Precis of rational powers in action[1]

Sergio Tenenbaum

Abstract: Human actions unfold over time, in pursuit of ends that are not fully specified in advance. *Rational Powers in Action* locates these features of the human condition at the heart of a new theory of instrumental rationality. Where many theories of rational agency focus on instantaneous choices between sharply defined outcomes, treating the temporally extended and partially open-ended character of action as an afterthought, this book argues that the deep structure of instrumental rationality can only be understood if we see how it governs the pursuit of long-term, indeterminate ends. These are ends that cannot be realized through a single momentary action, and whose content leaves partly open what counts as realizing the end. For example, one cannot simply write a book through an instantaneous choice to do so; over time, one must execute a variety of actions to realize one's goal of writing a book, where one may do a better or worse job of attaining that goal, and what counts as succeeding at it is not fully determined in advance. Even to explain the rational governance of much less ambitious actions like making dinner, this book argues that we need to focus on temporal duration and the indeterminacy of ends in intentional action. Theories of moment-by-moment preference maximization, or indeed any understanding of instrumental rationality on the basis of momentary mental items, cannot capture the fundamental structure of our instrumentally rational capacities. This book puts forward a theory of instrumental rationality as rationality in action.

Keywords: practical rationality, intrumental rationality, decision theory, extended action, intention.

1. *The basic structure of the theory*

Rational Powers in Action defends a theory of instrumental rationality that significantly departs from most contemporary treatments of this topic. In a nutshell, the theory proposed there, *The Extended Theory of Rationality* (*ETR*) takes intentional action to be the primary category of the theory (it's an "action first" theory, somewhat akin to "knowledge first" theories in epistemology). This

[1] This precis is largely based on a series of posts in the Brains Blog (https://philosophyofbrains.com/author/tenenbaums).

philinq XI, 1-2023, pp. 67-85
ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i1.442
Submitted: 14 November 2022
Accepted: 18 November 2022

is a departure from the dominant approach of assigning that primary role to momentary mental states. Changing the focus of the theory in this way turns out to have major implications, or so I argue in the book.

Here is a sketch of the domain of a theory of instrumental rationality: An ideally rational agent efficiently pursues a conception of the good life, a conception that is warranted in light of their knowledge. The theory of substantive practical rationality investigates the principles that guide a rational agent in choosing their conception of a good life, and the theory of instrumental rationality investigates the principles that guide a rational agent in the efficient pursuit of this conception of the good life. There is much to quibble with in outline of a theory of practical rationality, and, as will become clear momentarily, I myself find it too narrow. But let me bring two points to attention here: First, a theory of rationality so understood focuses on rational principles that *guide* agents (insofar as they act rationally), rather than on principles that merely *evaluate* agents, or principles that keep score on how agents are doing relative to a certain standard. In my preferred language, the theory describes the nature of (part of) the agent's rational powers or capacities. Second, a theory of instrumental rationality does not aim to be a full theory of practical rationality, as it leaves questions about the rationality of our basic ends or preferences untouched. It might be stupid, irrational, or ill-advised that I am intent on erecting a monument to Jakob Fries in my backyard, but this is no concern of the theory of instrumental rationality. Our theory concerns itself only with whether I am doing it coherently and efficiently.

Now, debates about instrumental theories of rationality often rely on very different conceptual apparatus. For instance, some of them take graded states as their starting points and propose formal theories, while others rely on binary states; some take risk and uncertainty as their central case, while others pay scant attention to such scenarios. While in epistemology there has been a raging debate about the relation between credences and beliefs, or between traditional epistemology and formal epistemology, this has happened to a lesser extent in debates on the conative side of the equator. So, it might help to sketch what I take to be the main components of this kind of theory of instrumental rationality:

> i. Basic Given Attitudes: A theory of instrumental rationality will take some attitudes as basic, both in the sense that, at least each in isolation, they (almost) never manifest irrationality, but they are also at the centre of the theory of instrumental rationality. On a standard reading of Hume, Hume thought that our passions are neither rational nor irrational (not even just from the point of view of instrumental rationality), and that reason was slave of the passions. Passions are not only beyond rational criticism, but whether you acted rationally or not depends on whether your rational powers were properly obedient to your passions. Interpreted

in this way, Hume took passions as the basic attitudes. Among the most popular candidates for being basic given attitudes are intentions, desires, and preferences. So, for instance (and ignoring complications), for a theory of instrumental rationality based on decision theory, the basic given attitudes are preferences. An isolated preference, say, for pushpin over poetry, is neither rational nor irrational (although, of course, it might be a member of an incoherent set of preferences).

ii. Standard Exercises: So, if the basic attitudes are the "inputs," the standard exercises are the attitudes that serve as the outputs of practical reasoning. The chapter in which Hume famously defends the view that reason is the slave of passions is called "Of the influencing motives of the will." Reason's forced labour is at the service of directing the will, and thus the standard exercises of instrumental rationality on this view are "willings." On a possible interpretation of decision theory, *choice* is the standard exercise of instrumental rationality; a rational agent *chooses* the option that maximizes utility. Other common candidates are intentions and decisions.

iii. Principles of Derivation and Coherence: A rational agent moves from basic attitudes to the standard exercises guided by certain rational principles. These will be the principles of derivation. Moreover, even if the theory does not put restrictions on the content of isolated basic given attitudes, it might rule out certain combinations of these attitudes. These are the principles of coherence. Means-ends coherence, the axioms of decision theory, principles of intention stability, all count as possible principles of this kind.

