Rational Powers in Interaction: Replies to Paul, Andreou, Brunero, Mayr, and Haase

Sergio Tenenbaum

I can hardly express my joy and gratitude in having such excellent philosophers pay such careful attention to my book. I am not surprised, but very pleased, that these are all fantastic comments (I am ashamed to confess that some part of me wishes that they were less challenging and easier to respond…). I certainly can't do justice to all of them here, but I'll try to answer at least some of them (I'm sure I would have amazing responses to all the other ones if I had a bit more time and space). Often when one receives such a large set of comments, one expects that one will spend some amount of time dispelling confusions and correcting mistakes. I am lucky enough that I have no need to do this here; all the commentators correctly describe my view, and very often they do a better job than I could do myself explaining them. So, I am in the fortunate position that I can go directly to the points of contention when discussing the comments.

Paul

Paul raises some significant challenges to my treatment of uncertainty. I first should immediately grant that this is an aspect of the theory that I hope to develop in more detail in the future; the book mostly tried to show that *ETR* had enough tools to approach the issues, and that an adequate treatment of risk and uncertainty contexts was within reach. But I am under no illusion that this is a fully developed discussion of the topic.

Let me start by trying to, as it were, contain the damage that Paul's criticisms might end up doing to theory. As Paul correctly points out, I take cases in which no relevant false information or uncertainty is involved to be the paradigmatic cases for a theory of practical rationality; the theory is first formulated on the assumption that the agent knows all the relevant aspects of the agential context in question (in particular, on the assumption that the agent knows that her ends can be achieved by her efforts, and she knows how to employ sufficient means to realize her ends). And although I don't favour allowing cases involving false

Submitted: 3 December 2022
ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i1.448
Accepted: 29 January 2023

beliefs to count as successful exercises of one's instrumental rational powers, I claim that not much hangs on this: we could reformulate the main principles of the theory in terms of belief and accept that instrumental rationality could be exercised not only in employing sufficient means to one's ends, but also in employing the means that are (possibly falsely) *believed* to be sufficient to one's ends. Paul thinks that if she's right about the difficulties that the theory faces in cases of uncertainty, this will also challenge the idea that the central principles of the theory should be formulated in terms of knowledge. But I think these are essentially different issues. After all, one could, for instance, formulate decision theory in terms of knowledge of (or belief about) probability distributions rather than credences. And, on the other hand, reformulating *ETR* principles of derivation and coherence in terms of beliefs (and thus allowing actions in light of false belief to be manifestations of our instrumental rational powers) would be of no help in dealing with cases of risk. For this reason, I am somewhat cavalier about intuitions about rationality in light of false beliefs ("if you are attached to these intuitions, change a couple of words in the principle," I would say), but I think it is essential that the theory can account for plausible judgments about instrumental rationality in risk situations. Most of Paul's concerns are indeed on how the theory treats cases of risk and uncertainty, but it might help if I start by first explaining why I think cases of risk and uncertainty are a much more serious threat to the theory and essentially different from the case of acting in light of false beliefs.

ETR's Principle of Derivation tells me, roughly, to take sufficient means to my end. When we reformulate this principle to something like "take (what I believe to be) sufficient means to my end," we get a very similar principle. We need to make a few adjustments to make sure that we are not enjoining the agent to change their beliefs when the going gets tough, but if we get this right the belief version of the principle will guide the agent to act in exactly the same way as the original version in cases of knowledge.[1] But the same is not true if we try to formulate the principle to accommodate uncertainty; there is no similar tinkering we can do to the principles to extend its reach to risk or uncertainty contexts. We could try reformulating the Principle of Derivation as follows: "take the means that are most likely to bring about the end." But this version of the principle is obviously invalid; given my other ends, it might be perfectly rational not to take the most likely means to some end. Perhaps the better route is to put forward a much weaker version of the Principle of Derivation such as: "take what *might* be sufficient means to my end." The original Principle of Derivation

[1] Arguably this version changes nothing in terms of how the principle guides the agent, but only in terms of a third person evaluation of the agent based on this principle.

was existentially quantified ("take *some* sufficient means"), but an existentially quantified version of this revised Principle of Derivation is obviously too weak: a principle that enjoins me to engage in *some* action that might be sufficient means to my end would give us at the best the anemic sense of "trying," but really not even that. In my pursuit of acquiring a house, it would suffice to do anything that I have a non-zero credence that it would ultimately lead to my having a house to count as pursuing this end rationally (it would be enough to strike a conversation with a rich person who, for all I know, would take a liking to me and offer to buy me a house). But a universally quantified version of the principle does no better: I am not required to do everything that *might* result in the success of my pursuit. This would be no different from first attempt to reformulate the principle.

It seems that any plausible version of the Principle of Derivation in this context would have to relativize it to the pursuit of other ends ("make it more likely that you φ without jeopardizing your pursuits of …"); I am not sure how this would be done without in effect jettisoning the principle in favour of something akin to orthodox decision theory. [2] However, if my arguments in the book are correct, the costs of moving to such a theory of instrumental rationality are prohibitive.[3]

My own view is that the introduction of risk or uncertainty changes the nature of the action; once you realize it is not within your power simply to do something, the nature of what you are pursuing has changed. At the very minimum you're no longer φ-ing, but trying to φ. In fact, in ordinary parlance, I can no longer say "I am driving to Rome" when I become uncertain of whether I'll be able to make it there (if, say, road blockades might have made the city inaccessible); I must now say "I am trying to drive to Rome," or something like that. But whether or not ordinary language confirms this view, the above facts give us enough "independent motivation for the idea that trying to E is a substantively different action from doing A;" that is, if what I say above is correct, the rational principles guiding the agent who is φ-ing cannot be guiding the agent who is trying to φ in exactly the same way. On the other hand, it is not enough for a theory of practical rationality to note this fact and simply postulate that actions in risk contexts cannot share an act type with actions done "under knowledge." The claim that these are different act types must also explain why certain principles guide, or seem to guide, a rational agent in risky contexts. I argue in the book that this is precisely what an understanding of trying within the *ETR*

[2] Note that it is also hard to see how any such proposal for revising the Principle of Derivation would be compatible with accepting some version of the Principle of Coherence. After all, we can rationally aim at incompatible objects that we are uncertain about its realization through our efforts: I can try to both go to Harvard Philosophy and to go to Yale Law next year if the chances of either happening are low. So, if this path is blocked, how else can we extend the theory to risk contexts?

