Foundations of Decision Theory and Free Will

Wolfgang Spohn

What do the foundations of decision theory have to do with the problem of free will? A lot, and I will explain how I see the topics to be connected.

To begin with, decision theory, at least in the interest of philosophers, is a normative theory about how we should rationally act – given our beliefs and desires. The normative evaluation of the beliefs and desires themselves is, in the traditional understanding, not the task of decision theory, although it is an urgent issue as well. Free will is an issue about the actual determination of our actions. So, a first connection is given by the fact that the normative ideal of rational action is not just a castle in the air, but also an empirical approximation. We discuss what our normative ideals should be. This is never fixed, but at any time the defeasible result of our normative dispute. And we strive to fulfill our normative ideals. This is the essence of norms; otherwise, our normative dispute would be pointless. If our striving were fully successful, the normative ideal would already describe empirical reality. Of course, we do not fully succeed. Still, we have a stronger or weaker tendency towards acting according to the norms. Hence, every empirical theory about the determination or causation of our actions must somehow take the normative ideal into account. I shall return to this important point.

So, the next question is: What is the normatively correct decision theory? This addresses the foundations of decision theory, which are hotly disputed in philosophy, since Richard Jeffrey opened a surprising alternative in his *Logic of Decision* in 1965, namely that between causal and evidential decision theory. Pace Jeffrey, I contend that causal decision theory is the correct one, and later in my talk I shall add what I call reflexive decision theory. What is causal decision theory? It is characterized by what I call the NoPA principle: no probabilities for acts. That is, I postulated in 1976 that a decision model must not contain probabilities (or any other epistemic evaluation) of the possible actions deliberated in that model. Isaac Levi spoke of "deliberation crowds out prediction", and Alan Hájek maliciously called it the DARC principle: "deliberation annihilates rational credences."

I have put forward various arguments in favor of this principle. However, my main point was always parsimony: In any given decision situation I have several options all of which I am able to take. This is the presupposition of any practical deliberation. If there were only one option, there would be nothing to decide and to deliberate. And then I somehow evaluate which of the options are the best ones. Of course, deliberation and evaluation usually

are a matter of course or at best implicit. It is not implied that we deliberate all the time. In any case, probabilities for acts play no role whatsoever in that evaluation, in determining rational action, and hence they have no place in modeling rational action. I have nowhere seen a decision rule that determines the optimality of actions with reference to probabilities for them; they even play no role in maximizing evidential expected utility.

Things are different with so-called probabilistic acts. There are various reasons for considering them. There is deliberate randomization; and the resulting action may be uncertain in the case of tryings and in the case of weakness of will. But let's not complicate things and stick to the basic case free of such additional considerations.

An important consequence of the NoPA principle is the ExA principle: acts are exogenous. That is, in any decision model, the action variables are exogenous variables. If we represent the model as an influence diagram, i.e., as a Bayesian net with action variables, then the action variables have no parents. In other words, actions have no causes in a decision model. This is not to say that they do not have causes at all. It only says that they have no causes in the agent's perspective, i.e., in the decision model representing the agent's mental state.

For inferring the ExA principle from the NoPA principle, we crucially need a theory deriving causal relations from probabilities. No more than the standard theory of causal Bayes nets is needed in my view. This may be contested, but let's not discuss this point. In any case, the ExA principle has the important consequence that actions can be evidentially relevant only for their causal consequences. This is why it explicates causal decision theory as opposed to evidential decision theory.

The main objection to the NoPA principle has always been: What's so difficult about having an epistemic assessment of my own future actions? Other people have it all the time, and I know myself much better than others do. So, what should prevent me from predicting my own actions? Nothing, of course. I can well predict my future unintentional behavior. And my future intentional actions, too. And they comprise a lot. It is not only deliberate action. It is also habitual behavior, because most of our habits are under rational control. In my understanding, it is, e.g., also action motivated by anxiety; the anxiety is accompanied with very negative utilities that virtually dictate anxiety-avoiding action. Of course, this is debatable; it's not so important.

