*philosophical inquiries*
www.philinq.it

# philinq
## philosophical inquiries

XI, 1
2023

Edizioni ETS

Essays published on “Philosophical Inquiries” are double-blind peer-reviewed.

Please visit www.philinq.it for submissions and guidelines.

*six-monthly journal / periodico semestrale*
Subscription (paper, individual): Italy € 50,00, Abroad € 80,00
Subscription (paper, institution): Italy € 60,00, Abroad € 100,00

Subscription fee payable via Bank transfer to
Edizioni ETS
Intesa San Paolo
IBAN IT 21 U 03069 14010 100000001781
BIC/SWIFT BCITITMM
reason: abbonamento “Philosophical Inquiries”

Registrazione presso il Tribunale di Pisa n. 1/13

*Direttore responsabile*: Alessandra Borghini

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

*Distribuzione*
Messaggerie Libri SPA
Sede legale: via G. Verdi 8 - 20090 Assago (MI)

*Promozione*
PDE PROMOZIONE SRL
via Zago 2/2 - 40128 Bologna

ISBN 978-884676565-9
ISSN 2281-8618

# Table of Contents

## Essays

## Focus

ISSN (print) 2281-8618-ETS

# Essays

# Freud and philosophy in Stanley Cavell

Raffaele Ariano

*Abstract*: This article offers a philosophical and historical assessment of the reception of Sigmund Freud in the work of Stanley Cavell. In the first half, I argue that every major theme in Cavell's philosophy entails a dialogue, sometimes explicit and sometimes implicit, with the Freudian model. To this end, I analyse the psychoanalytical motives in Cavell's therapeutic and later perfectionist understanding of philosophy, reframing of the problem of scepticism, and literary and film criticism. The second half of the article is devoted to the sources and interlocutors in Cavell's engagement with psychoanalysis, the most important of which are shown to be non-analytic and even non-philosophical, and in particular literary. Cavell, as I recount, had become a committed reader of Freud in 1947, well before beginning his training in professional philosophy. I thus contend that, in spite of the indifference or even hostility towards Freud that Cavell found in the academic circles in which he was educated and then taught, his reflections on psychoanalysis received their nourishment, outside the philosophy departments of American universities, in figures such as the literary and cultural critic Lionel Trilling.

*Keywords:* Cavell, Freud, Wittgenstein, psychoanalysis, Lionel Trilling.

## 1. *Introduction*

That Sigmund Freud is a major presence in the philosophy of Stanley Cavell is widely acknowledged. Already in the late 1980s, during the first surge of responses to Cavell's work, Conant remarked that "Cavell's most pervasive and sustained intellectual debt" might be shown to be to Freud (Conant 1989: 22). Recently, Assif went as far as to situate Cavell's work in the context of a yet-to-be recognized "strain of Freudians" amongst whom she also includes psychoanalysts such as Jonathan Lear and Christopher Bollas (Assif 2020: 12). Numerous further references could also be made (see for example Mulhall 1994: 216–17; Gould 1998: 41; Eldridge 2011). Cavell himself stressed this influence on many occasions. "The figure of Freud", he wrote for instance, "shadowed my work in philosophy from the time I first published an essay about

Submitted: 5 August 2021
ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i1.375
Accepted: 14 July 2022

Wittgenstein" (Cavell 2005: 213), which means from his 1962 seminal article on the *Philosophical Investigations* (Cavell 2002: 41–67). Surprisingly, however, an overall philosophical and historical assessment of Cavell's reception of Freud and psychoanalysis is still lacking.

In roughly the first half of my article, which will privilege thematic over chronological organization, my main purpose will be to show that there is virtually no significant aspect of Cavell's philosophy which remains untouched by explicit or implicit connections with psychoanalysis. I will look for such connections first in the 'therapeutic' conception of philosophy that Cavell builds mainly through a parallel between Freud's method and the method of ordinary language philosophy, in particular that of Wittgenstein. I will then explore similar connections in Cavell's notion of Emersonian perfectionism. Subsequently, I will highlight the Freudian undertones of Cavell's reframing of the problem of scepticism. Finally, I will sketch the complex role of psychoanalysis in Cavell's critical writings: this time Freud, more than the inventor of a method, would play the role of institutor of "an unsurpassed horizon of knowledge about the human mind" (Cavell 2004: 286).