[3] See chapters 3 and 4.

framework can deliver. I put forward an understanding of (non-anemic) trying in terms of its internal end: trying takes the object of trying itself as good, and this fact generates for trying to φ a partial preference ordering relative to this pursuit. Just as my pursuit of building a house (typically) generates a preference ordering relative to this end (relative to this end, I prefer to use concrete rather than straw for the house's foundation), trying to φ generates a preference ordering in which, for instance, I prefer to take means that are more likely, over means that are less likely, to result in my φ-ing. Trying, according to *ETR*, behaves like any other intentional pursuit in determining the actions of an instrumentally rational agent, and its internal structure generates constraints that mimic the constraints of decision theory precisely in the contexts in which decision theory is most plausible (and part ways with decision exactly in the contexts we tend to resist its prescriptions). I do not want to rehash the argument for the claims here, and I would be misleading the reader if I did not acknowledge that there is much more work to be done. The main purpose here was to provide a relatively concise answer to Paul's challenge which can be put in a slogan that is certain to rally the troops: "we treat φ-ing and trying to φ as different actions in order get a better account of the different ways in which rational principles guide us in the contexts of knowledge and uncertainty." I can almost see myself holding a sign with these words while marching on the streets.

Paul also challenges my account of the instrumental virtues and vices in the book. In particular, Paul complains that my concession that certain patterns of irresolution manifest an instrumental vice runs afoul of the Toleration Constraint. But, strictly speaking, I don't think this can be true. Roughly, the Toleration Constraint requires that a theory of instrumental rationality be as permissive as possible in terms of which ends it allows agents to pursue. However, the claim that there are instrumental vices in the book should not render any particular action or pursuit irrational, so it could not run afoul of this constraint. In fact, instrumental vices are ways in which people fall short of ideal rationality *without acting irrationally*. But, of course, the theory might still violate the spirit, without violating the letter, of the Toleration Constraint.

I hope it is fine to mention here an embarrassing fact about myself: I hold pens (or any writing utensils) in a non-standard way. Typically, this does not affect my capacity for writing: I can generally produce legible words in a paper by moving a pen with this non-standard grip. On the other hand, I cannot use fountain pens; the words will be too smudged to be legible. This is a limitation to my writing capacity: I can only write with certain kinds of pen. But it does not need to generate any failed writing on my part; as long as I stay away from fountain pens, I'll be as competent a writer as anyone else. Having an instrumental vice bears a similar relation to instrumental irrationality: it limits your capacity to pursue ends, but it does not imply that you ever pursue an end irrationally. Perhaps the existence of instrumental vices is of limited interest if we assume that instrumental rationality is the only form of rationality:[4] it would at most register that, for instance, cowards quickly give up on pursuits that they perceive to be dangerous insofar as they are rational. But if the final view of rational agency requires or enjoins the pursuit of certain ends, then cowardice would be something that would potentially condemn me to live a life in which I fail in some way: either by not pursuing what is good but dangerous or by akratically failing in pursuing those dangerous goods. Paul does anticipate this response on my part, but she thinks that a limited capacity to pursue ends need not be a vice: an agent who is akratic might profit from also being a coward if courage would lead her to pursue the lesser good (if, say, only cowardice is stopping her giving in to her temptation to rob a bank). But I don't think this shows that the incapacity as such here is not a defect. When failures multiply, it might be that one failure makes the other failure less unfortunate. If I am also allergic to ink pots, my incapacity to use fountain pens might save my life; yet, fortunate as this incapacity is in these circumstances, it is still a limitation to my ability to write on paper.

Andreou

There is much that Andreou and I are in agreement, and I find her views on these topics very compelling. Perhaps, the crux of our potential (why "potential" will be clear in a moment) disagreement is that Andreou finds that there is more structure in cases like Quinn's self-torturer than I do. In particular, Andreou proposes that categorical and relational appraisals play different roles in the theory of instrumental rationality. While, for instance, we can judge meals in terms of how they compare to each other and rank them from best to worst, we can also make various categorical judgments in assessing them: meals can be awful, bad, subpar, ok, pretty good, excellent, or superb. So far, I am indeed very much in agreement. As I said above in my reply to Paul, certain ends generate preference rankings that are internal to the pursuit of this end. As long as we are restricting ourselves to the end of a tasty meal, we can form a preference ranking relative to this end. I also think it is important that the ranking created in this manner will allow the agent not only to make relational appraisals, but often also categorical ones. A meal at Geranium in Copenhagen is not just better than a meal at my local taqueria. The meal at my taqueria is pretty good

[4] However, as long as there ends that are better (or better for you) to pursue, this will be enough to make an instrumental vice relevant to your life: even though you might be perfectly rational, it'll likely be a worse life due to this vice.

while Geranium is superb (or so I am told). These different categories play an important role, for instance, in end revision. The pandemic is over and I want to use the money I saved for a superb meal. Since there are no superb restaurants near me, I start planning my trip to Denmark. But I quickly realize that the expenses are too high. Given that I also want to buy a new car, and I want to have enough money to retire very comfortably, I cannot go take the flight to Copenhagen and foot the bill at Geranium while maintaining these ends. So, I need to revise one of these ends; perhaps, I will settle for eating an excellent meal at the new Japanese restaurant that is walking distance from my house. Or decide that as long as I retire moderately comfortably, that's good enough for me.

