie reasonableness$_A$ and prima facie unreasonableness$_A$ as relative to a special tautologous defeating condition.

We should note that given the characterization of defeating conditions of the first and the second kind, if *d* is systematically correlated with pf-frustration of the expectation to φ and with the pf-fulfillment of the expectation not to φ then *d* is systematically correlated with pf-fulfillment of the expectation not to φ and with the pf-frustration of the expectation to φ. This is to say that a defeating condition of the first kind is a defeating condition of the second kind for the contrary expectation. For example, lack of shooting skills is systematically correlated with the pf-frustration of the expectation to shoot the bulls-eye (and with the pf-fulfillment of the expectation not to shoot the bulls-eye). It is eo ipso systematically correlated with the pf-fulfillment of the expectation not to shoot the bulls-eye and with the pf-frustration of the expectation to shoot the bulls-eye. Believing that one's meeting is at 9am is systematically correlated with the pf-frustration of the expectation to be at the meeting at 8am. It is eo ipso systematically correlated with the fulfillment of the expectation not to be at the 8am meeting.

It should be emphasized that reasonableness$_A$ is relative to the way in which the performances are described. Consider an example. Suppose that Tamara has lost control over some of her fingers. She can move her index finger at will. She can also move her middle finger without problems. But she cannot move the remaining fingers at all. Given her condition, it would be, among other things: unreasonable$_A$ to expect of her that she move her thumb, and reasonable$_A$ to expect of her that she move her index finger. Would it be reasonable$_A$ to expect of her that she *not move* her thumb? It seems clear that the answer must be that it would be unreasonable$_A$ to expect of her that she not move her thumb. Her condition is systematically correlated with the pf-fulfillment of that expectation (and with the pf-frustration of the expectation to move her thumb). However, her moving her index finger is a way of not moving her thumb. So, it might appear problematic that though the expectation to move her index finger is reasonable$_A$, the expectation not to move her thumb is not. This impression disperses, however, in view of the fact that reasonableness$_A$ of expectations is sensitive to description. One and the same performance may be reasonable$_A$ under some descriptions but not under others.

B. Counterdefeating Conditions

So far we have considered conditions that render prima facie reasonable$_A$ expectations unreasonable$_A$. I would like to briefly mention a class of counterdefeating conditions which render prima facie unreasonable$_A$ expectations reasonable$_A$. Once again the class includes varied conditions. A large portion of it is occupied by special abilities possibly due to special equipment. It is prima facie unreasonable$_A$ to expect of a person that she perform a pirouette. But it would be reasonable$_A$ to hold a skilled skater to the expectation.[12] The reason why the expectation to perform a pirouette is prima facie unreasonable$_A$ is that such an expectation would be systematically pf-frustrated by most people. However, the expectation would not be systematically pf-frustrated by skilled skaters. Having a leg amputated will usually make the expectation to walk without support unreasonable$_A$. It will count as a defeating condition of the first kind: it will lead to the systematic frustration of the expectation. However, when the agent is equipped with a prosthesis the expectation would no longer be systematically frustrated.

We can amend our preliminary characterization of reasonableness$_A$.

> It is reasonable$_A$ to expect of α that α φ if either (a) no defeating condition (with respect to the expectation to φ) occurred, or (b) such a defeating condition did occur but it was countered by an appropriate counterdefeating condition.

C. The Fundamental Problem and the Evolution of Defeating Conditions

Before going on, we should consider the fundamental problem once again. The concern with the sort of account I am proposing is that it reverses the natural order of the concepts of action and responsibility. This is also evident here.

The concept of reasonable$_A$ expectations, or more precisely the notion of what it is reasonable$_A$ to expect of an agent, is to give us a way of understanding the concept of

---

[12] The example, together with the observation, is borrowed from Annette Baier ("The Search for Basic Actions," *American Philosophical Quarterly* 8, 1971, p. 164).

action. In order for thus developed account not to be circular, the concept of reasonable$_A$ expectations needs to be construed without presupposing the concept of action. I have already explained how the concept of a normative expectation can be construed without presupposing the concept of action. Indeed, even though I allowed that we speak of holding the agent to an expectation to *perform an action*, I have shown that we can interpret this phrase in an innocent way (allowing that expectations are prima facie fulfilled by performances in general: actions and nonactions alike). I now have to demonstrate that the notion of reasonableness$_A$ can acquire an equally innocent interpretation. I have already hinted at how to do so in section 1.A above, where the concept of prima facie unreasonableness$_A$ is understood in terms of systematic prima facie frustration of normative expectations, and in section 2.A, where the concept of defeating conditions is understood in terms of prima facie fulfillment and frustration of expectations. Let me say a little bit to motivate this construal. It will be best to use to the notion of a defeating condition as an example.

Take the concept of a hindering condition with respect to the expectation to φ, i.e. a defeating condition that is systematically correlated with the frustration of an expectation to φ. So, one might say, breaking a leg is systematically correlated with the frustration of the expectation to run the race. The crucial question that we must ask is what sort of concept of frustration is at stake.

We can distinguish at least three concepts of frustration. An expectation to φ is *agentively* frustrated by actions that can be described as not-φings. An expectation to φ is *non-agentively* frustrated by nonactions (mere happenings) that can be described as not-φings. Finally, an expectation to φ is *prima facie* frustrated (pf-frustrated) by any performances (actions and nonactions alike) that can be described as not-φings. Take the expectation to run a race as an example. It will be agentively frustrated when an agent decides not to run just because he does not feel like it and intentionally fails to run. It will be also agentively frustrated if the agent is called out on an emergency, and so fails to run the race without intending to do so but foreseeing that he will do so. The expectation will be non-agentively frustrated when the agent does not run the race but when his not running is not an action of his, as when he is lying comatose in the hospital

or when his leg is broken. The expectation will be prima facie frustrated in all these cases.

