The action-guidingness of rational principles and the problem of our own imperfections

Erasmus Mayr[1]

Abstract: The following comment discusses the supposedly action-guiding role of rational principles and the question to what extent our imperfections as human agents should influence what these principles are. According to Sergio Tenenbaum, the principles of instrumental rationality (as stated in his theory) are meant to be action-guiding rather than merely evaluative. In the first part of the comment, I look at how this action-guiding role is to be understood, especially when it comes to the pursuit of long-term, indeterminate ends. The second part of the comment raises the question of whether the principles included in Tenenbaum's Extended Theory of Rationality should be supplemented by principles for dealing with our own imperfections. I consider two possible sources for such further principles: the risk that we will behave irrationally later on and uncertainty about the effectiveness of the means we take.

Keywords: action-guidingness, procrastination, acting under uncertainty, indeterminate ends, extended actions

Sergio Tenenbaum's excellent new book 'Rational Powers in Action' (RPA, hereafter) raises a powerful challenge to mainstream theories of instrumental rationality. The challenge comes in two, mutually supporting, parts. Negatively, Tenenbaum points out that most of these theories share a number of questionable basic assumptions. This, at the very least, puts in doubt their claim to provide a general account of instrumental rationality, rather than one which can claim validity only for a severely limited field of application circumscribed by highly idealized background conditions. In particular, these theories do not sufficiently take into account the fact that most of our goal-pursuits are temporally extended and that most ends we pursue have an indeterminate nature. Both these features present major obstacles for a (i) maximizing and (ii) moment-by-moment conception of instrumental rationality. Positively, in developing his own alternative theory of instrumental rationality, the extended theory of ra-

[1] Work on this text was supported by funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 439616221 (Capacities and the Good).

ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i1.446
Submitted: 14 November 2022
Accepted: 22 November 2022

tionality (*ETR*), Tenenbaum shows how far we can get without adopting these extra assumptions. Even though *ETR* does not impose as many constraints on what a rational agent would do as, e.g., orthodox decision theory does, it still delivers a surprising amount of the results we would reach by way of the latter theory. Thus, thinking about a theory that sheds the questionable assumptions the latter theory subscribes to begins to look like a much more credible (and potentially fruitful) alternative than it otherwise would. Regardless of whether you agree with Tenenbaum's own positive theory, I think this should, in itself, be seen as an important achievement of this highly interesting book.

In the following, however, for reasons of space, I will only focus on two (I believe interconnected) issues for Tenenbaum's own positive theory. This is, first, the status of principles of practical rationality and, second, the question to what extent a theory of rationality should take into account our imperfections as human agents.

1. *The status of rational principles and their presumed action-guidingness*[2]

As Tenenbaum himself notes, there are three different 'job-descriptions' a theory of rationality could have. It could be merely evaluative, such that its "principles simply evaluate actions or mental states of the agent as rational or irrational, while making no claims about whether an agent is, or ought to be, guided by such principles" (RPA: 4). Alternatively, it could be intended to play a merely descriptive role, explaining how humans, by and large, act and make their decisions. Lastly, it can be meant to be 'action-guiding,' such that it "tries to describe the principles *from which the agent acts* insofar as the agent is rational" (RPA: 5). Tenenbaum's theory is meant to be of the third kind.

However, the way he conceives of the distinction between merely evaluative and 'action-guiding' principles is interestingly different from what most readers acquainted with the contemporary debate about rationality would naturally expect. For the latter, I take it, this distinction will be more or less the distinction between merely evaluative standards and normative principles. Merely evaluative standards need not be normative, primarily because they need not be (capable of being) action-guiding. They can be highly idealized, and there is no presumption that they cannot be appropriately applied to assess a person if she is incapable of meeting them. (The fact that I am utterly unable to hit the right

[2] In Mayr (2022), I also discuss the issue of the action-guidingness of rational principles, but from a somewhat different angle, focussing more directly on the difference between the two perspectives for assessing the agent's rationality in pursuing long-term, indeterminate ends. But there is, unavoidably, some overlap in the points raised in the following and in Mayr (2022).