Let us now look at a slightly different situation. Suppose I make similar plans: I want to buy a pretty good car, have a superb meal, and leave enough money for a comfortable retirement. I check prices and my investments, and, fortunately, I can pursue these three ends. As I am about to buy my car, I realize that I can get an excellent car for the same price (no similar adjustment can be made to my pursuit of the other ends). Now, in the words of the book, I have a *Pareto preference* for buying the excellent car: it provides a better realization of my end of acquiring a car, without infringing on the pursuit of my other ends. As long as this decision does not implicate any other (indeterminate) end of mine, *ETR* says that I must buy the better car, and this is, of course, very intuitive. So far, we are in agreement. And I find Andreou's insistence on distinguishing between categorical and relational proposals really important in this context. I think we do part company in an important juncture, though I am not completely sure. I suspect that Andreou is committed to the view that these categorical appraisals will necessarily (often?) apply to situations in which two ends apply *considered as a whole*, and I am skeptical about this. Let me try to explain what I take the disagreement here to be and why I remain skeptical about Andreou's approach in this case.

Here is one way we could conceive of how the self-torturer settles on a certain stopping point according to *ETR*. The self-torturer might have had at first the end of making as much money as possible and living an absolutely pain-free life. But once she is offered the deal of making money in exchange for moving up the settings of the torture machine, she needs to revise at least one of these ends. We can now rely on categorical appraisals (as Andreou suggests) in specifying the way that the self-torturer revises her ends. On the side of the money, she could have "making a little money;" "making a significant amount of money," etc. On the side of pain, she could have "no pain at all," "no more than an insignificant amount of pain," etc. Let us say that "making a significant amount of money" is compatible with "just a little pain." Our self-torturer now might revise her ends in this way and rationally choose a setting which are acceptable realizations of each end. She will stop at a setting in which she makes a significant amount of money but does not suffer more than a little pain. Of course, had she chosen to revise her ends in a different way, she would choose differently. At any rate, on this description, what she chooses is still the acceptable realization of her ends, but since Andreou thinks that this is not sufficient, I assume she thinks that there are categorical appraisals that are relevant beyond the ones I described. So, perhaps even when so specified (or if my ends are specified in even more indeterminate ways like "enough money" and "not much pain"), it will be possible to classify my options in different categorical appraisal groups. But I am not sure that this can be done. After all, on what would these classifications be grounded? Andreou says that they would vary from subject to subject so perhaps different strengths of desire for money and pain avoidance would generate different classifications? But I am skeptical that there is any notion of strength of desire that is relevant for a conception of instrumental rationality.[5] But perhaps what Andreou has in mind is closer to what I propose here than I am making it sound, in which cases our views are not so far apart after all.

In a nutshell, Andreou thinks that in considering the options that satisfy different ends in different ways, the categorical appraisals will apply to the choice situation directly. On the other hand, I think categorical appraisals are only relevant to the internal ranking generated by each of our ends. Thus, these appraisals apply to the choice situation only insofar as they can be relevant in determining what count as an acceptable realization of the end. These might in the be little more than notational variants.

I will briefly address the other issue Andreou raises. Andreou raises an interesting challenge at least to my stronger claim of nonsupervenience based on a distinction between whether I am irrational *in* the moment or *at* the moment. And although my failure might be not contained within what can be captured in an snapshot of an instant, it might be happening *at* the moment, since I might be doing something at the moment, like frittering away my life, that reaches beyond the moment (my frittering away my life cannot be fully contained in a moment) but suffices to qualifies me as irrational at the moment. My first reaction is that retreating to the weaker supervenience claim that Andreou identifies would not cause major damage to the view. But I am not sure this is necessary. Andreou is relying on the nature of action in progress, such that I can already be engaged in, say, crossing the street, before the action is completed or even if it is never completed. Both Andreou and I, following Thompson (2008), think that action in progress is central to our understanding of agency. However, it is also important to notice that not everything that we can say in retrospect that I have been φ-ing (because at that point in time is true that I φ-ed) is something

[5] I argue against such use of strength of desire in chapter 3.

that I could be in a position to say that I was φ-ing at an earlier time. Some actions do not generate an imperfective paradox. Even though I can be crossing the street without having ever crossed the street, I cannot be said to be killing someone (except metaphorically) if the person does not die at the end of the process. "Frittering away my life," I submit, is of the latter kind. And, in particular, I would not have been frittering away my life at any moment, if I did not "successfully" fritter away my life at the end. At any rate, I think it is correct to say that, unlike crossing the street, "frittering away my life" is not an action in progress that is accessible to me at the moment that say, I am watching TV instead of writing my book. Admittedly, if I were to do nothing but watch TV the rest of my life, it will be true that I had been frittering away my life;[6] but unlike the case of "crossing the street," if my life does not end up being "frittered away" (if, say, I clean up my act after a while), it is also not true that I was frittering away my life at the time I was watching TV. Thus, at the time I am watching TV early on, it is not settled that I am frittering away my life, and thus it cannot be something that I am accountable for.

Brunero

Brunero does a great job of characterizing in which ways my view depart from the received view. But he also raises important challenges to it. First Brunero thinks that my rejection of principles governing intentions, like the means-ends intention coherence principles (MECs), has counterintuitive consequences. Brunero correctly points out that the account of gappy action is supposed to capture some of the verdicts of irrationality often attributed to MEC by allowing for gaps before any proper parts of the action start and by relying on the fact that *ETR COHERENCE* applies to these gappy parts of the action as well. But he argues that this move could not help for cases in which an agent never moves from intending to φ to taking any means to φ, and that in such cases it is still intuitive to say that an agent who violates MEC is irrational. Since Brunero himself thinks that there is an easy patch here, most of my comments are on the second part of the paper.[7] But I do want to resist the claim that there is anything that is clearly left out by the theory I propose. The putative problematic cases

[6] Though, of course, my pursuing the end of refraining from frittering away my life is available to me, but it will be another end that, if I fail to realize, my irrationality might not be attributable to any moment at which I was being irrational relative to this end.