Ignoring the fundamental problem for a moment, let us ask what concept of frustration would fit the notion of a defeating condition. Take agentive frustration first. It is relatively clear that this is not the notion that is at stake. Breaking a leg is not systematically correlated with the agentive frustration of the expectation to run the race. (Someone who breaks a leg *might* also intend or have intended not to run the race, but breaking a leg seems to break the pattern rather than be a part of it.) What about non-agentive frustration? Here the intuitions seem to be quite clear: it fits like a glove. Breaking a leg is systematically correlated with the non-agentive frustration of the expectation to run the race. Someone with a broken leg is not going to run the race, and her not running the race will not be an action of hers. Her not running the race (because of a broken leg) is something that happens to the agent. We would be thus led to conclude that the concept of non-agentive frustration (frustration by mere happenings not actions) should be involved in the notion of a defeating condition.

And it is here, once again, that the fundamental problem arises. For the notion of non-agentive frustration just like the notion of agentive frustration presupposes the very distinction between actions and mere happenings that we want to explicate in terms of (among others) defeating conditions. In order to make sense of the notion of non-agentive frustration we need the concept of a mere happening (and so the distinction between actions and mere happenings). It would be circular to proclaim that we can understand the distinction between action and mere happening in terms of such a notion of defeating conditions.

In other words, we cannot characterize defeating conditions in terms of the concept of agentive frustration, for it misses the target. But we also cannot analyze defeating conditions in terms of non-agentive frustration — although we would be right on the target, the account would be circular. The only option that remains is to choose the concept of prima facie frustration. Only the concept of prima facie frustration would not render the account circular, for only that concept does not presuppose the distinction between actions and mere happenings.

But our characterization of defeating conditions is not quite sufficient. Here are two objections which rely on the simple fact that the set of agentive fulfillments and the set of non-agentive fulfillments are included in the set of prima facie fulfillments of an expectation. Consider the first fact. When an expectation is agentively fulfilled it is also prima facie fulfilled. So suppose that at a certain stage of the development of the concept of agency, it is discovered that there is a condition that is systematically correlated with the frustration of an expectation. As it turns out, however, the performances that it is correlated with are exclusively agentive (relative to the understanding of 'agentive' at that stage). In this case, it is still true that the condition is systematically correlated with prima facie frustration of the expectation but this is only because the set of agentive frustrations is by definition included in the set of prima facie frustrations. Were such a situation to occur, we would not have a reason to speak of a defeating condition. There would be little reason to speak of a condition that takes a certain kind of performance out of the agent's power. After all, when the condition occurs, the agent systematically performs only *actions*. Were the correlation to extend to cover not only what is recognized as actions but also what is recognized as mere happenings, there would be reason to suppose that a new defeating condition is at work, which would require us to change our conception of what is an action and what is not.

We can summarize this in the form of an informal principle. Let *d* be a potentially new defeating condition that is systematically correlated with the frustration of an expectation to φ. Let *D* be the set of existing defeating conditions, which determine whether an expectation is fulfilled and frustrated agentively (relative to *D*): when some condition of the set occurs, the expectation is fulfilled or frustrated non-agentively (relative to *D*).

Principle I :

> If *d* is systematically correlated with agentive fulfillment/frustration (relative to *D*)[13] of the expectation to φ but not with the non-agentive

[13] Note that the concept of agentive frustration and fulfillment are relativized to an existing set of defeating conditions. The use of such a concept does not lead to circularity.

> fulfillment/frustration (relative to *D*) of that expectation, then *d* is not a new defeating condition (relative to *D*).

For a different reason, the opposite situation does not engender new defeating conditions. Suppose that at a certain stage of the development of the concept of agency, it is discovered that there is a condition that is systematically correlated with the frustration of an expectation. As it turns out, however, the performances that it is correlated with are exclusively non-agentive. In this case, it is still true that the condition is systematically correlated with prima facie frustration of the expectations but this is only because the set of non-agentive frustrations is by definition included in the set of prima facie frustrations. Were such a situation to occur, we would not have a reason to speak of a new defeating condition either. Here, however, the reason is different. Given that the frustrations in question are all non-agentive, this means that in all these cases, some defeating conditions are already in play. The alleged new condition covers a terrain that is already covered by the existing set of defeating conditions.[14] It is not a new defeating condition.[15]

Principle II (Economy of Defeating Conditions):

> If *d* is systematically correlated with non-agentive (relative to *D*) fulfillment/frustration of the expectation to φ but not with agentive (relative to *D*) fulfillment/frustration of that expectation, then *d* is not a new defeating condition (relative to *D*).

The spirit behind the second principle could be in fact generalized. As long as the systematic correlation of a condition *d* can be fully understood in terms of the existing set of defeating conditions *D*, *d* is not a new defeating condition. The principle of economy of defeating conditions is intended to bar the introduction of "funny" agglomerative conditions.

---

[14] I can see one exception here. It may be that a bunch of defeating conditions correlated with the frustration of a variety of expectations is replaced with one defeating condition ('a syndrome'). Such a unification, however, would need to be supplemented with some theoretical benefits.

[15] This seems to be a general practice in law making. New laws are only introduced if the cases they are intended to cover are not already covered by any combination of already established laws.

Principle III (of Non-Composition of Defeating Conditions):

> If d_1 is systematically correlated with pf-frustration of the expectation to φ, d_2 is systematically correlated with pf-fulfillment of the expectation not to φ, then d_1-or-d_2 is not a new defeating condition with respect to the expectation to φ.
>
> If d is systematically correlated with pf-frustration of the expectation to φ, then not-d is not a new defeating condition with respect to the expectation to φ.