I can now give the first outline of *ETR*. According to *ETR*, both the basic given attitudes and the attitudes that constitute that standard exercises of practical reason are *intentional actions*. Its sole principle of derivation is a version of the Principle of Instrumental Reasoning and the only principle of coherence (that I argue follows from the principle of derivation) is a prohibition on engaging in the pursuit of incompatible ends. In particular, my view is that nothing short of having intentional actions as our basic given attitudes can provide a proper theory of instrumental rationality for extended agency (that is, agency through time) in which the agent pursues indeterminate ends (that is, ends such that not all the relevant aspects of the end are specified in advance). So, when I am writing a book, I am engaged in a pursuit that takes time and whose goal is not fully specified (how good does it book need to be? How long? When does it need to be done?). So *ETR* is a view of instrumental rationality insofar as we are concerned with the pursuit of indeterminate, extended ends. But this restriction does not really put any real limits on the scope of the theory: Examine your life and actions, and you'll find nothing but the pursuit of indeterminate ends in temporally extended action.

2. *Classical vs contemporary conceptions of instrumental rationality*

Kant thought there was a single principle of instrumental rationality, the hypothetical imperative, that connected the pursuit ("willing") of ends and the pursuit of means. I think Kant was far from unusual on that point; at the time, western philosophers take for granted that something like the hypothetical imperative is the core principle of instrumental rationality. At any rate, I will call a "classic" conception of instrumental rationality, a conception that takes the central principle of derivation to be a version of the principle of instrumental reasoning connecting the pursuit of ends to the pursuit of means.[2] On this conception, the principle connects temporally extended actions to temporally extended actions. That is, an instrumentally rational derivation always connects something one is doing to something else one is doing:

[END] I am making a cake (pursuing the end of making a cake).

thus

[MEANS] I am making the batter (pursuing the end of batter making).

Let us take decision theory, understood as a normative theory of instrumental rationality, as our paradigmatic case of a contemporary theory of instrumental rationality. The focus there is on momentary mental states (utility or preference) that determine a rational choice or decision. Decision theory, understood in this manner, enjoins us to choose the act that maximizes expected utility. So the "output" attitude of the theory is also a momentary mental state; namely, a *choice* (or decision). The notion of pursuing an end is replaced by a comparative, momentary, attitude (preference) and the relation between the decision (the standard exercise of our rational powers) and intentional action is not within the subject matter of the theory. ETR is a version of the classical conception. Certainly, decision theory has greatly contributed to our understanding of rationality (more on this below), but I argue that a classical conception such as ETR has distinct advantages as a *fundamental* theory of instrumental rationality.

The following vignette from the book is supposed to illustrate one of these advantages:

> While on the subway to work I space out and, before I know it, I've reached my destination. But there were many things I could have done between the time I boarded the subway and my final stop. At each moment, I could have chosen to grade a paper from my bag, or to read … [a] book, or play some electronic games on my phone. There were also slight improvements that I could have made to my seating arrangements …

[2] For a more precise formulation, see *Rational Powers in Action*, p. 44.

improvements that I could have weighed against the inconvenience and effort of moving from one seat to another. (*Rational Powers in Action*, p. 5)

On the decision theory model, each time I failed to consider these options, I risked falling short of ideal rationality, and, if some of these options had greater utility, I fell short of the ideal. Of course, the advocate of decision will accept that we don't really approach this ideal, and given our limited cognitive capacities and resources, we should use heuristics and not try to maximize utility at every juncture. In fact, given our limited cognitive resources, it is *impossible* for us to be ideally rational for any significant stretch of time. Yet, intuitively, my trip on the subway was perfectly rational: I was riding on the subway for the sake of going to work and I did this unimpeachably; this is exactly what a classical conception predicts.

My view is that decision theory's ideal is so distant from the reality of human agency because it does not allow for indeterminate and non-comparative attitudes. My ends of discharging my professional duties, reading novels, and enjoying mindless entertainment are neither fully determinate (they do not fully specify what counts, for instance, as an "acceptable" realization of reading novels) nor do they fully determine a preference ordering between various ways of realizing them (is a life with reading 3182 novels and barely discharging my professional duties better than one in which I read 3181 novels and do slightly more professionally?). Moreover, decision theory's restriction on the nature of what we care about or pursue violates what I call "The Toleration Constraint": theories of instrumental rationality should not prescribe what agents should pursue or care about, but only the efficient and coherent pursuit of what they care about; if a theory of instrumental rationality must allow that I prefer the destruction of the university over scratching my finger, it surely should allow the pursuit of indeterminate ends.[3]

Let us now ask how decision theory moves from a preference ordering to the rationality of particular actions. Suppose Mary prefers apples over pears; you now give Mary a choice between an apple and a pear. Does she choose the apple over a pear? We are tempted to say "yes" here, but, of course, it must depend on further details of the choice Mary is offered. If the apple was rotten and pear seemed passable, it is compatible with having a *general* preference for apples over pears that she chooses the apple over the pear on this particular occasion. It might seem that this just shows that we did a poor job in specifying Mary's preference: it should be a preference for fresh apples over pears. But given the non-monotonic nature of practical inference, for any way one specifies the pref-