notes when singing does not mean that my singing cannot be evaluated as terrible.) By contrast, for normative standards, we usually believe that it is, in some way, the person's 'fault' if she fails to comply with them because they are meant to be capable of being recognized by her and of guiding her actions (at least in normally favourable circumstances). The principle of 'ought-implies-can' seems, at least in some version, applicable when such normative, and not merely evaluative, principles are at issue.[3]

Tenenbaum's way of drawing the distinction between 'merely evaluative' and 'action-guiding' principles, by contrast, sidesteps the question of the normativity of rationality and is, instead, framed in terms of the exercise of the agent's rational powers:

> we have certain rational powers and capacities to act, and the theory of instrumental rationality is the theory of a subset of these powers. The principles of rationality are thus the principles that, in some sense, explain the agent's exercise of such powers. In the good case, a rational action is one that manifests this power. Cases of irrationality will be cases of failures to exercise the power, or improper exercises of the power. (RPA: 4)

If I understand Tenenbaum correctly, this conception of the role of rational principles plays an important role in connecting the two parts of the theory of instrumental rationality he envisages: On the one hand, the part consisting of rational principles (as spelled out in *ETR*), and, on the other hand, the part concerning the instrumental virtues. These two parts do not have completely different topics, but concern different subsets of one unified set of capacities "to pursue ends, whatever they happen to be" (RPA: 185).[4] One subset are capacities whose exercise can be explained in terms of compliance with rational principles; the second subset are those whose exercise cannot be fully explained in this way (see RPA: 176). If one believes that complying with principles of instrumental rationality is not all there is to being instrumentally rational, but still wants to hold on to the idea that there is *one single* topic of a theory of instrumental rationality, then Tenenbaum's approach of tying principles of rationality to the operation of rational powers is undeniably attractive.

But it does not, it seems to me, provide a full story about what 'action-guidingness' (in the relevant sense) really is or what is required for an action to be the result of a (successful) exercise of the rational powers in question. It is true that – together with other remarks of Tenenbaum's – it gives us *some* important indications in this direction. In particular, it seems clear that, for Tenenbaum,

[3] For standards of rationality, the connection between the applicability of the standards and the possibility of conforming to them is defended, e.g., by Kiesewetter (2017: 67).

[4] This is how Tenenbaum characterizes the "power of instrumental rationality," understood as a power of the will, in general.

the principles of rationality need not themselves explicitly figure in an agent's deliberations or thoughts when she acts on them. This, I take it, also follows from Tenenbaum's suggestion that the principle of derivation is "a generalization of explanations of instrumentally rational actions" (RPA: 45). What is required is only an understanding, on the agent's part, of the connection between her pursuit of the end and the action she performs.

> [I]f I type this sentence *because* I am writing a book, then my knowledge of the instrumental relation between typing this sentence and writing a book (…) *explains* my writing this sentence. From the first-person point of view, I infer the action (writing of sentence) from my awareness of my end of my writing the book and the instrumental relation between writing the book and writing this sentence. (RPA: 45).

This is a plausible account for many situations, especially when the instrumental action is, at this point, required for reaching the end in question. But instrumental principles, for Tenenbaum, also apply to the much wider field of actions undertaken in pursuit of indeterminate, long-term ends. And here the issue of action-guidingness becomes much trickier.

1.1. Action-guidingness in the pursuit of long-term, indeterminate ends: For momentary actions

The pursuit of (most) such ends has the following characteristic structure (see RPA: 100 ff.):[5]

> (1) I can only pursue this end by doing more specific things at some points in time. E.g., I will only manage to realize my end of reading *War and Peace* during the summer holidays if at some points in time I am actually reading some pages. But there are no specific moments at which I have to be reading any pages, because I could still do the reading later instead. Of course, at one point it will have become too late for me to finish in time. But, as Tenenbaum argues, there need not be any specific moment at which I had the 'last chance' to start (or continue) the reading such that I could have finished it in time.
>
> (2) Whenever I ask myself, during the course of the summer, whether I should start or continue reading, my current preferences at that moment and my other ends may speak sufficiently strongly against reading some pages 'just now' that it is rational for me not to start (or continue) reading then. E.g.,

[5] See also (Mayr 2022).

my desire to go swimming may each time be strong enough to make it rational not to do any reading 'just now' (even though I do not give up the end of reading the book during the summer holidays).