D. Are Defeating Conditions Causes?

So far, I have been speaking of defeating conditions occurring rather than causing the events that would otherwise be actions. There is no problem in supposing that defeating conditions sometimes do cause the events that would otherwise be actions. Sometimes it is in fact plain that they do. For example, when a spasm causes a hand to tremble and the spoon to fall out, the spasm causes a performance (the falling out of the spoon) that might appear as if it fulfills the expectation that the agent drop the spoon, and yet, the performance is something that happens to the agent in virtue of the fact that it has been caused by the spasm. Similarly, drugs may cause memory problems causing one to forget to pick up a child from school. A sudden wind gust may throw one forward causing one to fall onto somebody else. And so on.

Indeed, it might not be perhaps too outrageous to suggest that it is because defeating conditions frequently cause mere happenings (nonactions) that the language of causality suggests itself with respect to reasons causing actions. The mere happenings are events that might look like actions but are (typically) caused by defeating conditions. This might lead one to search for the corresponding "agentive" causes of actions.

Be this as it may, it is not clear what is gained by speaking of the defeating conditions as causing mere happenings. The problem is not only that the idea of causality is notoriously difficult to understand but rather that it is notoriously difficult to apply in at least certain kinds of cases. For while we have little problem understanding how the wind causes one to fall onto a crowd, we have progressively more problems in grasping

the sense in which one's oversleeping caused one not to go to the meeting, one's forgetting caused one not to do the homework, or the absence of power tools "caused" one not to build the bridge.

In the last case, the problem in construing defeating conditions as always causing mere happenings is perhaps most vivid. The fact that one does not have power tools suffices to make it unreasonable$_A$ to expect of one that one build a bridge. Suppose that John does not have the required power tools (through no fault of his own), so despite the fact that he intends to build a bridge and is contracted to do so, it would be unreasonable$_A$ to expect of him that he build the bridge. In such a case, we might be tempted to say that John's not building the bridge was caused by his not having the required power tools. But suppose that while it is true that John does not have the power tools, he does not build the bridge not because he does not have the required equipment but because he does not intend to do so in the first place. In fact the idea might appear ridiculous to him if he were confronted with it. In such a case, we might think that it is John's indisposition (lack of intention, preparation or what not) that caused him not to build the bridge rather than the lack of relevant tools. And yet, I think that both situations are exactly alike with respect to the defeating condition: it is because John lacks the power tools that it would be unreasonable$_A$ to expect him to build the bridge.[16] This may be independent of what actually *explains* his not building the bridge.[17]

I prefer therefore not to require that defeating conditions must cause mere happenings even though many of them do.

3. Some Objections

A. Are Desires Defeating Conditions?

When I want chocolate, I eat it. When I want to go for a walk, I usually go for a walk. Desires appear to be the paradigmatic examples of conditions that are systematically correlated with the fulfillment of the expectations they justify and with the

---

[16] In the second case, we might also have additional defeating conditions: his lack of skills, for instance.

frustration of the contrary expectations. Are they defeating conditions? Would they make the expectations unreasonable$_A$? I will answer this question more systematically in section 4. For now let me make three points.

Not all desires make expectations unreasonable$_A$, but some do. Compulsive desires do indeed make it unreasonable$_A$ to expect of the agent that she perform the action justified by the desire. A person's desire to wash her hands every five minutes makes it unreasonable$_A$ to expect her not to wash her hands, or to wash her hands. The washing of the hands is in this case beyond the compulsive-obsessive's control. But it seems clear that a person's non-compulsive desire for a walk does not make it unreasonable$_A$ in any way to expect of the person that she does or that she does not take the walk. In section 4, I will explain how to understand the difference between compulsive and non-compulsive desires.

One may develop some degree of skepticism with respect to the alleged systematicity of the correlation. Recall that in order for a condition to count as being systematically correlated with the frustration of an expectation say, it must be the case that most agents *under favorable conditions* would frustrate the expectation. The "favorable conditions" comprise the agent's cooperativeness, lack of extraneous pressures, etc. In order for a desire to φ to count as being systematically correlated with the frustration of an expectation not to φ, say, it would have to be the case that in situations where agents are cooperative, under no pressures, etc., they would systematically satisfy the desire rather than the expectation. It is plausible to think that this would be satisfied for some very strong desires, paradigmatically for visceral desires like thirst or hunger. It is less vivid with respect to other desires to walk on the beach, to climb Mount Everest, to do the most outrageous thing one can think of in a public place, to vote against one's convictions, etc. Such desires frequently do not lead to their fulfillment.

Finally, one could try to employ Principle I (see p. 110, above) to argue that desires ought not to qualify as defeating conditions. This is because to the extent that

---

[17] This underscores the point that the question of the nature of action and nature of action explanation are different issues.

desires tend to be correlated with the fulfillment of the expectations they justify, the fulfillment in question tends to be agentive.

For now, I will simply assume that non-compulsive desires should not be counted as defeating conditions.

B. Defeating Conditions and Frankfurt-Type Cases

Frankfurt-type cases purport to illustrate that there are situations where we would hold an agent responsible despite the fact that he could not have *done* otherwise.[18] Consider the following case: Jones decides to kill the mayor of the town. He carries out his plan to the letter, shoots the mayor who dies as a result. Unbeknownst to Jones, evil scientists have implanted a device into Jones' brain which, were Jones to decide not to kill the mayor (or waver after his decision), would have swayed Jones to kill the mayor anyway. The intuitions about cases of this sort have been almost uniform. Jones is responsible for killing the mayor. At the same time, it has been claimed, Jones could not have *done* otherwise: he could not have not killed the mayor (see Figure 2). The question for us is first of all whether the presence of the counterfactual intervener functions as a defeating condition in this case. I will argue that it does not.