The important point is that the prediction of future intentional action is tacitly undergirded by a prediction of the future decision situation from which that action springs. Uncertainty about what I will do is always accompanied by uncertainty about the future decision situation I will face. There is no principled problem with all this. But it entails an important restriction of the NoPA principle: namely, it applies only to the possible actions presently under consideration, or better, presently at issue.

But why should the NoPA principle apply to the actions presently at issue? What is the blind spot here? Again none. As said, the point is that the epistemic assessment of these actions does not contribute anything to the evaluation of the possible actions as optimal or suboptimal. Once one action is evaluated as optimal, I will take it (at least in the idealized rational picture), and this is easily predicted. However, this prediction is epiphenomenal; it follows from the decision and does not add anything to it. Hence, in the decision model representing my present decision situation probabilities for the actions at issue need not and do not occur. This applies to all my future decisions as well; they also do not contain probabilities for the actions at issue in future. Presently, though, I can properly predict my future decision situations and hence my future actions.

So, I stick to the NoPA principle. It entails the ExA principle. And this means that in a decision model, i.e., in the agent's practical normative first-person perspective, the actions at issue are uncaused, undetermined – in other words: free in the sense relevant to the free will debate. This is how the normatively correct decision theory, or what I take as such, is related to the free will debate. At the same time, I can admit, indeed I fully agree, that in the observer's theoretical descriptive third-person perspective, the agent's actions are caused and perhaps even determined. It would be silly to deny this.

I just characterized each perspective in three ways which I use interchangeably. The first-person perspective is a practical and normative one – the agent must act and think what to do –, while the third-person perspective is merely descriptive and theoretical; the observer's own actions are not at issue at all.

Hence, I may be called a compatibilist, but not of the usual sort. I establish the compatibility of freedom and determination by distinguishing two perspectives. Kant pursued the same strategy by distinguishing two kinds of worlds. I am free in the intelligible world, but determined in the phenomenal world. You find softer formulations in Kant that seem to claim merely two different points of view. Essentially, though, he seems to be serious about his dubitable metaphysics. Anyway, let's not engage in the Kantian cosmos. By all means, I think my two perspectives are intelligible. Of course, the distinction of two perspectives is not unusual at all; think, e.g., of Daniel Dennett's physical and intentional stance. However, let's dispense with a more elaborate comparative discussion.

Clearly, I have established compatibility in a very simple, indeed cheap way. Apparently, I have just deferred the issue of free will to the relation between the two perspectives. One perspective must be the primary one, and then my compatibilism tips over either to libertarianism or to hard determinism. But then it seems clear which perspective is the primary one, namely the observer's theoretical perspective. He has the complete causal picture. And the interventionist theory of causation precisely explains how we get from the complete picture to the agent's causal picture: namely precisely by truncating the complete

picture with respect to the actions at issue, i.e., by deleting precisely the causal arrows pointing to the actions. Thus, the agent's picture is simply an incomplete version of the observer's picture.

So, do I grant that the empirical perspective is the primary one, the only one that counts in the end? So that we are stuck with the problem of free will as badly as ever? No, I do not grant this. None of the two perspectives is primary; they are on a par and have a rather complicated relation. Let me explain this point in a bit more detail.

First, there is no fundamental mystery whatsoever for the observer's empirical perspective as to how actions are caused. Ideally, it is precisely the practical deliberation that causes the ensuing action. This formulation suggests that we actually deliberate all the time. As said, this should not be implied. We should rather say that the action is caused by the mental set-up represented by the decision model appropriate for the case at hand. Clearly, this causal explanation is the only one consistent with our self-understanding as beings necessarily endowed with a normative perspective.