(3) However, when I always decide against reading 'just now,' in light of my current preferences, I will not reach my overall end – and because I have not given it up, I will turn out to have been instrumentally irrational over the whole period of time.

The interesting feature of such pursuits of indeterminate ends is, as Tenenbaum argues, that, though "[s]uccess in the pursuit of an indeterminate end depends on a series of momentary actions and is measured in terms of patterns of activity extending through time (...) there is no measure of the rationality or success of any particular momentary action with respect to the end" (RPA: 101). But this raises the question of how the principles of instrumental rationality could guide the rational person's actions in the pursuit of such ends. For the sake of simplicity, I will just focus on the Principle of Instrumental Reasoning (Sufficient), which derives, for the pursuit of some end A, the taking of some set of jointly sufficient means for pursuing A (RPA: 44). That is, this principle not merely rules out doing anything which would make reaching the end impossible; it also includes doing things which positively contribute to the end-pursuit. It is the latter element of the principle (let's call it 'Positive Contribution') which I am interested in here.

As long as I have the aim of reading *War and Peace* during my holiday, I must, if I am rational, take some jointly sufficient means to realizing that end. But neither my overarching end of reading *War and Peace* nor the principle of instrumental reasoning tells me to read some pages from *War and Peace* at any specific moment during the holidays: Whenever I am deliberating about what to do now, they leave it open to me whether to read or not. So how can the latter principle help me translate my overall aim into the "series of momentary actions" by which I would pursue it?

Tenenbaum holds that pursuing a long-term, indeterminate end brings with it a rational permission to take means to pursuing this end even when taking these means is not, at this moment, necessary for pursuing this end and even when doing so goes against what you prefer doing overall at this moment (RPA: 106). But this rational permission does not help the agent who is puzzling about whether to read another chapter or go swimming *now*. For, from the perspective of momentary decision-making (the "punctate perspective," in Tenenbaum's terminology), it is *only a permission*: The agent is not required to take advantage of it, but may always rationally decide against doing so and in favour of performing her "(Pareto) preferred momentary action" (RPA: 77).

What seems problematic here is not the fact that the principle of instrumental reasoning and the agent's long-term aim do not completely determine what the agent has to do (at least not with regard to 'Positive Contribution'), but leave her with several options. As far as rational principles are concerned, this is presumably true for all, or almost all, cases anyway: There are (almost) always different courses of action I could decide upon and still count as fully rational. If I have sufficient reasons to have coffee, but no further reasons for choosing either cappuccino or latte macchiato, then, ceteris paribus, I am rational whichever I choose. Rational principles do not tell me to choose one over the other; I am only rationally required to choose one or the other. So, the fact that the principle of instrumental reasoning does not provide fully specific guidelines about what to do is not a problem in itself.

The puzzle is rather the following: My success in pursuing my long-term, indeterminate end depends on momentary actions, and the principle of instrumental reasoning, which governs my end-pursuit if I am rational, is meant to be action-guiding (and to be so, I take it, with regard to 'Positive Contribution,' too). This suggests that this principle should be action-guiding for my momentary actions, by which I would pursue my end. That the principle should be action-guiding for such actions will seem independently plausible to many philosophers anyway: For it is a fairly widely held view that action-guidance pertains to specific situations in which to decide 'what to do now.'[6]

But in order to be action-guiding for momentary actions, it seems, the principle of instrumental reasoning must provide *some* "measure of the rationality or success of any particular momentary action with respect to the end" (loc.cit.). It must constrain in some recognizable way what I may do – even though it may not constrain it in such a way as to leave open only one permissible option. But when we have the structure in place that Tenenbaum describes for long-term, indeterminate actions, the principle of instrumental reasoning, together with my long-term end, does not seem to really constrain what I may do. Here, for *any* momentary decision about 'what to do now,' it is both rational to do something contributing to the end-pursuit or to postpone doing so. (This is the point of Tenenbaum's rejection of the claim he calls 'Culprits': RPA 136). So how are my actions rationally constrained? (This is very different from the coffee case earlier, where the principle does clearly constrain my choices, even if only down to a set of options with several members.)