[3] Of course, there are non-orthodox versions of decision theory that allow for preference gaps, imprecise preferences, and so forth. I argue in the book that these solutions don't address the central problem: decision theory (as a normative theory) starts from the wrong basic attitudes.

erence, I (or someone more creative than me) will find an instance of the options so specified, in which the agent would have the opposite preference. So, the agent might prefer a succulent, ideal pear, over a dry, low quality, fresh apple. Moreover, Mary cannot get the apple just by mentally choosing, she needs to go out in the world and grab it, and how she does it is relevant to her rationality. Even if Mary prefers this specific apple over this specific pear, not all ways of picking it up manifest rational agency. If Mary climbs an electric fence and predictably loses her sense of taste, she did not act rationally. I argue in the book that decision theory has no satisfactory way of moving from the rationality of a choice to the rationality of an action, but a theory of practical rationality should be able to determine whether *actions* are rational or irrational. Under *ETR*, the action itself is supposed to manifest rationality by pursuing sufficient means to an acceptable[4] determination of the end I am pursuing; and given the nature of the principle of instrumental reasoning, an action that pursues an end while undermining another (if I pursue my end of eating delicious apples by crashing my car into an apple tree), also manifess irrationality.

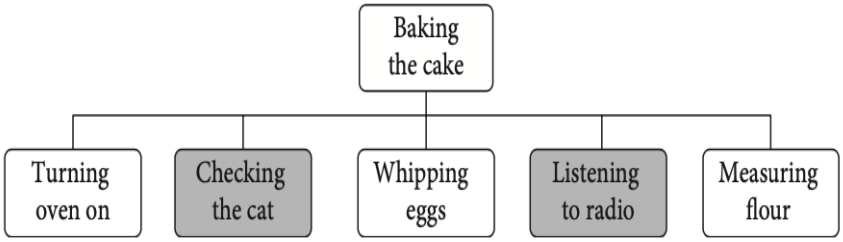
Finally, given the nature of the attitudes at the center of contemporary theories, they evaluate the rationality of an agent at a specific point in time. If the only attitudes relevant to the evaluation of actions are the ones the agent has at the time of the action, we have a time-slice theory of rationality. Now, not all philosophers in this tradition accept time-slice rationality. Philosophers like Michael Bratman (1987, 1999, 2006, 2018), David Gauthier (1997), Richard Holton (2009) Edward McClennen (1990), and Sarah Paul (2014) try to account for the rationality of choice over time by arguing that the rationality of the agent at a particular point in time might depend on their past actions and attitudes; in other words, they allow for diachronic rationality. But on *ETR*, the central attitudes are themselves extended; if I was writing a book between 2010 and 2020, whether I pursued this end rationally depends on what I was doing throughout this entire period. This is obviously not a form of time-slice rationality, but neither is it an endorsement of diachronic rationality, at least if such endorsement implies that the rationality of an attitude at a time depends on the agent's attitudes *at times prior to (the onset of) this attitude*.

In fact, *ETR* differs from both the time-slice and diachronic conceptions, in that on *ETR*, the rationality of an agent through an extended period of time t_0–t_n does not even supervene on the rationality of the agent at each moment in the interval between t_0 and t_n. This *nonsupervience claim* I argue constitutes be a major advantage of *ETR*.

[4] More on the notion of "acceptable" below.

3. ETR *and nonsupervenience*

Let me start with a bit more detail on the structure of *ETR*. Suppose I am intentionally baking a cake. According to *ETR*, this action is an end that I am pursuing and thus the principle of instrumental reasoning enjoins me to pursue sufficient means. The pursuit of various means for the sake of the end of baking a cake are thus manifestations of my instrumental rational powers. Baking a cake is an action that takes time; the means to the end of baking a cake are also further extended actions. But baking a cake is also what I call a "gappy action"; not everything I do in the entire interval is a means for baking the cake. I might turn the oven and then stop do something else, then whip the eggs, stop to listen to the radio for a minute, and then measure the flour. The diagram of my baking the cake might look something like this:



Of course, the actions I take as means are themselves extended and they themselves could be gappy. Underneath our "Whipping eggs" cell, we could have "grasp the whisker," "whisk the eggs," "check the cat again" (shaded), "return to whisking," and so forth. The shaded cells represent the actions that are performed while I am baking the cake but not for the sake of baking the cake. However, they are also partially "controlled" by the end of baking the cake. I act irrationally if I perform an action during the gaps that is incompatible with my baking the cake; that's why if you call me and ask me to help you move, I'll say "sorry, I can't; I am baking a cake." For the same reason, I cannot listen to the radio for too long; if I do, the whipped eggs will turn to mush, or I will need to leave go to work, or I'll eventually die of old age. So how long can I listen the radio for? Well, my end of baking the cake is indeterminate in many ways: for instance, it is left undetermined how tasty it needs to be, or how late it needs to be ready. It seems plausible that there is no exact moment such that both (i) if I continue listening to the radio for even one more millisecond there'll be no acceptable completion of the cake, and (ii) if I otherwise stop then, I'll be able to

realize my end properly.[5] In fact, if I enjoy listening to the radio, it might be that at any particular moment I prefer to keep listening to the radio rather than continue the baking of my bake.[6] Given that going back to cooking a millisecond later will make no difference to my baking, it seems that I prefer to listen to radio for a millisecond more. Yet, if I keep this pattern going, I'll end up not baking my cake.