Of course, decision theory only provides the basic pattern. Even if this pattern were roughly true, we should inquire how it precisely works, what the underlying mechanisms are, how that mental set-up of beliefs and desires is caused in turn, etc.; these inquiries should proceed on a neurophysiological and psychological (and social) level. This is already an inexhaustible research agenda. There is, moreover, the well-known concern that decision theory is not even nearly empirically correct, and thereby empirical research gets even more demanding and involved. The book of Christian List tells a lot about this. A particular concern is that this mental set-up is not well described in terms of standard probabilities and utilities, as is done in standard decision theory. Our gradings of beliefs and desires may need other, vaguer representations, which need not have a purely empirical character, as they do, e.g., in prospect theory, but certainly have a rational format, too. Rationality is not bound to the standard representations.

However, I am not so impressed by such concerns. As said, in the first place, decision theory is a normative theory, designed for the normative perspective. Its rational shape is up to normative argument. The important point now is that this normative theory, whatever its final shape, is at the same time an indispensable reference point for all empirical theorizing. If you look at cognitive psychology, you will find very few researchers who think that they can pursue their business while ignoring the rational ideal. Rather, psychologists spend huge efforts into investigating the many ways in which we deviate from the ideal. Usually, they take the ideal as given; it's either deductive logic or probability theory or standard decision theory. The many alternatives philosophers are pondering about are hardly known in cognitive science. Now take my claim into account that the rational ideal, the correct normative decision theory is an open issue. This entails that seeking that

rational ideal must be part of empirical research. Empirical research can ultimately not avoid engaging in normative research. Psychologists may prefer to delegate this research, but they must not ignore it. This is my reason for the irresolvable entanglement of the normative and the descriptive perspective. The third-person perspective cannot be completely executed without the amendment of the first-person perspective.

Let me bring home this point in a still sharper way. As said, our actions are caused; of course, they are. They are mentally caused, as just outlined. They are even physically caused. I feel comfortable as a type-type identity theorist: mental states supervene on physical states (perhaps to be taken in some suitable wide sense); and depending on one's notion of a property or a state, this entails that mental states are (identical to) physical states. Likewise for causation. Causation is on all levels, not only among elementary particles. There are causal relations among mental states (and between mental and physical states), and if mental states are physical states, these mental causal relations are physical causal relations; I do not see particular problems with supervenient causation.

These are apodictic claims stirring up a philosophical snake pit. Even to start defending them is far beyond the scope of this talk. However, there is no necessity to do so. The dialectics is rather that I think I can grant my imaginary opponent the strong materialist position of a type-type identity theory; I do not have to try my luck with some sophisticated doctrine about the relation between the mental and the physical that may open some tangled argumentative leeway. I am even willing to grant that our normative point of view supervenes on our physical constitution; if our normative conceptions differed from what they presently are, the distribution of matter would have to differ, too, from what it presently is.

I may even grant that the normative facts themselves, insofar they exist, supervene on the physical facts. I am uncertain whether it is at all legitimate to speak of normative facts. Perhaps normative truths are those that are maintained in a kind of Peircean limit of normative inquiry, in analogy to the Peircean limit of empirical inquiry. If we admit talking in this way, then an a priori normative truth, as, e.g., Kant thinks of his categorical imperative, trivially supervenes on physical facts, simply because it does not depend on such facts at all. Indeed, two contingent situations that are to be normatively evaluated in different ways must, it seems, differ also in physical detail. At least it belongs to our practice of normative evaluation to equally evaluate two physically indistinguishable situations. And then this holds also for the normative truths themselves in the Peircean limit evaluation. In any case, I can grant all of this.

The point I want to make now is that even the ontological professions of an identity theorist, which I share, do by no means determine our empirical third-person perspective. They do not decide the primacy of the descriptive over the normative point of view, nor

do they undermine the ineliminability of the normative point of view. Why is this so? Let me explain the point in a slightly different way:

It is a well-known philosophical maneuver to turn ontological considerations into epistemological ones with the help of Laplace's demon. By knowing the ultimate ontological inventory of our world, the distribution of matter (at a given time) and the fundamental physical laws governing it, the demon can apparently know everything that is, and he can apparently explain every past and predict every future action and even every normative conception we will have. He seems to be the incarnation of our epistemologically perfected descriptive perspective, and there is no place for the normative perspective in that perfection. Indeed, there is no place for free will in this picture.