The structure of these cases can be captured as follows. In the ordinary case (without the counterfactual intervener), we can suppose that it would be both reasonable$_A$ to expect of Jones that he kill the mayor and reasonable$_A$ to expect of Jones that he not kill the mayor. Does the presence of the counterfactual intervener render it unreasonable$_A$ to expect of Jones that he kill the mayor? One might think that it does. After all, given the presence of the counterfactual intervener it is determined that the mayor will die at Jones' hands. It would thus seem that the presence of the counterfactual intervener is systematically correlated with the pf-fulfillment of the expectation that Jones kill the mayor. However, there are good reasons not to treat the

---

[18] Belnap and Perloff ("Seeing to It that: A Canonical Form for Agentives," in (eds.) H.E. Kyburg, Jr., R.P. Loui, G.N. Carlson, *Knowledge Representation and Defeasible Reasoning* [Dordrecht: Kluwer, 1990], pp. 175-199.) point out that there are two interpretations of the phrase "could have done otherwise." On the stronger, to say that α, who φed, could have done otherwise is to say that it was possible that α see to it that α not φ. On the weaker: it is to say that it was possible that it was not the case that α see to it that α φ. Frankfurt-type cases are directed against the stronger interpretation of the phrase.

presence of the counterfactual intervener as a defeating condition. There are at least two ways to argue for this conclusion.

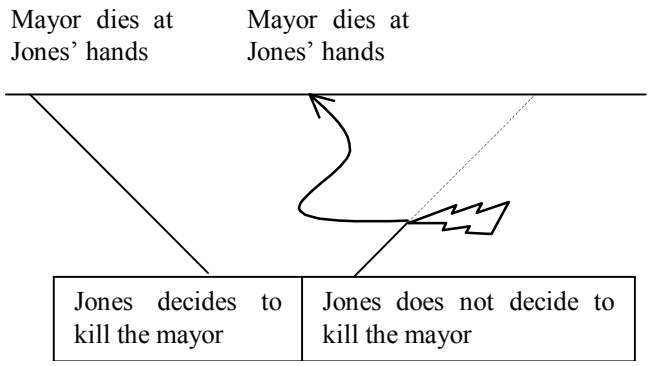

Figure 2. The structure of Frankfurt-type cases.

The first way to argue that the presence of the counterfactual intervener does not defeat the reasonableness$_A$ of holding Jones to the expectation that he kill the mayor is similar to the impact the case has had on the literature of the subject. The lesson that is sometimes drawn from Frankfurt's cases is that they show that our conception of conditions of responsibility is based on what *actually* happens rather than on what might happen.[19] Indeed as the case is described the counterfactual intervener does not affect the course of events in the actual sequence. This is different for the possible sequence. Were his intervention to occur, it would make it unreasonable$_A$ to expect of the agent that he kill the mayor. Insofar as it is the supposition of the examples that the counterfactual intervener will not intervene, we appear to have no reason for thinking that it would be unreasonable$_A$ to expect of the agent that he kill the mayor.

---

[19] The most prominent representatives of this actual-sequence approach to responsibility are Harry G. Frankfurt, *The Importance of What We Care About. Philosophical Essays* (Cambridge: Cambridge University Press, 1988); John Martin Fischer, "Responsiveness and Moral Responsibility," in (ed.) Ferdinand Schoeman, *Responsibility, Character, and the Emotions* (Cambridge: Cambridge University Press, 1987), pp. 81-106 and John Martin Fischer, *The Metaphysics of Free Will. An Essay on Control* [Oxford: Basil Blackwell, 1994].

One might object to this response that it presupposes that we think of defeating conditions as causes of mere happenings, which is what I decided not to require (section 2.D). Moreover, it might be argued that what it shows is that the interference by the counterfactual intervener does not defeat the reasonableness$_A$ of holding Jones to the expectation to kill the mayor in the actual case, but it would in the possible case. But the original question was not whether the interference (*C*) is a defeating condition but rather whether the presence of the counterfactual intervener is (*K*-or-*C*). After all, given the fact that the counterfactual intervener is present, it is settled that the mayor will die at Jones' hands. The only way to dismiss this alleged defeating condition as bogus, the objection continues, would be to suggest that it is not a causal condition, while the interference by the counterfactual intervener is. As I explained in section 2.D, the issue of deciding what is and what is not a causal condition is rather delicate. Rather than offering an account of the matter, I propose an alternative (though ultimately not unrelated) explanation why the presence of the counterfactual intervener is not a defeating condition.

There are exactly two avenues to the mayor's death at Jones' hands envisaged in the example. First, Jones might decide to kill the mayor (*K*) and so kill him. Second, Jones might not decide to kill the mayor, in which case the counterfactual intervener will take over (*C*), leading Jones to kill the mayor. In the first case, the expectation is agentively fulfilled, in the second case it is also fulfilled but non-agentively. The case is constructed so that either *K* or *C* occurs (this is what the presence of the counterfactual intervener amounts to). Given the presence of the counterfactual intervener (*K*-or-*C*), the expectation that Jones kill the mayor will be prima facie fulfilled. Since the correlation is not with exclusively agentive fulfillment, *K*-or-*C* does not violate Principle I. Principle II would exclude *K*-or-*C* if *K* and *C* were themselves defeating conditions. While *C* is a defeating condition, *K* is not (see section A, above, and section 4, below).