This problem is aggravated by another consideration pertaining to the presumed action-guiding character of the principle. It seems that when a principle

[6] E.g. Weirich (2018: 82): "To be action guiding, rationality must target first acts in a current decision problem."

is action-guiding, the agent must be able to determine, at the time she acts, whether she complies with this principle or not (at least under normally favourable circumstances). If she was only able to determine is in hindsight, she could not herself apply this principle in making her decision and in performing the action in question. This suggests that it must be facts which obtain at the time of the action itself which determine whether the agent complies with the principle and whether – when the principle at issue is a principle of rationality – she is rational or not. It cannot be the case that this can only be determined 'post factum' or depends on new facts which only came to obtain after the action had been performed. For then, the agent could not be guided, in his deliberation and action, by this principle.

This does not mean that, in applying a principle which is action-guiding, the agent may not be called upon to use her own assessment of what is going to happen later. E.g., in determining whether she has to do X now, the agent may need to rely on her own assessment of whether there is going to be another opportunity for doing X later on. But in such a case, it seems to me, whether the agent has complied with the principle or not does not, strictly speaking, depend on what really happened later. It depends on her own expectations, beliefs (at least reasonable ones), and knowledge at the time she acted or decided – i.e. only on features concurrent with her action or decision.

However, on Tenenbaum's view, whether the principle of instrumental reasoning is violated or not does sometimes depend on developments taking place only *after* the (non)performance of the momentary action by which I (would) have contributed to the end-pursuit. This is a consequence of his principle 'Sufficiency' and becomes even clearer in his application of this principle to a case of early-stage procrastination in an extended pursuit of an indeterminate goal. 'Sufficiency' states: "For my actions to be instrumentally rational in relation to the end of φ-ing (…), it is sufficient that I φ-ed (…) through my actions in the knowledge that so doing would result in my having φ-ed" (RPA: 130). In the case Tenenbaum discusses later, he starts writing a book and, in the beginning, falls into a "pattern of potential procrastination," such as spending too much time watching football to get the job done. Realizing that he will fail to reach his end of writing a book if he proceeds in this way, he adopts some intermediate policies about how to write the book, and finally succeeds. Tenenbaum does not interpret this case as one where he initially behaved instrumentally irrationally in his end-pursuit, while he was procrastinating, and only behaved rationally from the time he adopted the new policies. Rather, he was, on his view, instrumentally rational throughout:

> [Sufficiency] determines, plausibly, that whether particular tweaks and fine-tunings add up to a manifestation of irrationality depends on whether my end *has been accomplished.* (...) If my adopting intermediate policies delivers a decent book after a certain time, *I ended up* hitting on an acceptable set of choices, one that happens to include these seemingly procrastinating actions in my first days at the job. (...) Since the outcome was good, and it was non-accidentally brought about by my acting with the aim of writing a book, there is no reason to think that my actions exhibited any kind of failure to comply with the principle of instrumental reasoning. (RPA: 196 f., my emphases.)

I must admit that I am not really persuaded by Tenenbaum's concluding assessment of this case. It does seem much more natural to me to say that Tenenbaum was irrational during the period of his procrastination and later corrected this failure on his part.[7] But, more importantly, I find the idea of action-guidingness hard to reconcile with the claim that his compliance with the principle of instrumental reasoning during this first period depended on what the pattern of his actions would be later. For, during this first period he didn't know what this pattern would be. As the case is told, during the time of procrastinating, he couldn't already rely on his finding a workable pattern later on. But then, at the time of procrastinating, he couldn't tell whether he was complying with the principle of instrumental reasoning or not – he could only do so in hindsight. And how can the principle then have been action-guiding for him at that time?