This pattern is ubiquitous. Just to give another example, next time you are wasting time on Twitter (or some other website) planning to get back to work soon, ask yourself "will it make a difference to my professional life if I read just one more tweet?" The answer is invariably "no." Yet, as we know all too well, we can easily waste the day online if we keep going. It is tempting to say: "there must be a moment in which it is ideal to stop; the moment in which I'll have done the maximum amount of radio listening without compromising my cake," but I argue in the book that this is an illusion; the theory of instrumental rationality cannot pick out an exact point. In a nutshell, there are various points in which I have clearly left myself enough time to bake an acceptable cake and clearly did an acceptable amount of radio listening. If I stopped at any of these points I acted rationally, and if I stopped at any point in which I clearly did not bake an acceptable cake or in which I cut off my radio listening clearly too soon, then I manifested irrationality.[7] Since there is no such exact last moment, it would be a gratuitous demand of a theory of instrumental rationality to say that I *must* stop at a specific point. On the other hand, it would also be self-defeating if the theory said that I *must* keep listening to the radio as long as this is my most preferred alternative. Thus the principle of instrumental reasoning must issue:

> (a) permissions not to choose a most preferred alternative in order to pursue an indeterminate end.
> (b) requirements to exercise some of these permissions.

Anything more would be a demand to ask to pursue something beyond the sufficient means to my end; anything less would make it impossible to pursue indeterminate ends. Once we notice this general structure of the rational pursuit of extended indeterminate ends, a number of consequences follow. First is the *NONSUPERVENIENCE THESIS* I mentioned above:

[5] Or if there's such a moment, I have no way of knowing it

[6] Note that according to ETR, preferences cannot be the *basic* given attitudes. But my ends my generate preference orderings. More on this later.

[7] And of course, there might be borderline cases in which it is not determined (knowable) whether I stopped at an acceptable point.

The rationality of an agent through a time interval t_1 to t_n does not supervene on the rationality of the agent at each moment between t_1 and t_n.

Since there is no "last moment" in which I can exercise a permission to stop listening (given the indeterminate nature of my end of baking a cake), I could always keep failing to exercise these permissions until it's clearly too late to bake a cake. At each momentary snapshot in the interval, I would have acted rationally, and yet I would not have acted rationally throughout the interval.

Next, we get a vindication of "satisficing." In pursuing multiple indeterminate ends, the agent often must be guided by the pursuit of "enough" of it (enough money, enough professional success, a good enough cake, enough fun). Satisficing is a rational ideal for us, not because of our limited cognitive capacities, but because given the structure of indeterminate ends, maximizing is literally impossible. In our cake baking vignette, there is no best combination of baking and listening to radio; I could always listen to the radio for one more millisecond.

Finally, future-directed intentions turn out to be dispensable. What Bratman[8] takes to be characteristic of our planning agency, turns out to be a much more general feature of the pursuit of any action extended through time (and thus of the pursuit of any action). The rational requirements that supposedly apply specifically to future-directed intentions are an immediate consequence of the principle of instrumental reasoning applied to extended agency.

At this point you might be tempted to say that this is all wrong-headed: "there *must* be a last moment in which I can stop listening to the radio without compromising my baking, and decision theory gets it right that I maximize utility (and thus act rationally) only if I stop at this point." In the book, I argue against this thought by focusing on a particularly sharp instance of this general structure: Quinn's puzzle of the self-torturer. The self-torturer (ST) is given the following series of choices: for $100,000, a weird scientist will permanently attach a device to ST's body that gives her electric shocks of varying degrees of intensity. The machine has many settings corresponding to increasingly more powerful shocks. The settings move very gradually (but irreversibly): adjacent settings are (nearly) indistinguishable to ST, but very high settings deliver extremely intense pain. ST is paid 100,000 every time she moves up a setting. Whichever setting she's in, ST seems to have compelling reason to move on to the next one; after all, she cannot (can barely) notice any difference in pain level, but she pockets an extra $100,000. But it cannot be rational for her to keep moving up the settings. After all, at the higher settings, she would be in agony and would gladly return all her earnings (and probably pay much extra) to have

[8] See references above.

the device removed. When should ST stop? For decision theory, there must be a last setting $\mathbf{s_n}$ such that stopping at $\mathbf{s_n}$ is permissible, but stopping after this point is not. I argue that this is an extremely implausible conclusion. Although the argument is complex,[9] the central problem is that decision theory cannot preserve a plausible constraint on any solution to the puzzle; what I call *nonsegmentation*. In a nutshell, nonsegmentation says that in a one-shot version of the puzzle, I must (or am at least permitted) to accept the money. Suppose that due to my back pain I am already at a pain level equivalent to $\mathbf{s_n}$. I am now offered $100,000 to be part of a study testing a cosmetic product that will move me to a pain level equivalent to $\mathbf{s_{n+1}}$. I *cannot tell the difference* between these two pain levels,[10] and I was really looking forward to be able to afford a new kitchen renovation. It seems completely unwarranted to say that it would be irrational of me to accept the money, but this is what any theory that rejects nonsegmentation is committed to. On the other hand, *ETR* has no problem explaining why nonsegmentation holds. In the original puzzle, I can exercise the permission in (a) above, because, to use the language of the book, my end of a relatively pain free life is *implicated* in the series of choices; however, the pursuit of money in the one-shot case does not encroach on the pursuit of the better anesthetized life.[11]