However, the case does violate Principle III. What is special about the example is the fact that two conditions are identified, either one or the other occurs, and both of them are systematically correlated with the pf-fulfillment of the expectation to kill the mayor. It follows from Principle III that the condition *K*-or-*C* is not a defeating condition with respect to the expectation that Jones kill the mayor. The condition is not a new defeating

condition, for it relies on the disjoining of systematic correlations we have a good understanding of. Hence it is reasonable$_A$ to expect of Jones that he kill the mayor.

In this case, the presence of the counterfactual intervener does not make unreasonable$_A$ either the expectation that Jones kill the mayor or the expectation that Jones not kill the mayor. The presence of the counterfactual intervener is not properly construed as a defeating condition. However, the actual intervention by the scientist would be construed as a defeating condition, were it to occur.

In Appendix A, I shall discuss how this approach can be used to shed some light on the debate concerning the so-called asymmetry thesis.

C. Unintentional Omissions

I want to close this section by noting that the account thus far is too poor to capture the concept of reasonableness$_A$. This is best illustrated with respect to unintentional omissions.

Here is a familiar scenario. An employee is expected to be at a meeting at 9am, but he oversleeps. As indicated, I want to insist that despite the fact that the agent is sleeping at 9am, it would still be reasonable$_A$ to expect of him that he be at the meeting. Yet this is not the result that the concept of reasonableness$_A$ thus far developed yields. Surely being asleep at the time one is expected to be at the meeting is systematically correlated with the frustration of the expectation to be at the meeting. It would thus appear that being asleep is a defeating condition with respect to the expectation to be at the meeting and so renders the expectation unreasonable$_A$. We will see in the next section how to avoid this conclusion.

4. Defeating Defeating Conditions

So far I have adopted a relatively straightforward characterization of defeating conditions as those conditions that are systematically correlated with the frustration or fulfillment of an expectation. I have suggested that this idea of defeating conditions constitutes a way of delimiting our understanding of what it means to say that it is within the agent's power to do something. But there is a complication.

Let us take an expectation of α that α φ. Let us suppose that *C* is systematically correlated with the frustration of the expectation to φ. Intuitively, when *C* occurs it is not "within the agent's power" to φ. A question that might be reasonably raised is: Is it "within the agent's power" to see to it that *C* does not occur?

It seems clear that there are such cases. It is a well-known fact that drinking an immoderate amount of alcohol will reliably result in a loss of much control: it is systematically correlated with the frustration of a range of expectations (to drive safely, to behave responsibly, etc.). According to the account so far, given that a person has ingested an immoderate amount of alcohol, it will be unreasonable$_A$ to expect her to drive safely, or to behave responsibly. So, looking forward a little, if she drives unsafely it will not be a breach of expectation, it will not be something she did. Nor will it count as her doing it if she abuses someone while drunk. But the account is too simple-minded. We need to add a normative condition on what counts as a defeating condition, by allowing for the possibility of there being circumstances where the defeating character of a defeating condition is itself defeated.

Let *C* be systematically correlated with the frustration of the expectation to φ. The defeating character of *C* will be itself defeated if it is reasonable$_A$ to expect of the agent that she bring it about that *C* does not occur. We can thus enrich our characterization of reasonableness$_A$.

> It is reasonable$_A$ to expect of α that α φ if and only if either (a) no defeating condition (with respect to the expectation to φ) occurred, or (b) such a defeating condition did occur but it was countered by an appropriate counterdefeating condition, or (c) such a defeating condition did occur and it was unreasonable$_A$ to expect of α that α bring it about that it not occur.

Let us note that according to this characterization, conditions that are systematically correlated with the frustration of an expectation to φ will not defeat that expectation's reasonableness$_A$ *unless* it is also unreasonable$_A$ to expect the agent to prevent them from occurring. So, suppose that α is reasonably$_A$ expected to φ and a condition *C* occurs which is systematically correlated with the frustration of the expectation to φ. Suppose

further that it is reasonable$_A$ to expect of the agent that *C* not occur. It follows that the expectation of α to φ is reasonable$_A$; its reasonableness$_A$ has not been defeated by *C*.

Let us consider two examples to illustrate the point. Let us first take the example mentioned above. A person is reasonably$_A$ (and surely reasonably$_N$) expected to drive safely (φ). She is at a party and has far too much to drink (*C*). It is well known that a state of drunkenness systematically interferes with agents' fulfillment of the expectation to drive safely. However, it is also reasonable$_A$ to expect the agent not to drink too much. If this is so then it is also reasonable$_A$ to expect of her that she drive safely despite the fact that she is no longer in a state to fulfill the expectation.

An employee is reasonably$_A$ (and reasonably$_N$) expected to be at an important meeting (φ). He oversleeps (*C*). His oversleeping makes it impossible for him to be at the meeting, and would thus appear to make the expectation that he be there unreasonable$_A$. However, it is reasonable$_A$ to expect of him that he not oversleep. (And in support of our so thinking we cite the fact that the agent can put alarm-clocks all around him, alert his neighbors, go to bed early, etc.) If so then (in absence of further conditions, to be explained below) it straightforwardly follows on our account that despite the fact that while the agent is asleep it is not "within his power" to be at the meeting, it is still reasonable$_A$ to expect of him that he be there. The fact that he oversleeps does not defeat the reasonableness$_A$ of the expectation that he be at the meeting because it is reasonable$_A$ to expect of the agent that he not oversleep.

As Figure 3 shows, the structure is indeed complex. We can see a further complication arising if we consider the case of the employee who oversleeps and so fails to come to the meeting but who oversleeps because he has been drugged. In such a case, our intuitive attitude toward such a person changes. We are right in believing that it is not within the agent's power to come: it is unreasonable$_A$ to expect the agent to be at the meeting.