1.2. Action-guidingness in the pursuit of long-term, indeterminate ends: Over time

In the last sub-section, I have voiced some concerns about how the principle of instrumental reasoning could be action-guiding for the momentary actions by which I pursue long-term, indeterminate ends, especially with regard to what I have called 'Positive Contribution.' But Tenenbaum might respond, at this point, that the principle was never meant to be action-guiding for those momentary actions. (Contrary to what, in the last subsection, I took to be a plausible consequence of the fact that the success of pursuing the long-term, indeterminate end depends on what momentary actions I perform.) Instead, it was only ever meant to be action-guiding for the overall pursuit of the long-term, indeterminate ends *over time.* The rational agent manages to comply with the demand to 'do enough' in the time she is pursuing the aim, and is guided in this by her understanding that she has to 'do enough.'

[7] Does the principle 'Better Chance' (RPA: 215), that rational agents will choose the means with the higher chance of success, allow Tenenbaum to explain this remaining charge of irrationality? Not as far as I can see, since it will always, this principle notwithstanding, be permissible for the agent to choose his "(Pareto) preferred momentary action" (RPA: 77), and that's what we can assume Tenenbaum to have done when he was procrastinating by watching too much football.

This response would fit well with Tenenbaum's insistence that we must distinguish between two different perspectives "in evaluating actions in the pursuit of long-term, indeterminate ends" (RPA: 77): A 'punctate' one, which evaluates the (momentary) action in relation to the agent's ends and preferences at that moment (though including the 'rational permission' mentioned earlier), and an 'extended' one, which evaluates, over time, whether the agent has 'done enough' to successfully pursue his long-term indeterminate, ends. Does not the evaluation from the extended perspective constrain the agent's behaviour at least over time, since in order to be rational she must show a pattern of behaviour over time which is suitable for successful end-pursuit?

This answer would evade the first half of the problem raised for action-guidingness for momentary actions in the last sub-section. But not only would it directly lead to a further question for Tenenbaum: How is the rational agent guided by the instrumental principle in exhibiting the right pattern of behaviour, without being guided in her single momentary actions that jointly constitute this pattern? While I do not have any positive answer to this question, there is no reason for thinking that this question is unanswerable. It would just be interesting to see what Tenenbaum's own answer would be.

Furthermore, the second half of the problem for action-guidingness from the last sub-section seems to remain. Let us look again at Tenenbaum's case of early-stage procrastination in his book-writing project described in the last sub-section. If the principle of instrumental reasoning is meant to be action-guiding over time, it seems, then at the periods at which it guides the agent's behaviour, the agent must be able to determine whether she complies with the principle or not. And, we would expect, this must be true for the whole period during which the agent is meant to be guided by this principle. But, if we look at the procrastination stage, Tenenbaum's own verdict that he was not acting irrationally during that time depends on changes which occurred only after that period and which he could not in advance rely on to occur, i.e. on the fact that he later hit upon an efficient way to pursue his project. So, again, it seems that it could only be established 'in hindsight' – whether the agent, during this first period, was acting rationally or not – which seems hard to square with the supposed action-guidingness of the principle of instrumental reasoning.

If this latter problem for action-guidingness indeed remains, how could Tenenbaum react to this? There are at least two options for him here:

First, he could accept that the principle of instrumental reasoning cannot be action-guiding after all for the pursuits of long-term, indeterminate ends, at least not with regard to 'Positive Contribution,' if these pursuits share the features (1) to (3) presented at the beginning of sub-section 1.1. The principle might still be action-guiding in other contexts and for the pursuits of long-term,

indeterminate ends in other respects (e.g., when it comes to ruling out courses of action which would make reaching the end impossible). But with regard to 'Positive Contribution,' it would merely be an evaluative standard.

Second, Tenenbaum, while maintaining the feature of action-guidingness for the principle in all contexts, could modify his assessment of the agent's rationality for cases such as the early-stage procrastination case he discusses, by changing his assessment "that there is no reason to think that my actions exhibited any kind of failure to comply with the principle of instrumental reasoning" (RPA: 197). For instance, he could accept that during the procrastination period, the agent was temporarily irrational, at least as long as he could not (yet) expect that he would do later what was required for reaching his end.