These are some of the advantages for *ETR*. But some features of the theory might appear problematic: *ETR* seems to have no place for comparative attitudes, and thus, arguably, no place for acting under risk. On the other hand, decision theory shines exactly in cases of risk and uncertainty. In the book, I argue that *ETR* can appropriate the resources from decision theory in the contexts in which decision theory is most plausible and provide important explanations of why decision theory proves to be implausible in other contexts.

4. ETR *on comparisons and risk*

4.1. Preferences

Let us assume that at the start of your adult life you have only one end; namely, singing. Your whole life is dedicated to it. But then, one day you discover the joys of marathon running, and now you have two ends: singing and running marathons. As you go out for your first training run while singing, you realize

[9] The argument first appeared in a paper co-authored with Diana Raffman (Tenenbaum and Raffman 2012).

[10] Are they then different pain levels? I am assuming they are, but we could make the same point in a more longwinded manner, by just focusing on the changes to the physical causes or the physical realizers of the pain.

[11] It is worth mentioning that I argue in the book that the puzzle does not depend on crossing vague thresholds; you can create a very similar structure by relying on repeated gambles instead.

that as you huff and puff, your singing suffers. You stop to hit the right note, but then you realize you are no longer training as you should.

You have arrived at the realization that your two ends are incompatible, at least in their unrestricted version: you cannot have both the ends of singing as much as possible, and being as good a marathon runner as possible. Since it is, according to *ETR*, incoherent to pursue incompatible ends, you must give up or modify at least one of them. You could give up singing altogether or marathon running altogether, or you could have as an end to sing a lot and be a decent marathon runner, or to be a committed marathon runner and sing from time to time. *ETR* is completely neutral on the question of *how* you should revise these ends; it only says that you *must* revise them. So far, this seems right to me; in fact, I argue that attempts to say that it matters how *strongly* you desire each of these things will quickly collapse into a form of normative hedonism. But hedonism is not a theory of instrumental rationality; it is a substantive view about intrinsic value. However, in some cases, comparisons are important for the theory of rationality, and it seems undeniable that often what I prefer is relevant to my rational agency. Moreover, comparative attitudes seem particularly important in contexts in which I face risk or uncertainty: how can we evaluate prospects with radically different outcomes if we can't compare the value of these outcomes? *ETR* seems to be embarrassingly silent on these arguments.

However, *ETR* says that comparative attitudes are not the *basic* given attitudes, but not that they cannot be given attitudes. In particular, if *ETR* can show that the basic given attitudes it postulates generate preference orderings in specific contexts, then it can simply appropriate the resources of decision theory in these contexts. The book argues that the contexts in which *ETR* generates preferences turn out to be exactly the context in which decision theory verdicts seem most plausible. Here are three ways in which our ends generate preference orderings.

> i. Preference Relative to an End
>
> Most of the ends we pursue have a certain internal structure. So if my end is to build a house, there will be better and worse houses, and thus better and worse realizations of the end. Although I will have realized my end if I build an acceptable house, in pursuing this end I am guided by its internal structure. If no other ends are even implicated, then a rational agent pursuing the end of building a house who faces the question of whether to build it from sticks, straw, or bricks, will not be in a Buridan's ass situation: the nature of the end determines that they use bricks, even if a straw house is an acceptable one.[12]

[12] My explanation of why the end has this structure is based on my views that all our actions are done under the guise of the good; the structure is inherited from the nature of the good you are pursu-

ii. Pareto Preferences

In some cases, an action of mine advances many ends without implicating or being in any way relevant to any other ends. So going on a hike might advance my ends of spending time with my loved ones, exercising, and appreciating natural beauty. And let us assume that my going on a hike is not relevant for any other ends I might have. But now suppose there are two hikes, one of which (the Glacier Lake hike) is more beautiful, quieter, and more strenuous without being out of reach. The Glacier Lake hike is a better realization of every single one of the ends I am pursuing in going for a hike, and thus I have a Pareto preference for the Glacier Lake hike over the unnamed hike; the rational pursuit of these ends determines that I hike at Glacier Lake.

iii. Reflective Preferences

Here are two ends I am constantly pursuing: the end of following my Brazilian team and the end of ensuring the welfare of my children. Here I am watching an important match for my team, when I notice that my child is in distress and needs my immediate attention. There is no question in my mind which end I need to pursue; I must tend to my child's needs. This is not because my desire for the welfare of my children is in some way stronger at that moment, but because I have a higher-order end that I am also pursuing; roughly, the end of giving priority to the pursuit of my child's welfare over the pursuit of my end of supporting my team.