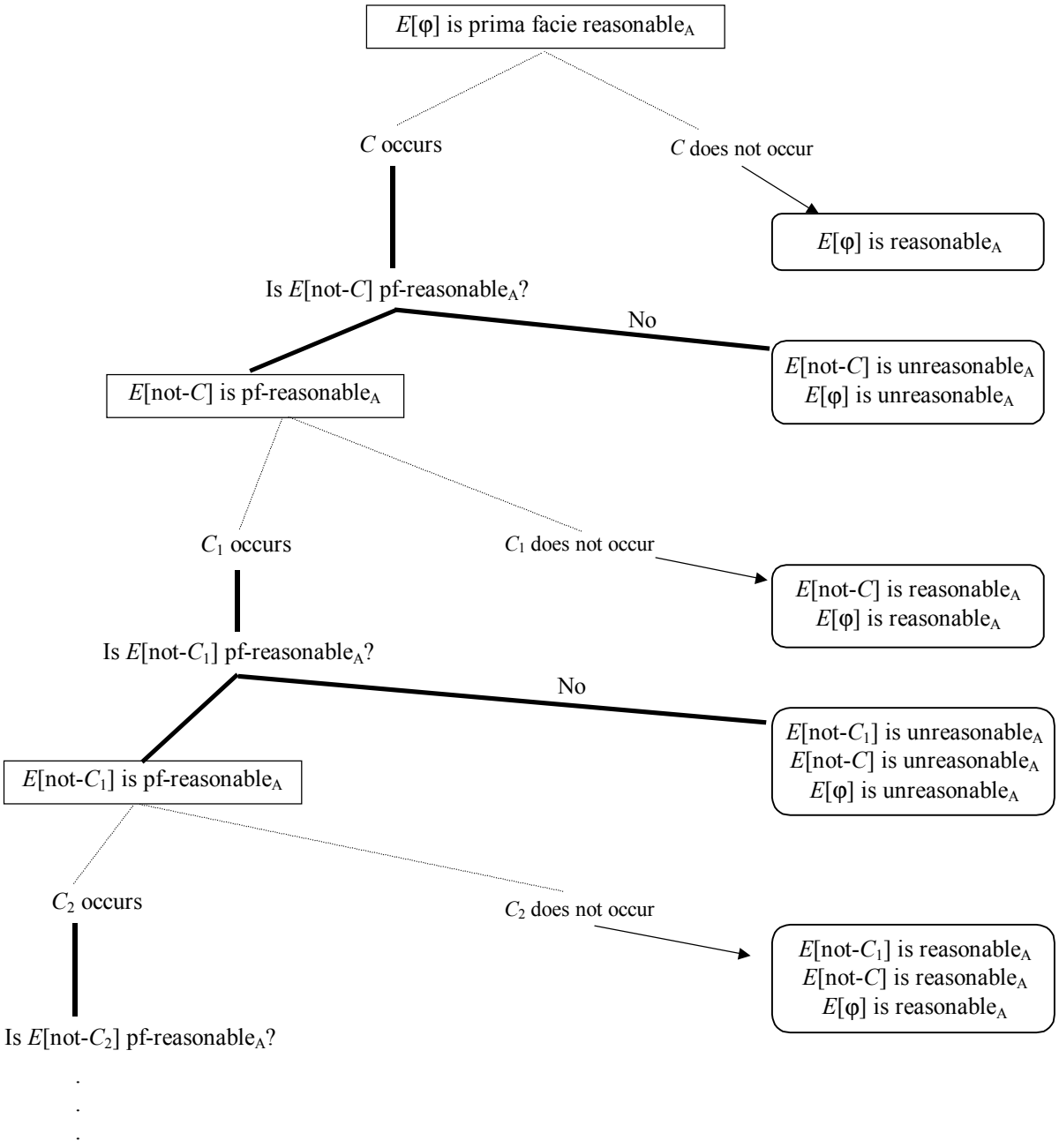

Figure 3. The structure of dependence of reasonableness$_A$ of expectations on further defeating conditions. Condition C is a defeating condition with respect to the expectation to φ (E[φ]); C_1 is a defeating condition with respect to the expectation to prevent C from occurring (E[not-C]); C_2 is a defeating condition with respect to the expectation to prevent C_1 from occurring (E[not-C_1]).

The reason why such a complication is possible is this. For any potential defeating condition (with respect to φing), a condition that is systematically correlated with the frustration of an expectation to φ, it must be ascertained whether or not it is reasonable$_A$ to hold the agent to the expectation that he prevent the condition from occurring. Assuming that it is reasonable$_A$ to expect of the agent that he prevent the condition from occurring, we have not one but two expectations in play. And just as there are possible defeating conditions to the first expectation, so there are defeating conditions to the latter.

Abstractly, let us assume that it is reasonable$_A$ to expect of α that α φ. *C* occurs. *C* is systematically correlated with the frustration of the expectation to φ. However, *C* does not defeat the reasonableness$_A$ of the expectation to φ because it is reasonable$_A$ to expect of the agent that he prevent *C* from occurring. This second expectation (that the agent prevent *C* from occurring) also has potential defeating conditions, however. Conditions C_1 and C_2 are systematically correlated with the frustration of the expectation that he prevent *C* from occurring. However, it is unreasonable$_A$ to expect of the agent that he prevent C_1 from occurring, but it is reasonable$_A$ to expect of the agent that he prevent C_2 from occurring.

Suppose first that C_1 occurs. Since it is unreasonable$_A$ to expect of the agent that C_1 not occur, C_1 defeats the reasonableness$_A$ of the expectation that *C* not occur. We will remember, however, that the reason why *C* did not defeat the reasonableness$_A$ of the original expectation that α φ was that it was reasonable$_A$ to expect of α that *C* not occur. Now, however, condition C_1 has defeated the reasonableness$_A$ of the expectation that *C* not occur. As a result, the occurrence of *C* defeats the reasonableness$_A$ of the expectation of α that α φ.