[7] A slightly different adjustment that would be better at preserving the spirit of the view would say that the process of φ-ing was interrupted before any proper bodily manifestation of the process had taken off. I don't think this would conflict with the restriction of the aim of the book to bodily actions, but it is beyond the scope of this response to establish this point.

are cases of what Davidson (2001) calls "pure intending." In the swimming race case, if the agent, for instance, turns down a friend's request to go to a party that day, or schedule a meeting at a different date, due to the conflict with the swimming event, then the agent has already taken means in the pursuit of the end of swimming, so the account has no problem in explaining that the process of swimming in the competition has started.[8]

Is the agent who intends to φ, but never does anything in light of the fact that they intend to φ, irrational when they fail to intend the means to φ-ing? I am not sure much hangs on what one says here, and as Brunero recognizes, there are difficulties in formulating a precise principle that will allow for permissible delays in having the instrumental intention and so forth. Yet, I do feel the pull to think that there is something incoherent about Brunero's swimmer. But I am not confident that we need to think of this as a case of practical irrationality even if we want to do justice to this intuition. Wallace (2001) and others have suggested (roughly) that instrumental incoherence is just a form of theoretical incoherence: namely, it is simply the incoherence among the *cognitive* attitudes implied in intending. If intending implies belief,[9] then incompatible intentions are incoherent because they imply the existence of incompatible beliefs. Although I obviously don't think that instrumental incoherence can be reduced to theoretical coherence, I think in some cases we are willing to ascribe irrationality to the agent in such cases because of the implied incoherence in these beliefs. I think this is supported by the fact that once we allow cases in which I intend to φ without believing I will φ, it seems coherent to intend the end and not intend the believed necessary means. I think it is plausible to say that I intend to reach the top of Everest even when I do not believe I will (or in some views, even if I believe I won't) because people with my level of fitness often fail to reach the top. But now I also believe that buying a very expensive equipment is necessary for my reaching the top. However, believing so is (arguably) compatible with my thinking *I might be wrong about it*. But now I could keep my intention and buy the budget equipment and still intend to reach the top while hoping that I am wrong that the expensive equipment is a necessary means to success. Needless to say, linguistic intuitions differ here, but the point is that I can have 'reaching the top' as my aim, believe that buying the expensive equipment is a necessary means to it, and coherently not buy it. These are, of course, difficult issues and there is a burgeoning literature on this topic. But I just want to point out that the relevance of these putative counterexamples to *ETR* can be challenged on independent grounds.

[8] See below for more details on this point.

[9] Not a view that Wallace accepts. I am assuming it momentarily for ease of presentation.

Brunero raises the case of Principled Patty as a possible challenge to the *ETR SUFFICIENCY.* Here too he correctly anticipates my response to the purported counterexample, so it would be best if I approach this case as a challenge to my views on the role of trying in the theory of practical rationality. I think Brunero's challenge here can be fruitfully framed as continuous to Paul's, that is, as a concern about whether the discussion of contexts of risk and uncertainty can deliver what it promises, especially through its reliance on distinguishing trying to φ and φ-ing. So, I'll examine directly the objections that he has to treating cases like Principled Patty and others using this framework.

Let me start with the claim that this treatment distorts the nature of the agent's instrumental reasoning by "having *trying* as *the object of pursuit*." It is important here to clarify one aspect of *ETR*. As Brunero correctly points out, *ETR* takes intentional action to be the basic attitude grounding the exercise of our rational powers and also as their "outputs." So, the basic manifestation of this power is when I am φ-ing as a means to ψ-ing. I have chosen to represent the intentional actions in question as "pursuing the end of φ-ing" for two different reasons. First, I am largely in sympathy with a view defended by Thompson (2008), Ferrero (2017), and Moran and Stone (2009). Suppose I aim to make an omelette. I can start by checking the fridge for missing ingredients, then going to the store, then start breaking eggs, and so forth. When did I start making an omelette? Most of us would hesitate in saying that I was already making the omelette when I opened the fridge, but these authors (very roughly) argue that there is a continuous process that has already started at my opening the fridge and that the breaks we make here (that we count, say, the breaking of the first egg as the beginning of making the omelette) are somewhat arbitrary, or grounded on reasons that have little to do with the metaphysics of action.[10] Although I am in full agreement here, for the purposes of the book, I am committed only to the weaker version of the view: for the purposes of practical rationality, we should regard this process as a single process that has already started at least when I started walking towards the fridge. Using this formulation allows us to incorporate this point without doing any violence to the English language, as there is little disagreement that I was pursuing the end of making an omelette as soon as I started walking towards the fridge. Secondly, this "notation" allows us to separate attitude ("pursuing the end") and content ("making an omelette"), and thus helps in presenting intentional action as an attitude. However, exactly for that reason, "pursuing the end of φ-ing" is not going to appear as such in deliberation, as the object of deliberation is always only the *content* of the attitude. When I deliberate about whether *p*, I do not take as

[10] See Anscombe (2000) for a similar claim.

a premise "I believe that if *q* then *p*," but "if *q* then *p*" itself; the attitude is "backgrounded" (borrowing an expression from Pettit and Smith 1990). So, indeed, Brunero is correct that "pursuing the end of trying" does not figure as the object of the agent's reasoning. The object is simply "trying to φ." And it is unproblematic, and I think correct, to say that "trying" does appear as part of the object of the pursuit: if I am running a race as fast I can and I conceive of the object of my pursuit as "winning the race," it seems that I would have to engage in some morally dubious action (or give up my end) when someone points out that I can only ensure that I win the race if I off my competition. Here I would naturally say that what I am doing is "running as I far as I can and hoping it will work out" or "trying (as hard as I can) to win the race."[11] I think these considerations also help answer Brunero's concern that it would not be enough that I show that the end of trying to get a hire is the end pursued in the case of uncertainty but that I need to exclude also the end of getting a hire as the end pursued. Brunero thinks this is particularly implausible in the case in which the Dean does let me proceed with the hire. But even if I got the hire intentionally in such a case (which many philosophers would deny), *ETR* is only committed to the claim that "getting the hire" could not be the action *guiding me* in the pursuit of means; it would not be a basic attitude for a theory of instrumental rationality. An analogy might be helpful here. If my Leader gives me the order "get a hire from your Dean!" I would have to explain to my Leader that I don't know if I can get a hire from my Dean, and a reasonable Leader would revise the order to "Well, then try your best!". In the realm of instrumental practical reason our ends are like our Leader whose orders we follow by taking sufficient means to it.