These three types of preference provide some structure, though they will typically be localized. They might generate fine-grained preference orderings among possible means of building a house, but they will say very little about choosing among competing ends for which we have not formed reflective preferences, or at least not reflective preferences that are fine-grained enough. But this is not necessarily an area where decision theory excels; this is the terrain of "incomparability" and "incommensurability" where the tools provided by decision theory break down. The main problem so far is that it is not yet clear how this limited ordering will help us understand the nature of rational agency under risk. In order to do this, *ETR* needs a bit more equipment.

Given our reflective powers, we can think of the ends that we are pursuing as a totality, and engage in their coordinated pursuit. This is what I call, "the end of happiness," the end of pursuing all our ends well. There are certain means to this end, means that we pursue not for the sake of specific ends but as means to whatever we might be pursuing. So if I decide to follow my doctor's advice that

ing. But *ETR* is not committed to this explanation.

I should exercise more, I might not be doing so for the sake of any particular end, but as a way of better pursuing many, or all, of my ends.

The same holds when I am making decisions about how to invest my money. Health, wealth, and the cultivation of my talents are, among others, *general means* to the end of happiness. The pursuit of these ends also has an internal structure that generates a preference ordering internal to the end. But note that these ends are amenable to much more fine-grained ordering. A house can be better or worse in many dimensions, but wealth, at least if we ignore liquidity, seems to generate a very clear and detailed ordering that can be summarized by the economic principle, "the more, the merrier." Health is more multidimensional, but at least there are some broad categories that suggest a clear ordering, such as life expectancy. Decision theory is particularly compelling in exactly these areas and so if we can incorporate the insights of decision theory in our pursuit of general means, we might have the best of both worlds. But to do so, it is not enough that *ETR* generates a preference ordering in such domains; it needs also show that it can incorporate decision theory's treatment of risk, or at least something like it.

4.2. Risk

In the height of the pandemic, I started engaging in (what seemed to me at the time) the temporally extended action of travelling to Rio de Janeiro.[13] After calling a few airlines and looking into COVID travel restrictions, it became increasingly clear to me that I did not know whether it was possible to fly to Rio from Toronto and back in the dates available to me. As soon as I realized that I did not know that it was possible for me to travel, the action of *travelling to Rio* was no longer a possible action for me. In decision theory, I weigh the utility of each possible outcome by the probability that it will obtain in order to determine the utility of an act. But under *ETR*, my state of knowledge changes the range of actions open to me. I could no longer be engaged in travelling to Rio, even if I could be engaged in various related pursuits: improving my chances of going to Rio; pursuing opportunities to go to Rio; leaving open the possibility of being in Rio in the following month; and so forth. *ETR* does not imply that a rational agent will now engage in any of these related actions. Again, this seems the right result; instrumental rationality should not require any specific revisions to my end when I realize it is not in my power to ensure that I will be in Rio in the near future. However, an option that I do have is to make a rather minimal revision in my end, and pursue instead the end of *trying* to travel to Rio. Just like the end

[13] Or at least preparing to travel to Rio de Janeiro; in the book, I argue that for our purposes, it is not relevant when the proper action of travelling to Rio de Janeiro begin.

of building a house, trying has an internal structure: I am arguably not even trying to dance the tango if I just move my right foot distractedly to the side a couple of times; I am doing better if I attentively follow these instructions, and possibly even better if I watch an instructional video. I argue that the internal structure of trying gives rise to very basic risk principles, such as, for instance, that, *ceteris paribus*, a rational agent trying to φ faced with a choice between two ways of trying to φ will choose the one that is more likely in resulting in their φ-ing. Such basic principles are obviously a far cry from the powerful principles of decision theory. But let us take our end of making (enough) money. In various circumstances, an obvious means to this end is *trying* to make money. The end of trying to make money will inherit its internal structure from the end of making money, but it doesn't determine a particular way of balancing, for instances risky attempts of greater gains and safer bets at lower ones; for this we need *reflective* preferences in which agents can give different kinds of priority to one over the others. Risk functions of classic decision express possible forms of these reflective preferences. More liberal approaches to incorporating risk, like Lara Buchak's (2013) risk-weighted expected utility model, provide us with a wider menu of reflective preferences; I argue in the book that *ETR* will likely allow attitudes to risk even more permissive than the ones allowed by Buchak's theory. But the important point is that, under *ETR*, these risk attitudes are ways of making more determinate the internal structure of the indeterminate end of trying to make money.

This strategy has its limits. Let us take, for instance, the Allais paradox.[14] One of the options in the Allais paradox is *making a million dollars*. This choice is often represented as "100% chance" of getting a million dollars, but I argue this is wrong; this option should be represented as a case of *knowing* that you will make a million dollars. This makes this option essentially different from the others, and turns it into an option that cannot be governed by our end of (merely) *trying* to make money. So *ETR* cannot rule out that a rational agent will choose to make a million dollars even if their reflective preferences (their risk function) would otherwise determine that they choose the "risky" option. But this is a welcome consequence; most of us choose in this manner, and it seems perfectly rational. In the book, I argue that *ETR* is also more promising in dealing with purported cases of bias such as the endowment effect or mental accounting.

[14] Allais (1953). For an overview of the Allais Paradox, see the Wikipedia entry on the topic (https://en.wikipedia.org/wiki/Allais_paradox).