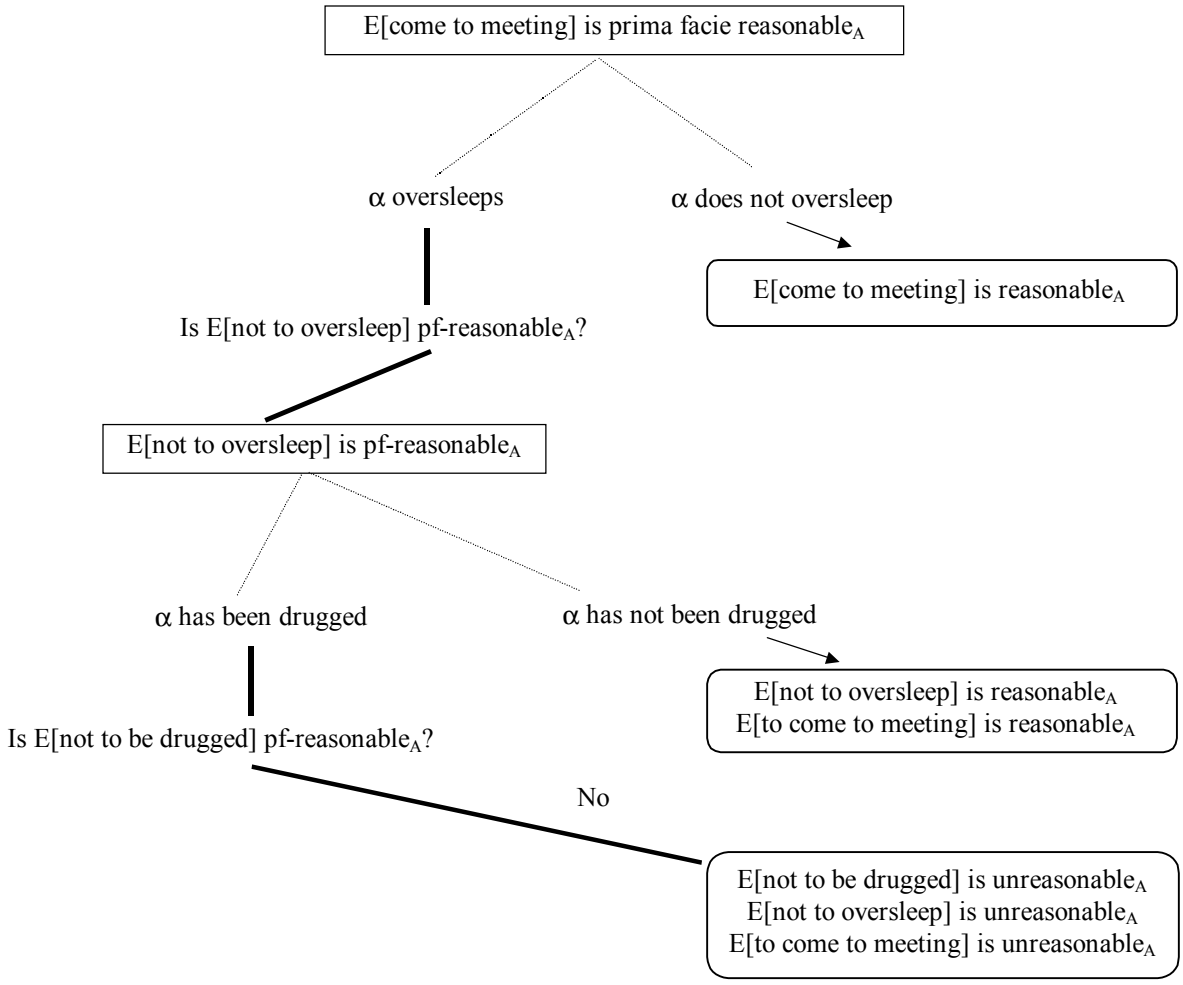

Figure 4. The structure of dependence of reasonableness$_A$ of the expectation to come to the meeting on the conditions that the agent overslept and that he overslept because he has been drugged.

Although complicated, this is the structure exhibited by the case of the employee not coming to the meeting (not φing) because he has overslept (*C*) as a result of being drugged (C_1) (see Figure 4). It is prima facie reasonable$_A$ to expect the employee to come to the meeting (φ). He oversleeps (*C* occurs). Oversleeping is systematically correlated with the frustration of the expectation to come to meetings. However, it is prima facie reasonable$_A$ to expect of the agent that *C* does not occur (not to oversleep). The agent has been drugged, as a result of which he oversleeps (C_1 occurs). Being drugged is systematically correlated with the frustration of the expectation not to oversleep. And, it is not prima facie reasonable$_A$ to expect of the agent that he not be drugged. So, the agent's being drugged defeats the reasonableness$_A$ of the expectation that he not oversleep. Since now it is no longer reasonable$_A$ to expect of the agent that he not oversleep, his oversleeping does defeat the reasonableness$_A$ of the expectation to come to the meeting. In other words, the expectation to come to the meeting is no longer reasonable$_A$ in view of the fact that the agent has overslept as a result of being drugged.

The well-like character of defeating conditions (captured in Figure 4), which may be subject to defeat by further conditions, is responsible for much of the open-ended nature of our attribution of actions. One way in which this feature has been manifest in the literature is by the necessity of introducing the open-ended qualifier "in the right way."[20] It is also sometimes captured by the introduction of the standard of *due care*. In the above terms, the standard of due care (relative to a certain expectation) comprises all those conditions (systematically correlated with the frustration of the expectation) where it is prima facie reasonable$_A$ to expect of the agent that they not occur (and so not interfere with the fulfillment of the expectation).