My own inclination would be to go with the second option (since 'Sufficiency' seems too permissive to me) – but I am very interested to see what Tenenbaum's own stance on that issue would be.

2. *Principles for imperfect agents*

I now want to turn to the question to what extent the possible imperfections of the subjects of a theory of instrumental rationality can and should influence what principles of rationality such a theory should include. These principles are (at least also) meant to apply to human beings, and we humans are imperfect in many ways: In particular, we are not always perfectly rational, and we do not always know all relevant facts and how things will work out. Both of these imperfections are ones we are ordinarily aware of and which we should take into account in how we act. Does this give rise to new principles we should include in our theory of instrumental rationality or to a modification of old ones? In the following, I want to look at two possible sources of such additions or changes: The first is possible uncertainty about whether we will act rationally in the future; the second is uncertainty about our chances of successfully reaching our ends by the means we take.

2.1. Dealing with the risk of our own future irrationality

We cannot always rely on ourselves to be fully rational in the future. Tenenbaum allows that this may influence what we should (rationally) do. For instance, while a more perfectly rational agent would not need intermediate policies in order to pursue a long-term, indeterminate end – and would not adopt such policies because they make him less flexible – , we often have to adopt them (RPA: 193) and even sometimes have to make them strict rather than vague ones (RPA: 196). The reason for this is that, as we realize, we will not otherwise manage to successfully pursue our aim.

But the need to cope with our own deficits of rationality seems to go further, and to extend to cases where there is no certainty, but only sufficient risk of my acting irrationally later on. Take again my project of reading *War and Peace* over the holidays. I am in the first week and ask myself whether I should start reading – or rather go swimming and postpone the reading. I know that I am an inveterate procrastinator with regard to reading novels, and that on all of the following days the prospect of going swimming will be no less attractive than it is today. If I do not read now, I might still do so on later occasions: It is not impossible. But, knowing me, it is not too likely, either.[8] More realistically, I will be as little motivated to read as I am now and procrastinate further. Under such circumstances, it does seem to display a lack of instrumental rationality to postpone the reading to these later occasions, since I cannot rely on my taking advantage of these occasions.[9] Even though, in this case, I may still eventually reach my aim (because, e.g., unexpectedly, I later break my leg and cannot go swimming any more[10]), there does seem to be something rationally criticisable about the way I pursued my end. For I knowingly risked failure in the pursuit and let success too much slip 'out of my control.' While it was not 'just luck' that I succeeded, since, after all, I did the reading myself, I made myself too much a hostage of fortune to escape rational criticism. Thus, protecting ourselves against and reducing the risk of our own future irrationality (by reducing the chances for it) seems to be required by instrumental rationality. (How much we should do so depends, of course, both on how well our own rational capacities work and on how important the end in question is for us.)

(Interestingly, in a different context, Tenenbaum seems to accept the underlying idea that we should take into account not just the certainty, but also the risk of our own future irrationality (RPA: 179). But he does not pursue the idea of how this should shape the pursuit of our ends, beyond its speaking against taking up certain activities in the first place.[11])

[8] For a discussion of such cases see also Mayr (2022).

[9] Can Tenenbaum explain this by appeal to his principle 'Better Chance,' that, in cases of uncertainty of success, the rational agent will take, ceteris paribus, the option offering the better chance of doing X? (cf. RPA: 215). I don't see how he can. First, as stated above (fn. 7), 'Better Chance' does not seem to help in cases where the agent pursues long-term, indeterminate ends and, on each particular occasion, prefers doing something else to taking the means contributing to doing X. Second, 'Better Chance,' as stated, only covers cases where "doing X is more likely to result in A's F-ing than doing Y" (RPA: 215). This is not true in the case discussed above: Whether I read some pages today or tomorrow, the contribution to successfully finishing reading the novel will be exactly the same.

[10] Would reaching the aim in such a case be a mere accident – in which case Tenenbaum could explain the charge of irrationality by appeal to his nonaccidentality condition (RPA: 137)? It doesn't seem so, since, when I read all parts of the novel intentionally and in the knowledge that this will lead to my having read the whole novel, it is no mere luck or accident that I end up having read the whole novel.