I think it is this powerful picture that is the profound reason why scientists who have made yet another exciting discovery about our inner workings are led to raise the problem of free will again and again. We all have this inclination. Therefore, it is important to understand how seriously deceptive this picture is. We need to understand how wildly unhuman the demon is. The point is not that in our indeterministic universe even the demon would not get far. Ontologically, we may well assume strict determinism. The point is rather that neither we nor the demon are capable of specifying the supervenience relation that is only claimed to obtain in our ontological professions and that this incapability has very different, though converging reasons for us and the demon.

For us, the problem is perhaps not so much complete knowledge of fundamental physical laws, although we still seem far from it. For us, it is rather the demon's complete knowledge of particular physical facts (at a given time) and his perfect computational capacities. Both are entirely fictitious for us. It is safe to predict that we shall never exactly compute complex molecules in quantum-mechanic terms and that, despite bold declarations of neuroscientists, we shall and can never have more than the roughest understanding of the physiological supervenience base of mental processes like, say, producing and grasping the sentence just delivered.

In particular, we have to proceed from the above simple explanation of our actions that was the only one being consistent with our having a normative perspective. We may and should specify, qualify, and amend it in multifarious ways. Of course, we also evolve our normative point of view; we seek ever better and more complete answers to our normative questions. At the same time, we thereby promote our empirical perspective; as said, our normative conception serves as well as our empirical ideal. We often do what we should, and we often do not; we often fail and often live up to our normative ideal. Every empirical theory about our behavior must respect this point, by taking the normative theory as an idealization (just like, say, frictionless motion) and by complementing the ideal by various error theories. Any empirical theory that simply neglects our normative point

of view is bound to be incomplete and inadequate.

The demon has the complementary problem. Well, not necessarily, the demon might also be an eliminativist and thus not care about supervenience. However, I take the eliminativist's prediction that our intentional idiom will eventually dissolve to be a totally incredible phantasy. Hence, if eliminativism is no option, then it will not do for the demon to know everything there is to know on the basic ontological level of physics. He is still entirely ignorant of all relevant supervenience relations. If he wants to know what water is, he must first know our notion of water; then, of course, it is easy for him to establish that water is H_2O. If he is to predict whether or not I am happy tomorrow, he must also know how happiness supervenes on all the physics he knows; and in order to know this he would first have to acquire the complex notion of happiness. Likewise for all the other mental concepts we have. In particular—this is my point—he would need to have and exercise a normative perspective by himself; otherwise, he could never grasp what our normative discourse is all about.

From both sides, we thus arrive at the same conclusion. The demon needs to have a normative perspective, even if his sole aim is to complete his empirical picture. We have the normative perspective and have to respect it as an empirical ideal even in doing empirical psychology. Hence, the agent's normative perspective is ineliminable from the observer's descriptive perspective. You cannot complete empirical psychology without engaging into normative considerations.

Certainly, I do not want to suggest that the normative part of psychology is in any way dominating. For instance, if psychologists investigate the complex phenomenon of dyslexia, any kind of normative theorizing would be beside the issue. The same holds for large parts of psychology. I only claim that one can never exhaust psychology in this spirit.

Of course, this observation spreads from psychology to all human affairs, economics, social and political science, etc. Therefore, even from the descriptive perspective one is committed to the normative perspective. As said, the two perspectives are irresolvably entangled. No perspective can be distinguished as the primary one. To conclude, my compatibilist solution of the problem of free will in terms of two perspectives stands. It does not tip over to the libertarian's or the hard determinist's side. When we claim to have a free will and that our actions are first causes, we speak the truth from our ineliminable normative first-person perspective.