5. *Instrumental principles and instrumental virtues*

We generally think that a theory of instrumental rationality provides us with principles of rationality and that an agent is rational insofar as they comply with these principles. If the theory is a guiding or explanatory theory of rationality, then it claims that an agent is rational only insofar as she is guided by (or only insofar as her actions are explained by) these principles of rationality. But this can't be all there is to a theory of rationality, at least if a theory of rationality should determine what constitutes an ideally rational agent. An agent could always comply with all the principles of instrumental rationality, in all their actions, and yet fall short of ideal of rationality because they do not have all the virtues constitutive of instrumental rationality. Or so I argue.

Let us start by examining the virtue of courage. According to an Aristotelian conception of courage, this virtue can be manifested only in the pursuit of good ends; on this view, the daring burglar does not manifest courage. On a Kantian conception, the actions of the burglar do manifest courage.[15] I find the Kantian conception more intuitive, but I will not argue for it here; I will just assume this understanding of courage. On the Kantian conception, being courageous seems to be an aspect of being instrumentally rational; a coward often falls short of pursuing the means to their ends. So perhaps this is the problem of cowardice: if you are a coward, you will routinely fail to comply with the principle of instrumental reasoning. However, this is not quite true. Let us tell a story with two cowardly heroes: Sticker and Shifter. Our heroes learned of the location of the Holy Grail and set off to bring it to their country. At some point in their quest, they found out about the scary rabbit in their path that threatens to devour anyone who continues towards the Holy Grail. Both Shifter and Sticker are cowards, but their cowardice is manifested in different ways.

Sticker sees the frightening rabbit but hangs on to his end of retrieving the Holy Grail. But, out of fear, he never actually advances any further towards the Holy Grail. Sticker just spends the rest of his life taking a few steps towards the bunny, losing his nerve, and going back to his hiding place. Shifter reacts to the news of the rabbit quite differently. Once she hears the tails about the bunny, and the fate of those who dared to face it, she tells herself "Well, who needs this trinket?" abandons her end of retrieving the holy grail, and heads back home. Sticker violates the principle of instrumental reasoning: he is obviously still pursuing the end of fetching the Holy Grail, while not taking the necessary means to his end. But the same is not true of Shifter. For her, the failure to pursue the

[15] Of course, they are not *virtuous actions*. It is important to note in our discussion below that I am not committed to the view that an action that manifests only instrumental virtues is a virtuous action.

means to retrieving the Holy Grail and the abandonment of the end were concomitant. Thus she is always in compliance with the principle of instrumental reasoning; after all, reason does not tell us never to abandon our ends. In fact, Shifter might do this consistently: conscious of her cowardice, she always abandons an end as soon as she realizes that she'll need to face some danger in order to realize this end. So her cowardice never leads her to violate the principle of instrumental reasoning.

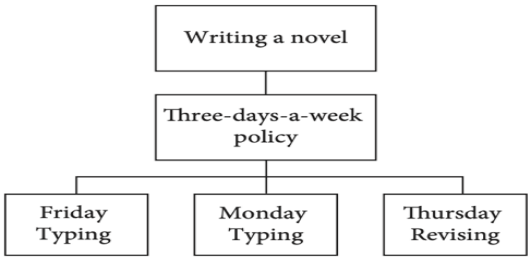
Yet, Shifter still falls short of ideal rationality. Why? In a nutshell, our capacity for instrumental rationality is a capacity to pursue our ends efficiently, *whichever ends we happen to have*. Cowardice is a limitation of this general capacity. Of course, our capacity to pursue ends has many limits. If a putative end requires that I travel faster than the speed of light, it will be beyond my reach. But cowardice is a limitation internal to my will. Shifter *could* just face the rabbit; it is within the general powers of her will. But because she is a coward, she expects she won't. Roughly, instrumental vices are internal limitations to our rational powers to pursue whatever ends we set for ourselves; the instrumental virtues are their contrary.

Of course, *ETR* is not the only theory of rationality that can accommodate the existence of instrumental virtues that are not reducible to compliance with principles of rationality. But *ETR* brings to light a particularly important instrumental virtue: what I call the virtue of "practical judgment." Let us say I am writing a novel. I need to ensure that in the course of the time I give myself to write the novel, I will engage in enough actions that will jointly constitute sufficient means to the writing of an acceptable novel. The *nonsupervenience thesis* ensures that, for the most part, rationality does not compel me to take these means at any particular time during this interval. I could take today off, and this is fully compatible with my action of writing a novel. And the same goes for next day, and the next day. And, again, at each time I might have a Pareto preference for just taking the day off. But again, if I keep doing this every day, at some point it will be clear that I will not be able to write my novel in the available time.

Extended agency gives rise to a problem of managing the pursuit of our ends through long periods of time, when at each particular time we might prefer not to take means to this end. As mentioned above, I am rationally permitted throughout this interval to act against my preferences so as to take the necessary means to write a novel, and I must exercise enough permissions. But at no particular moment am I rationally required to be engaged in the writing of the novel. This predicament poses no problem for an ideally rational agent: they would just exercise some of these permissions and take enough means to their end. An ideal rational agent thus exhibits the virtue of *practical judgment* to the highest degree. The virtue of practical judgment is roughly our capacity to pur-

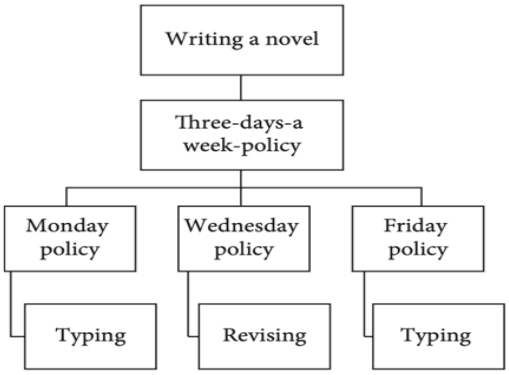
sue indeterminate ends through extended periods of time even when they leave undetermined the specific means for their realization.