Brunero also thinks that understanding risk contexts in terms of "trying" will cause problems for *ETR COHERENCE* as trying to φ and trying to ψ are not incompatible even when φ-ing and ψ-ing are incompatible. But at first sight this is a welcome consequence of the view, as in these cases of uncertainty, we do often try to do incompatible things. In Bratman's celebrated video game case (Bratman 1987), an agent wins the game if they hit either target A or target B, but if they are about to hit both the game shuts down, no target is hit, and the agent loses the game. But given their limited skill and the relatively low likeli-

[11] It is worth mentioning that unlike in the case of "crossing the street," "trying" is always intentional, and, arguably, unlike "making an omelette," we can say that you are trying to φ as soon as the process that ends in you having tried to φ begins. So, there is no difference between saying "pursuing the end of trying" and "trying" and thus "pursuing the end of trying" will always seem like another activity, but it's really no different than trying (this is also why in the book, when moving to trying, I generally say just "trying" rather than "trying intentionally" or "pursuing the end of trying"). For similar reasons, I think "trying to try" and "trying" do not describe two different actions, but nothing in the book commits me to this view.

hood of hitting either target, the agent tries to hit target A and tries to hit target B, knowing full well that it is impossible to hit both. I think Brunero would accept this point, as he never suggests that *ETR COHERENCE* should apply in full generality to such cases, but rather he provides an interesting putative counterexample that would show that *ETR COHERENCE* cannot capture the incoherence in this particular case. The case in one in which I am trying to get a hire and if there is a hire, as the Chair, I'll be in the search committee. It seems that *ETR COHERENCE* would allow me to try to get a hire and try to be out of the committee even though engaging in both these activities would be incoherent. But I think once the example is properly characterized, we can see that *ETR COHERENCE* can explain why these actions are incompatible. The incoherence here can't be just the fact that I am trying incompatible things in the sense above; the video game shows that there is no general incoherence here. What is special about this example is that there can only be a question of my being in the search committee if the hire is approved and if it is approved, I'll be thereby in the search committee. So, what is the end that I am pursuing? It can't be "trying to ensure that I am not in the search committee when there is a hire" as ex hypothesis this is impossible, and at least in the sense of "trying" in play here, I cannot try what I know to be impossible. The only thing I can do is try to prevent a hire (that is, not to get a hire), but trying to φ and trying not to φ are indeed incompatible actions.

I'll just briefly address the phobia case. It's hard for me to have a clear view on this charming example because I think a lot depends on one's understanding of the pathology at play. There are readings of phobia that I am no longer in control in my actions, and in such cases I am hesitant to say that I could be either rational or irrational. Phobia might also involve irrational belief: the pathological emotion gives rise to an irrational belief that the Math Hall is dangerous. This is obviously a case of irrationality but not instrumental irrationality. But perhaps the phobia simply makes it so unpleasant to go the Math Hall that the agent does adopt the end of avoiding the Math Hall. Here I grant that *ETR* cannot rule out that this is *instrumentally* rational, but I think no theory of rationality should. I think the most threatening understanding of Phobic Patty for *ETR* is one in which Phobic Patty is a case of *akrasia* (even if an unusual form of *akrasia*). In such a case, she fails to comply with some kind of enkratic principle.[12] And here I need to grant to Brunero that I still have misgivings about how to accommodate enkratic principles (or show that they are not a proper part of the theory of instrumental rationality). I have tried to do this in the book. Given that I am running out of space, I can conveniently refer the reader to these pages (164–7), rather than try to persuade her here of my success.

[12] For my own views on *akrasia*, see Tenenbaum 2007, 2018.

Mayr

Mayr also raises a number of important challenges to my view. Let me start with the procrastination case. Mayr challenges whether in this case we can still think of *Sufficiency* as action guiding. As Mayr correctly points out, the principle does not simply tell the agent not to do something incompatible with the action they are engaged in. *Sufficiency* also enjoins the agent to take some sufficient means to it; the principle needs to be action-guiding with respect to the agent's "positive contribution" (as Mayr puts it) to the pursuit of this end. Before getting to the main point, let me try to first respond to a side issue. Mayr disagrees with my verdict that the person who postpones writing their book for a period of time, but then picks up their pace and ends up with an acceptable realization of their end of writing a book, never acts irrationally (not even at the time that they were procrastinating). More precisely, according to *ETR* such an agent is rational throughout the entire period they are writing the book *relative to the end of writing a book*. But how could it be otherwise? The book was written in the end (and it was a fine book, and it was not too late for the publishers, etc.). It was no accident that the book was written; it was written by my successfully taking the means to this very end. In which sense then, could I have been irrational in relation to the pursuit of this end? Certainly, often procrastination does involve irrationality in relation to *other* ends; I might have engaged in sub-optimal actions (with respect to my Pareto preferences) by staying home to write my book, but done nothing in the direction of writing the book. Or, more specifically, I might have taken a particular means to my end of writing a book (such as staying at home to write ten pages), and have been instrumentally irrational with relation to the pursuit of these means (I never actually took the steps needed to write ten pages). But none of this shows that there is something wrong with the original case: if I did write the book, not through luck but through my competent pursuit of this end, and I did not undermine any of my other ends, I acted rationally.