Of course, this point does not exhaust the problem of free will. In a way, I have so far addressed only the problem of free action. We are free to do what we want to do. But this is of no avail if we are not free to want what we want. If our will is determined, so are our actions. So, is the will itself free? Well, certainly not in the way our actions are free. We cannot simply decide to will what we will. We don't choose desires from a menu of

potential desires as we choose meals from a menu of meals. We don't choose them at all in any good sense of choice. And thus the problem of free will reemerges.

As far as I see it is here where we find most of the compatibilist efforts. Compatibilists usually do not take the route of distinguishing two worlds, perspectives, or stances, but rather prefer to call actions free when they are caused in an appropriate way. And then a rich and most interesting dispute emerges about what may count as an appropriate way:

An action must be an action, not mere or physically coerced behavior, i.e., it must be intentional under some description so that a basic form of a decision-theoretic explanation applies, as just discussed. This is only a minimal condition, though. Extorted action is still intentional action, but it is not free. Similarly, an action must not be internally coerced or compulsory in order to be free. It must not merely satisfy rigid desires or utility functions as those of a wanton, the figure introduced by Harry Frankfurt. Free actions must rather be responsive to reasons in a wider sense of "reason" than a merely instrumental one (whatever the wider sense may be). Or the first-order desires must in turn be responsive to or under control of certain second-order desires; this was Frankfurt's central idea. Or, maybe, they have to be genuine desires in the sense of surviving cognitive psychotherapy. Such views may also be related to suitable senses of autonomy. The subject must have had the opportunity to develop her own aims or desires in a sufficiently self-determined and reflected way, or she must take a stance toward them and accept them as her own, etc. Such key words gain ever greater importance in the work of Frankfurt. Or even more directly, the (first-order) desires may have to have the right kind of content. They must conform to moral duty or even to the categorical imperative; this is the Kantian perspective. Or they must be humanly adequate in respecting our rational nature or in perfecting our virtues; this is the Aristotelian perspective. And so forth. Thus, the manner of argument often tends to be normative and not empirical; it is about the normative foundations of freedom. The claims are that our actions should be guided by moral motives, they should be responsive to reasons in a more comprehensive sense, they should be governed by second-order desires. And so on.

This is an interesting and highly relevant, though quite vague body of literature occupied with completing our basic decision-theoretic account of actions. I guess I can agree with a lot of it. However, what has all of this to do with formal decision theory? Well, I said at the beginning that decision theory is an open normative enterprise. As announced, we need to open it towards what I call reflexive decision theory. The label "reflexive" signals that we reflect on our own attitudes, i.e., go second- or higher-order. One facet of this is that we reflect on our future decision situations, i.e., our subjective views of them, the future decision-relevant mental set-ups. Earlier I said already that we take this reflexive stance, at least implicitly, when we predict our future actions not presently at issue.

However, we do not only take a predictive attitude towards our future decision situations, we also take a strategic attitude. Let me explain this point a little bit:

For 70 years economists discuss what they call endogenous preference change like, e.g., aging or becoming addicted or bored, i.e., preference change that is not the result of new information. We can foresee such change. But does such foresight influence our behavior? Yes. The main or the only proposal economists have come up with is the rule of so-called sophisticated choice. This takes a merely predictive attitude towards future preference change. In future I will do whatever is rational given my changed preferences; and this prediction constrains my present choice. I now choose that plan that is optimal among all plans compatible with this prediction, optimal according to my present preferences.