Human beings tend to fall short of the ideal of perfect practical judgment. In particular, we often need to engage in what I call "intermediate policies" or intermediate actions.[16] So if I am writing a novel, I might need to rely on a more specific policy, for instance, a work schedule in which I commit myself to write at least 2000 words per week, and to read an average of 100 pages per day. The intermediate policies can be more or less specific (2000 words per week or 300 words per day), and they can be more or less vague or precise ("I will read roughly the equivalent of two books every few days" or "I will read 120,000 characters per day"). The more specific and stricter my policies are, the easier it is for me to ensure that I will not mismanage the pursuit of my ends. On the other hand, the policies that are less specific and more vague allow for more flexibility. If my writing policy involves never leaving home on Wednesdays between 9 and 7, I will lock myself out of pursuing ends that would require my being away during these times. Here is a little diagram illustrating the more and less flexible writing policies. At the top, we have the end of writing a novel, and at the bottom the actions that I perform as means of writing the novel. In between the two, we have the more or less specific policies I adopt in order to pursue this end:



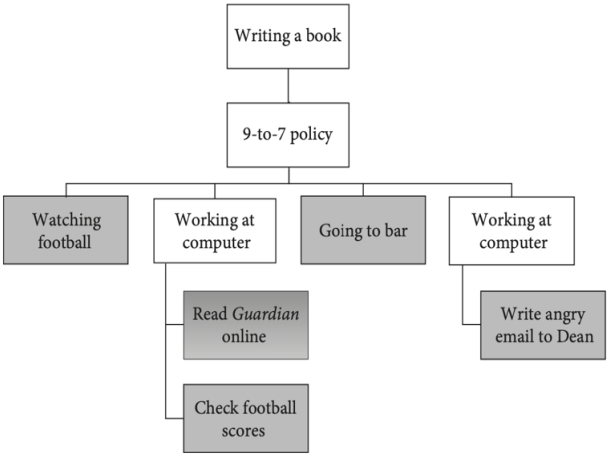
More flexible intermediate policies

[16] One of the claims of the book is that, for the purposes of a theory of instrumental rationality, policies are just instances of extended action.



Less flexible intermediate policies

Although "virtue of practical judgment" is a technical term in the book, the corresponding vices are easily recognizable. The person who needs very specific and strict policies manifests the vice of inflexibility; these are the people who cannot enjoy a beautiful sunny day in March outside because their self-imposed work schedule does not allow for this kind of exception. But even more popular is a vice that corresponds to a more general inability to take the means to our indeterminate ends. Even very specific intermediate policies need practical judgment to be carried out successfully. My quite strict policy of working on my book from 9 to 7 (allowing only a couple of breaks), still leaves room for failures of practical judgment. My attempt to implement this policy might look like this (and note that making the policy stricter would not necessarily solve the problem here):



This vice of implementation is a readily recognizable one: I argue in the book that this just is the vice of procrastination. We are prone to procrastinating not (just) because we have a tendency to discount the future, hyperbolically or otherwise. The structure of the pursuit of indeterminate ends, the fact that I can act rationally at each moment yet fail to act rationally through the resulting interval, makes avoiding procrastination particularly difficult. In fact, we manifest the virtue of practical judgment to a high degree when we are able not only to avoid procrastinating, but to do so without manifesting the vice of inflexibility. A theory of instrumental rationality should not only put forward the correct principles of instrumental rationality but also allow us to describe and explain the nature of the core instrumental virtues. The Extended Theory of Rationality, I argue, gives us a compelling picture of these principles and their relation to the instrumental virtues.

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

References

Allais, M., 1953, "Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Américaine," in *Econometrica*, 21(4): 503-546.

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

Bratman, M., 2006, *Structures of Agency*, Oxford University Press, Oxford.

Bratman, M., 2018, *Planning, Time, and Self-Governance*, Oxford University Press, Oxford.

Buchak, L., 2013, *Risk and rationality*, Oxford University Press, Oxford.

Gauthier, D., 1997, "Resolute Choice and Rational Deliberation: A Critique and a Defense," in *Nous* 31(1): 1-25.

Holton, R., 2009, *Willing, Wanting, Waiting*, Clarendon Press, Oxford.

McClennen, E., 1990, *Rationality and Dynamic Choice: Foundational Explorations*, Cambridge University Press, New York.

Paul, S., 2014, "Diachronic Incontinence is a Problem in Moral Philosophy," in *Inquiry* 57(3): 337-55.

Tenenbaum, S. and Raffman, D., 2012, "Vague Projects and the Puzzle of the Self-Torturer," in *Ethics* 123(1): 86-112.