Some Desires Render Expectations Unreasonable$_A$, Others do Not. I have already answered one of the objections set out at the end of the last section. I have shown that sometimes when a person unintentionally omits to do something it would be still reasonable$_A$ to expect of him that he do it. Let me answer the second one. The objection

---

[20] Donald Davidson, "Freedom to Act," in *Essays on Actions and Events* (Oxford: Clarendon Press, 1980), pp. 63-81. The context of Davidson's discussion may appear different because it is introduced with respect to the causal theory of action. But that it is not as different as it might appear will become clear.

was that desires, which are systematically correlated with the fulfillment of the expectations they justify, would count as defeating conditions and hence render the expectations unreasonable$_A$.

We can now see the limitations of that objection. It will be indeed the case that a desire to φ, if it is systematically correlated with the fulfillment of the expectation to φ, renders the expectation unreasonable$_A$ if it is unreasonable$_A$ to expect of the agent that she prevent her desire from affecting her action. It is arguable that there are some desires like that. Compulsive desires, for instance, are desires that we would intuitively think beyond the agent's power to control. The compulsive-obsessive's desire to wash his hands every five minutes is not something that he can control. But this is not so for most other desires. Even if someone's desire for chocolate is very strong, so whenever he has it he submits to it and eats chocolate, it would (or at least might) be reasonable$_A$ to expect him not to eat the chocolate. It might be reasonable$_A$ to expect him to control the desire. Let me illustrate with a somewhat too graphic example.

Let us stipulate that there is a pretty much stable pattern for a particular type of chocoholic. He occasionally gets an urge to eat chocolate, which stimulates his thinking about it, which further strengthens his desire for chocolate, and so on. Finally, the desire becomes strong enough to move him to search for chocolate. Once the chocolate is in his sight, only force could prevent him from eating it. The piece of chocolate is doomed, he can do nothing about it.

Most desires do not function like that. The connection between the desire and the action is usually not so strong. In fact, it is intuitively implausible to think that there is any single type of mental state that is so strongly tied to action. But even in this scenario, where it is clear that not so much the desire on its own but the desire together with the sight of chocolate is systematically (inescapably) correlated with the fulfillment of the expectation to eat chocolate, it would be reasonable$_A$ to expect the agent not to eat the chocolate. Why? Because it would be reasonable$_A$ to expect the agent to prevent himself from ever getting to the stage where he sees the chocolate, which overwhelms him. It would be reasonable$_A$ to expect him to counter the thoughts about chocolate with thoughts about an experiment he should conduct instead, for example.

It is, of course, possible that there are desires that the agent cannot control in such a manner. A compulsive desire to wash one's hands every five minutes may be systematically correlated with the fulfillment of the expectation to wash one's hands. Yet, it may be that the agent can do nothing to prevent the desire from leading to action, i.e. that all attempts to counter the desire (whether by thinking other thoughts or by engaging in other activities) may be systematically frustrated.

• • • • •

I have suggested in Chapter II that one of the basic commitments of a responsibility-based approach to action is to develop an account of practical responsibility that would be significantly different from moral and legal responsibility. It is in part because H.L.A. Hart did not offer such an account that his theory has been subjected to sharp criticism, which pertained not only to the details but to the very core of his account. One of the fundamental charges that responsibility-based accounts of action face is the fundamental problem: the objection that such accounts rely on an initial mistake — they take the concept of responsibility to characterize the logically prior concept of action. In Chapter II, I have promised to develop a concept of practical responsibility that would be immune to the fundamental objection. The account is now complete.

I have argued that

> an agent α is practically (task-)responsible for φing if and only if it would be reasonable$_A$ to expect of α that α φ.

In Chapter III, three tasks have been accomplished. First, I have distinguished between normative and predictive expectations (between what it means to expect of α that α φ and to expect that α will φ). Second, I have indicated that despite the fact that what complements the normative expectation appears to be an agentive statement ('α φs'), this does not necessarily mean that the account that appeals to normative expectations falls prey to the fundamental problem. For I have declared that in the first instance, we shall assume that an expectation of α that α φ will be fulfilled not only by α's actions but, more generally, by α's performances (which include actions and mere happenings). Third, I have argued that we ought to focus on practical normative expectations, not on

specifically moral (with a moral justification) or legal (with a legal justification) ones. In this way, the concept of responsibility is broader than its specifically moral or legal counterparts.

In Chapter IV, I have then undertaken the task of characterizing the sense in which normative expectations must be reasonable. I have distinguished two senses of reasonableness: agent-centered reasonableness$_A$ and specifically normative reasonableness$_N$. Roughly, expectations are reasonable$_N$ if there are good reasons for holding the agent to them; expectations are reasonable$_A$ if it is "within the agent's power" to do what is expected of him. I have not attempted to analyze the concept of reasonableness$_N$, for as we will see in the next chapter, this concept is less important to the distinction between actions and mere happenings.

In the present chapter, I have proposed an account of reasonableness$_A$, which is meant to elucidate the meaning of the metaphor of what is "within the agent's power." I have argued that those normative expectations that are not unreasonable$_A$ are reasonable$_A$. Normative expectations are made unreasonable$_A$ by the occurrence of defeating conditions, conditions that are systematically correlated either with the fulfillment or the frustration of a given expectation (sections 1-2). I have also pointed out that a defeating condition can be itself defeated if it is reasonable$_A$ to expect of the agent that she prevent the defeating condition from occurring (section 4).