I find this proposal of the economists wanting in several respects. An important shortcoming is that they don't discuss epistemic change in this context, which may have many causes, not only changes through learning or information which are amply discussed, but also forgetting and other disadvantageous changes. If economists had attended to this, they might have learned that we do not merely predict future decision situations, we may and must take an evaluative attitude towards them, we may and must evaluate the future changes as more or less favorable or unfavorable. There are not only negative, but also positive addictions. The stereotype of the latter is getting addicted to good wine (which in this context does not count as a negative drug). Your behavior rationally depends on this evaluation. You may fight your addiction or avoid it in the first place, or you may cultivate it. You may hate your senile preferences and suppress them, or you can happily accept them. This is a kind of second-order evaluation of future decision situations. I don't know whether this evaluation can be rational or irrational. But there is no doubt that we actually have such evaluations. Clearly rational behavior depends on them. We can rationally promote favorable and try to reduce unfavorable future preference change.

This is certainly plausible as far as it goes. Another matter is to spell this out in precise detail, to exactly model this second-order evaluation and a decision rule respecting it. This is what I call reflexive decision theory. I have precisely elaborated a version of it, resulting in a decision rule which I call reflective choice. Again, we may have a normative dispute about it. But the main achievement is to have at least one explicit proposal on the table.

To be honest, it's the only proposal that formally and substantially accounts for higher-order evaluative attitudes. In deontic logic you find iterations of the deontic operator. This is not very illuminating in my view. There are attempts at grasping higher-order preferences, but they don't get very far, as far as I know. Harry Frankfurt emphasizes the importance of second-order wants or volitions and their relation to first-order wants. But neither he nor any of his scholars have worked out the logic behind this.

It should be clear by now how reflexive decision theory is related to the problem of free

will. Reflexive choice is not applicable to the wanton and his fixed and unchangeable desires. Rather, the presupposition of reflexive decision theory is that our will, though not a willful choice, is at least flexible and capable of being influenced, not only by external uncontrollable factors, but by fellow humans, and indeed by the agent herself. Our will is not simply fixed and determined. We partially determine it by ourselves, and we can do it in rational ways. In this way, reflexive decision theory formally approximates the afore-mentioned ideas of autonomy that seemed crucial for an account of free will. And to iterate, it is the only formal approximation I know of.

To resume: This is how the foundations of decision theory are connected with the problem of free will. You should accept causal and reflexive decision theory as normatively correct, and you thereby get two constructive responses to two important aspects of the problem of free will.

Let me add a final remark: The issue of free will seems intimately connected also with our practice of blame and praise and with our conception of responsibility and guilt. Indeed, the philosopher-scientists who fall prey to hard determinism are concerned—with an excited shudder, it appears to me—about the revolutionary consequences of their discoveries for these conceptions, or rather about the refutation of these conceptions. This has always been the primary line of attack of the hard determinists. The consequences are rash. Prima facie, the connection is surprising. The problem of free will is called a metaphysical problem, while guilt, blame, and praise are moral issues. What has the one to do with the other? Well, the connection lies in our normative premise and practice. Attempting to directly draw these negative consequences simply means committing a naturalistic fallacy.

To amplify: There is generally the issue how we should treat our fellow humans, whether friends or foes, helpless or criminals. This is a normative issue, indeed a moral issue. Our practice of blaming and praising, of holding people accountable for their behavior and its consequences, and of shaping our treatment of fellow humans accordingly is our answer to this normative issue. It is our normative attitude to relate all these ways of treating people to the extent in which the fellow humans were free to do what they did. Of course, our moral and legal ascriptions of accountability are most complex. Free action is only one facet. It is very difficult to delineate the causal consequences of one's actions for which one is responsible from those for which one is not. Accountability also depends on what one should and could have known or attended to, although one didn't. And so on. All of this is part of our normative practice, specified in great detail in our legal system. It is not a general practice. Our ancestors had different conceptions, and other cultures do so as well. We may consider it to be our moral progress that we connect these moral issues so tightly to the issue of free will. But this is up to normative dispute. If our opinions about

free will should change, our normative attitudes may be renegotiated. In short, this connection is part and parcel of our normative point of view. However, nothing about it follows from the metaphysical issue of free will as such.