[11] Another way in which this idea might get a foothold in his theory is a comment he makes in

The need to reduce this risk may also be the reason why sticking to earlier decisions and policies is rationally required more often than Tenenbaum allows for. This is suggested by an illuminating discussion of Michael Bratman's proposed solution to Quinn's Self Torturer case. Bratman argues that, when the agent has settled in advance on stopping at some point (rather than continue minimally increasing the pain in exchange for more money), she should rationally stop at this point. For "She can ask: 'If I abandon my prior intention to stop at [a_{25}], what would then transpire?' And it seems that she may reasonably answer: 'I would follow the slippery slope all the way down to [a_{1000}] [the last setting].'" (Bratman 1999: 81, quoted after RPA: 109 (incl. the added changes)). Tenenbaum responds that this reasoning only works when the agent "has reason to believe that she will either stick to her plan or continue to the end of the slippery slope. (...) But why should she believe that?" (loc.cit.) Indeed, if the agent can rely on herself to stop before the pain becomes too intense, then there seems to be no reason for her to stop at the planned point. But, on the one hand, given the unbearability of the pain when she doesn't stop in time, even the risk of not stopping, if it is significant enough, speaks strongly in favour of 'playing it safe' and stopping at the pre-settled stage. And I suspect that this is the scenario that Bratman envisages: i.e., that there is a real danger of the agent's not stopping later on. On the other hand, even when the agent can be confident that she will still 'stop in time,' this is strictly speaking not a case where she first rationally adopted a future-directed intention or plan that she may now rationally disregard.[12] It is rather a case where adopting the plan was not needed in the first place. We realize that the problem that adopting the plan was meant to solve did not exist at all and that we therefore can give up this plan now. But this is not a case of being permitted to abandon an intention that, at the time, was formed on a sufficient rational basis. In fact, Bratman himself may accept that not stopping at the pre-determined point is rationally permitted here, since, as he suggests, the requirement to stick to our future-directed intentions is plausibly restricted to cases where there is "both initial, supposed support for that intention and constancy of view of the grounds for that intention" (Bratman 2012: 76).

I take Bratman to understand the situation under discussion to be one where the problem originally existed and has not disappeared in the meantime. (Cf. his description of the case as one where "His prior decision to stop at [a_{25}] was his best shot at playing the game without going all the way," Bratman 1999: 81,

passing on the necessity of the "temporal management of our ends" (RPA: 124).

[12] Unless the agent has realized in the meantime, i.e., only after adopting the plan, that she can trust herself to stop in time; but that is, as far as I understand it, not the situation Bratman or Tenenbaum envisage.

quoted after RPA: 109 (incl. the added changes).) Insofar as this is true, stopping at the pre-envisaged point does indeed seem to be the choice recommended by instrumental rationality – notwithstanding the fact that, as Tenenbaum rightly points out, the antecedent is not always true, and then stopping at this point is not always rationally required.

These kinds of cases suggest that there may be further rational principles, not included in *ETR*, which apply to us because we must cope with the imperfections of our own rationality. Maybe such principles even require intention-persistence under specific circumstances. Accepting this need not really be a problem for Tenenbaum, though, as long as these principles are not basic ones we would have to add as such to *ETR*, but derivative ones. But I wonder whether such a derivation is possible for all plausible principles for dealing with our own potential irrationality. My guess is that we will have to add at least some basic principle which prohibits running too high a risk of failure in our end-pursuits by relying too much on ourselves to do what is required later on.

2.2. Uncertainty of success

Our own future irrationalities are only one imperfection of ourselves we have to cope with. Another one is lack of certainty about whether we will successfully reach our ends by the actions we take as means. This brings us to Tenenbaum's discussion of the cases of action under risk in chapter 9. Tenenbaum's treatment of these cases rests on his view that doing X is not the same thing as successfully trying to do X. Instead, when an agent realizes that she cannot take "means she knows to be sufficient for her ends of φ-ing [she] must revise her ends, and among the possible acts still available to her will be the act of trying to φ. But for our purposes, trying to φ is an essentially different action from φ-ing" (RPA: 210).