But whatever our intuitions are here, Mayr presses a more fundamental objection to the view; namely, *Sufficiency* cannot be action guiding because it does not determine how to pursue my extended action through momentary actions. Mayr gives the example of his end of reading *War and Peace* at the beach. Since he procrastinates in reading novels, at each moment he'll prefer not to read. So, the theory will not say at any moment that he needs to read the book. But this does not show that *Sufficiency* is not action guiding. Suppose Mayr does start reading the book at some point on the beach. He is reading it now *for the sake of* (realizing) the extended end of reading the book. His action was the pursuit of (part of) a sufficient means to his end, so he was guided by the principle.[13] And at every moment this end (together with his other ends) will determine what the possible choices are, in particular, in the case as described, the principles of instrumental rationality will make both reading *War and Peace* and just relaxing on the beach permissible. Mayr thinks this is not enough; he complains that "this rational permission does not help the agent who is puzzling about whether to read another chapter or go swimming *now*." But this is just the nature of any rational situation in which more than one course of action is permitted. If I am rationally permitted to watch either *Saturday Night Fever* or *The Colour of the Pomegranates*, no rational principle will help me when I am trying to decide which one to watch.

But there is obviously more to Mayr's concern. Another way of putting his concern, I think, is the following: *Sufficiency* tells him to choose reading *War and Peace* often enough. But how can a principle guide us to read a book often enough without telling us more precisely *when* (at which moments) to read it? However, I think that "Do it enough times" represents the full guidance that our instrumental rational powers can provide at this point (at least without further complications, as we'll see momentarily). Given non-supervenience, any guidance that specified the exact moments which Mayr should dedicate to reading would put arbitrary constraints in his pursuit of this end; after all, there are other means of achieving the same end. I think the suggestion that there is no more specific guidance might seem less intuitive than it is because we're looking at a very wide timeframe. Indeed, in such cases, I suggest that given our limited nature, we will often find it difficult to pursue our ends without the help of more specific "intermediary policies" (as I call them). Given the difficulties in ensuring that I leave enough time to read *War and Peace* with all the distractions of a beach vacation, I might realize that I need to settle on some such intermediate policy ("I'll read in the mornings," or "I'll read at least 10 pages before breakfast every day"). But let us look at a different example. Suppose I always wanted to skydive, and finally signed up to go. At some point the guide will open the plane's door and will ask me to jump. The guide will not expect me to jump me immediately, it will give me time to collect myself. But I can't take too long. If I do, they'll have to close the door and start moving back to pick up other customers. So, the time comes, the door is open, and I wait; I am rather scared and I need to collect myself. At the same time, I need to jump soon enough. But no guiding theory of rationality could specify a precise moment in which I must start bending my legs to jump. The only guidance I can have here is exactly that I need to jump *soon enough*. I can start asking myself: "Must I

[13] In the way described in the book that does not involve explicitly formulating the principle to oneself. See Brunero's contribution on this point.

jump now?" but a theory of rationality could not specify a moment and enjoin me to jump at that particular moment.

Of course, this takes us to another aspect of the same concern. In my original case, I conclude after some extended period of not doing any actual writing that I need to change the way I am doing things and possibly adopt some specific implementation policies of writing the book. But how could I conclude that I need to switch course, unless we accept that I have been acting irrationally? My concern at this point must be that the lack of writing is part of a pattern that is likely to continue into the future. Here I agree with Mayr that realizing that an irrational pattern is likely to happen in the future requires me to change the way I pursue this end; in particular, it requires that I adopt some intermediate policies, and, depending on what I expect from myself, they might have to be rather strict policies (I might need, for instance, to adopt a policy that I *never* look at social media until I write at least a thousand words). However, this is not a problem for *ETR*; in fact, this is course of action dictated by Sufficiency. Once I expect I will act irrationally if I don't adopt stricter policies I will not be pursuing sufficient means to my ends.

Once I plug the data that I expect myself to act irrationally, or even that it is possible or likely that I will act irrationality, I think the theory can give the right results; I must now reason as taking these future actions not as further actions in which I will engage but as part of my circumstances. However, this requires that I treat my future self's activities in the same way I treat a chance of rain in the forecast. And there is here a deeper issue hiding beneath the surface. Sometimes treating my future actions this way seems like a cheat. If I give up too soon on my dream of skydiving because I decide that I am coward and I'll never jump, I seem to be treating myself in an objectionable way. On the other hand, if I ignore my limitations completely I fail to face reality.[14] These are difficult issues and part of the reason that I don't address them in the book is that I am not sure that they are part of the theory of instrumental rationality; they're rather more general questions about how to relate to our finitude that appears in the other contexts (similar issues arise if I take into account my vicious nature in making decisions about whether to engage in certain actions). I think that Mayr might not agree with me here but I must deploy again the excuse of limited space and leave this issue for another occasion.

Similarly, I can only make a couple of brief comments here in response to Mayr's concerns about my views on the rationality of trying. Suppose you were pursuing the end of φ-ing but now realize that you do not know whether it is within your power to φ (if, for instance, you were pursuing the end of meeting your friend at her office, but you are no longer sure that she'll be there). I argue

[14] See Marušić 2015 for related issues.

that it would be perfectly rational to abandon now this end, instead of trying to φ. Mayr thinks that it would be irrational not to try to meet my friend in such an example if, for instance, my friend is likely to be at the office and this end is important to me. First let me soften the blow of this conclusion by noting that, on my view, *it is always rational to abandon one's end* from the point of view of the theory of instrumental rationality. Since the theory of instrumental rationality does not require to pursue any particular end, it cannot also require you not to abandon an end (except if it is instrumental to the pursuit of another end that I still have). And, as Mayr points out, I accept that this might be substantively, though not instrumentally, irrational. But since these considerations are unlikely to satisfy Mayr, let me make a few additional remarks. When Mayr stipulates that meeting my friend is "important enough," what should we understand by "important" here? If "important" refers to how it contributes to another end of mine (I am cultivating my relationship with my friend; I give priority to the end of spending time with my friend; etc.), then it could be irrational not to try to meet her in such a situation according to *ETR*. If "important enough" refers to my views about what I ought to do or what has value, this might be a case of *akrasia*, and here I can only give the same very limited response I gave to Brunero on how the theory is supposed to handle *akrasia*. Finally, if "important" refers to the strength of my desire to meet my friend, and if we thought this has relevance in determining the agent's instrumental rationality, we'd accepting a view about the nature of the basic given attitudes that is incompatible with *ETR*.[15]

Haase

Once again, I'll only be able to address some aspects of Haase's criticisms; many of the issues raised in his paper are questions that I'd want to continue to think about and hope to come back to them in future work. But let me start with a framing issue. Haase presses me at various points on my possibly incautious quoting of Hegel; I argue that my view vindicates to some extent the idea that the rational is the real and the real is the rational. In a nutshell, according to *ETR*, action in the material world is the immediate manifestation of our rational powers and our rational powers extend all the way to the external, material world (rather than stopping at our minds and being connected to the rest of the world via some "brute" causal relations). And Haase is right that this in no way captures the full extent of how Hegel conceives of the identity between the real and rational. But he takes me to task at various places for not being able to live up to even this limited version of the dictum. Haase is aware that I try to remain

[15] I do argue that strength of desire cannot serve as a basic given attitude in chapter 3.

agnostic about various questions on the metaphysics of action, but he thinks that the dictum commits me to viewing the theory of rationality as a metaphysics of action. As he puts it: "Where the real is the rational and the rational is the real, the theory of practical rationality and the metaphysics of action should be one and the same." But I don't think this is true even in a very expansive reading of the dictum. One can accept that the extended world is the material world and the material world is the extended world without committing oneself to Cartesianism about matter. Even though all matter is by nature extended (and every extended substance is matter), there might be dynamic aspects of matter that are not accounted by its nature as extended. Moreover, even if the dictum had this implication, since the book presents a theory of only one part of practical rationality (its instrumental part), it would still seem possible to remain agnostic about the metaphysical issues. So, I will approach Haase's various distinctions that he thinks the theory misses by asking if they should make a difference for the theory of *instrumental* rationality.

Haase disputes whether the rational really reaches all the way to fully determine reality if the rational action is always in progress. If I am writing a book, the book is not there, and once the book is there I am no longer acting. My aim was to have a book written, but my rational powers seem to start just short of this product: all that they can determine is the process of writing it. I think Haase would agree with my taking action in progress to be the focal point of a theory of agency, as this is where, if I can be pardoned the half pun, most of the action is. On some conceptions of intentional action, the completed action seems to be outside of the scope of agency. But I am inclined to reject this view (and, I think, I would be agreeing with Haase here). I confess that my thoughts on this matter are rather tentative, but I do want to make room for the idea that my rational agency *does* extend all the way to the *completed* action. In particular, I think that, in the relevant sense, my action is only completed by my awareness of its completion. Suppose I am pouring soup into my guest's plate, trying to fill their bowl. At some point, I'll have filled (enough of) the bowl. But suppose I distractedly continue pouring the soup, and now the soup has overflowed and it's dripping onto the dinner table, *my action of filling the bowl* has not completed. In cases of telic actions, my activity stops only after I am satisfied that the process has been completed. Just as God needed to be satisfied that He saw that that which He created was good before he could move to the next item of creation, finite beings can only conclude their telic actions by representing them as completed. Of course, much more needs to be said in this matter, and I confess not being sure that these thoughts will hold up under scrutiny. But I think something roughly in this direction must capture the fact that our agency does extend all the way into the completed action.

Haase finds troublesome that *ETR*'s implied (or at least allegedly implied) metaphysics of action is what he calls "monolithic." Here again I want to insist that the real question is whether the various distinctions we might want to make in other theoretical endeavours are relevant for the theory of instrumental rationality. I find Haase discussion of engaging in a pleasant activity such as eating gummies fascinating, but I am not sure that it does pose a challenge to the theory of rationality. *ETR*, of course, allows that extended actions can vary greatly in their extension: some last a few seconds; some are pursued indefinitely. I could, for instance, eat gummies as a constitutive means of my elaborate "snack time activities." Here eating each gummy is a constitutive part of the larger activity ("I'll have a few gummies, some milk, and end with a cup of espresso") and, arguably, each brief pleasure is connected to the larger pleasure of the afternoon snack time, just as the pleasure of listening to each note of "Lavender Haze" is connected to the pleasure of listening to the whole song. But often, my eating gummies, is indeed a case in which, each gummy eating is its own activity, and indeed past gummies "are nothing" to my current casual gummy eating. I think *ETR* is actually well-placed to explain the difference: in the snack case, ensuring the availability of gummy bears, for instance, is essential. The same is not in the case of the casual eater. In other words, the *ETR*'s principle of derivation will classify them as different activities. I think a similar thought both explains why a smoking addiction is not the same as having a continuous atelic end as in the case of singing (or being a singer). For the addicted smoker, each new craving is a new end, and the addicted smoker only procures the means to future smoking out of sympathy for her future self whom she predicts will experience similar cravings. On the other end, the singer procures singing lessons (which are painful now but will pay off in the future) for the sake of the end she is *now* pursuing.