Tenenbaum's latter claim about the nature of trying will not seem compelling to all readers. Many theorists, I take it, will want to insist that we have a continuum between doing X in the knowledge that you can do it, and trying to do it, because full certainty can never be achieved anyway, and the only possible difference between the two cases is one of degree of certainty. However, Tenenbaum's point seems to me, in a crucial respect, correct: Lack of knowledge that I can do F can (and often does) change the nature of what I am doing.

But it seems hard to accept the consequence Tenenbaum draws from this, namely that we can draw no inference as to the instrumental rationality of an agent who, on learning that the envisaged means may fail to lead to the aim, (and who can therefore no longer decide to reach this aim, but only decide to try to do so) does not (even) try to reach this aim. "Suppose I was on my way to meet Mary at her office, and I now realize that Mary might not be in her office. (…) Nothing about my basic given attitudes here determines whether I will, insofar as I am

rational, engage in the action of trying to meet Mary at her office" (RPA: 210).

This does seem too permissive: If meeting Mary was important enough for me, and if there is still a way to try to meet her which is not too costly and has a reasonable chance of success, then my realization that my intended means is not 'foolproof' hardly allows me to drop my project altogether and not even engage in an attempt to meet her. The jump from 'doing F' to 'trying to do F' may (often) involve a change in the nature of what I am doing, but with regard to my instrumental rationality, the difference does seem to be one of degrees, not a fundamental one, and a demand of instrumental rationality to do F will, maybe slightly weakened, regularly 'transform' into a demand to try to do F when I realize that I cannot be certain whether my means will be successful or not.

Interestingly, Tenenbaum might be able to reach this – to my mind, highly plausible – result by a different route, at least for agents who reliably recognize the reasons which apply to them. Since he subscribes to the 'guise of the good' view of the pursuit of aims, there will, for any end we are pursuing, have to be reasons which speak in favour of doing so, when our beliefs about our ends are correct. These reasons may regularly also support trying to reach this end when one lacks knowledge about how to reach it and therefore cannot decide to do F. Trying to F may be different from and only a 'second best' compared to doing F, but if the latter has value, the former may, normally, have some (at least derivative) value, too. If this is true, I will indeed normally be rationally required to try to do F, when I cannot decide to do F for lack of relevant knowledge, in order to comply with those reasons. This, however, will not follow from principles of *instrumental* rationality, but rather from the (substantive) reasons in favour of doing F in the first place. To me, this latter feature seems to be a crucial drawback of the alternative explanation. We would – and should – expect it to follow from principles of instrumental rationality and from my 'basic given attitudes' in the situation that I should try to do F in cases of (non-dramatic) uncertainty when the aim is of sufficient importance to me.

These considerations suggest a further addition to the principles of rationality included in *ETR*, which would allow us to infer, when we realize that we don't know any sufficient means for doing X, that we should (under the specified circumstances) still try to do X (or adopt the end of trying to do X).[13]

Erasmus Mayr
Institut für Philosophie, Friedrich-Alexander-Universität Erlangen-Nürnberg
erasmus.mayr@fau.de

[13] For very helpful discussions of an earlier draft, I am indebted to Stefan Brandt and Christian Kietzmann. For comments on the proofs, I am indebted to Dorothee Bleisch, Patrick Faralisz and Ufuk Özbe.

References

Bratman, M., 1999, *Faces of Intention*, Cambridge University Press, Cambridge.

—, 2012, "Time, Rationality, and Self-Governance," in *Philosophical Issues* 22: 73-88.

Kiesewetter, B., 2017, *The Normativity of Rationality*, Oxford University Press, Oxford.

Mayr, E., 2022, "Rational Powers in Action: Instrumental Rationality and Extended Agency, by Sergio Tenenbaum," in *Mind* https://doi.org/10.1093/mind/fzac015

Tenenbaum, S., 2020, *Rational Powers in Action*, Oxford University Press, Oxford.

Weirich, P., 2018, "Rational Plans," in J.L. Bermudez, ed., *Self-Control, Decision Theory, and Rationality: New Essays*, Cambridge University Press, Cambridge: 72-95.