The case of Bartleby is doubtless interesting and I cannot here exhaust what there is to be said about it. If I understand Haase's reading of Bartleby, Bartleby's "I prefer not to" is a form of refusal to engage with the business of instrumental rationality; a steadfast avoidance of any form of pursuit of ends (pursuits that give rise only to frustration or further pointless pursuits). I think Haase is correct that the Toleration Constraint requires me to accept this kind of refusal to act as a possible direction that the will of an instrumentally rational agent might take. I don't think Bartleby's end could literally be described as the end of not pursuing ends (this end can only be realized by immediate suicide),[16] but as the end of avoiding, as much as one can, the business of practical reasoning. But Haase suspects that this response gives up the dictum as the rational in this case is no longer the real; after all, such an end aims exactly at the absence

[16] More on this momentarily.

of self-actualization or self-realization. Or at least “this would be tantamount to giving up on the thesis that instrumental rationality is rationality in action;” after all, nothing happens in the world when Bartleby’s will takes this direction. Of course, this is not quite true of Bartleby himself; Melville’s character has to interact with the world. In order to remain put in the office, he has to keep answering the entreaties of his boss, even if he does it always with the same phrase. Bartleby also had to eat some ginger nuts (and I imagine drink from time to time) so that he could maintain his steadfast non-cooperation. Perhaps these activities were self-betrayals, but even so, they were exactly self-betrayals in the engagement of the will with the world; a failure not to keep one’s spirit untainted by the vicissitudes of external reality. But couldn’t we imagine a more “successful” version of Bartleby, one that simply declines to answer the lawyer’s requests instead of repeating a polite refusal, and one who steadfastly stays put until starvation takes his life away? Here, however, I am in great sympathy with the passage from Korsgaard (2009) that Haase cites; a refusal to act is, as such, a failed project, not only as a project not to act, but also as project not to engage with the world. In refusing to cooperate with the lawyer, Bartleby is not only acting but *interacting*, engaging with the lawyer, his office, and the building, even if just by his insistent silent and his maintaining his position in his physical space at any cost. Our reconceived Bartleby’s withdrawal from the world might be a degenerate case of interaction with the world, but an instance of it nonetheless. The perfect withdrawal could only happen by rendering oneself unconscious in some way; that is, by withdrawing from agency altogether. But this would not be a case of *irrationality*, but simply of non-agency, no more problematic for a theory of rationality than a case of somnambulance.

Haase also has doubts about my conception of the instrumental virtues. First, a small correction: Haase says that I take the coward to be instrumentally irrational, but I would prefer to say simply that the coward falls short of ideal rationality; it’s a limitation of the capacity itself, rather than a defective manifestation of it (just as we can say that a dart thrower has limited skills even if, due to the fact that all her attempts are from close range, she has always hit bullseye flawlessly). Much of what I say in response to Paul on this issue also addresses Haase’s concerns, or so I hope. So, here I’ll focus on a couple of additional issues. Haase points out that in choosing certain courses of action, in developing my skills, in short, in living my life, I must foreclose some options. Why wouldn’t I be committed to any such choice as a case of falling short of ideal rationality, just like the coward, who cannot choose ends that require bravery? I argue in the book that there is a difference between a limitation in our capacity to act that is external to our will one that is internal to it. That I cannot fly makes certain ends impossible for me to pursue, but it is not a shortcoming of my will.

On the other hand, if, because I am too cowardly, I would not fight oppression even if I were to set it is an end, my cowardice *is* a shortcoming of my will. In a nutshell, the problem is not that I *could* not fight oppression but that I *would* not. Haase's examples seem to me clear case of external limitations. It might seem different because a choice of mine foreclosed some possibilities, but that *being both a basketball player and a football player* is not a possible end for me due to my physical limitations. Choosing that, given my limited physical powers, I will train my football skills is an *exercise* of the will, rather than a shortcoming of it. Finally, Haase disputes my claim that the cases of procrastination that I present can be confidently said that are cases of instrumental irrationality. My argument was that, unlike cases of alleged momentary instrumental irrationality, we cannot say that the agent simply abandoned her end (writing her book) in favour of another end (say, watching TV) because for some of the actions she undertakes (typing on the computer), the only possible end that she can be pursuing is exactly the one for which she is not taking sufficient means. Haase does an excellent job in finding other ends that I could have been pursuing rationally in simply engaging in the (unsuccessful) process of writing a book. But I don't see why we should accept that for every case of procrastination there will be such an alternative end explaining my actions. In fact, if we replace "writing a book" with "writing a grant proposal," I find it all the more plausible that I would end up not completing the process of writing the grant, and absolutely implausible that there would be any non-instrumental end I would be pursuing in the process of writing itself.

Conclusion

I know these remarks fall embarrassingly short of fully addressing these insightful sets of comments. I feel humbled to have this group of amazing philosophers engaging so thoughtfully with my book. I would like to end by expressing one more time my deep gratitude to all my wonderful critics here and to Luca Ferrero for making this possible.

Sergio Tenenbaum
Department of Philosophy, University of Toronto
sergio.tenenbaum@utoronto.ca

References

Anscombe, G. E. M., 2000, *Intention*, Harvard University Press, Cambridge, MA.

Bratman, M., 1987, *Intention, Plans, and Practical reason*, Harvard University Press, Cambridge, MA.

Davidson, D., 2001, *Intending. In his Essays on Actions and Events: Philosophical Essays Volume 1*, Clarendon, Oxford, 83-102.

Ferrero, L., 2017, "Intending, Acting, and Doing," in *Philosophical Explorations*, 20 (sup2): 13-39.

Korsgaard, C., 2009, *Self-Constitution: Agency, Identity, and Integrity*, Oxford University Press, Oxford.

Marušić, Berislav, 2015, *Evidence and agency*, Oxford University Press, New York.

Moran, R. and Stone, 2009, "M. Anscombe on Expression of Intention," in Constantine Sandis, ed., *New Essays on the Explanation of Action*, Palgrave Macmillan, London, 132–168.

Pettit, P. and Smith, M., 1990, "Backgrounding desire," in *Philosophical Review* 99 (4):565-592.

Tenenbaum, S., 2007, *Appearances of the Good*, Cambridge University Press, Cambridge.

—, "The Guise of the Guise of the Bad," in *Ethical Theory and Moral Practice* 21(1): 5-20.

Thompson, M., 2008, *Life and Action*, Harvard University Press, Cambridge, MA.

Wallace, J., "Normativity, Commitment, and Instrumental Reason," in *Philosopher's Imprint* 1(4).

Edizioni ETS
Palazzo Roncioni - Lungarno Mediceo, 16, I-56127 Pisa
info@edizioniets.com - www.edizioniets.com
Finito di stampare nel mese di febbraio 2023