In the second half of my article, which will follow an inverse chronological order, I will survey and assess both the philosophical and non-philosophical sources and interlocutors in Cavell's reception of psychoanalysis. I will begin with Cavell's texts of the 1980s and 90s, whose main sources prove to be feminist literary and film critics, and thinkers coming from the 'continental' tradition, especially French post-structuralism. Then I will move to considerations on the first two decades of Cavell's production and also sketch out the academic environment in which he received his education. The scarcity of philosophical interlocutors on Freud at UCLA, Berkeley and Harvard, where Cavell came of age between the late 1940s and early 60s, suggests that at that time his interest in psychoanalysis was mainly being pursued through non-academic channels. Finally, an account of the circumstances of Cavell's first encounter with Freud's work in 1947, as well as other information scattered through Cavell's interviews and autobiographical writings, suggests the hypothesis that Cavell was influenced by the writings on Sigmund Freud of literary and cultural critic Lionel Trilling. I will pursue this hypothesis through a brief overview of the similarities between Trilling's treatment of Freud in the 1940s and 50s and that of Cavell in later decades. Overall, the second half of my article will contend that the most sensible way to account for the seeming contradiction between the limited standing of psychoanalysis in the philosophical debate in which Cavell was raised and the momentous role it would come to play in his own mature work is to focus on non-analytic and even non-philosophical sources, especially literary ones. As I hope will become apparent, Cavell's philosophical allegiance to Freud is an important aspect of what he called his “lifelong quarrel with the profession of philosophy” as it stood in the English-speaking half of the philosophical world (Cavell 1984: 31).

## 2. *Philosophy as therapy and self-knowledge*

The most natural place to start my overview is with Cavell’s interpretation of Wittgenstein, his main philosophical point of reference since his doctoral dissertation. In “The Availability of Wittgenstein’s Later Philosophy” (1962), Cavell argues that two crucial aspects of the *Philosophical Investigations* are habitually overlooked: their purpose of fostering self-knowledge and literary style. For Cavell, the aim of a method consisting of reminding ourselves of the statements we ordinarily make about things (see Wittgenstein 1997: I § 90) is to ask the person they are directed to “to say something about himself”, and thereby to produce self-knowledge (Cavell 2002: 61). Here the parallel with Freud is broached for the first time:

> So the different methods are methods for acquiring self-knowledge; as – for different (but related) purposes and in response to different (but related) problems – are the methods of ‘free’ association, dream analysis, investigation of verbal and behavioral slips, noting and analyzing ‘transferred’ feeling, and so forth. Perhaps more shocking, and certainly more important, than any of Freud’s or Wittgenstein’s particular conclusions is their discovery that knowing oneself is something for which there are methods – something, therefore, that can be taught (thought not in obvious ways) and practiced (61).

The purpose of self-knowledge explains why Wittgenstein’s writing, both in its style and in the literary genres it combines and reworks, is so peculiar. Rather than philosophical arguments and demonstrations, we find a pastiche of literary devices (confession, dialogue, rhetorical questions, jokes, parables, etc.) whose aim, rather than to build theories and systems, is to change the reader. At this point, Cavell puts forth a second parallel with Freud:

> his writing is deeply practical and negative, the way Freud’s is. And like Freud’s therapy, it wishes to prevent understanding which is unaccompanied by inner change. Both of them are intent upon unmasking the defeat of our real need in the face of self-impositions which we have not assessed (§ 108), or fantasies (‘pictures’) which we cannot escape (§ 115). In both, such misfortune is betrayed in the incongruence between what is said and what is meant or expressed […]. Both thought of their negative soundings as revolutionary extensions of our knowledge, and both were obsessed by the idea, or fact, that they would be misunderstood – partly, doubtless, because they

> knew the taste of self-knowledge, that it is bitter […] the ignorance of oneself is a refusal to know (67).

The 'psychoanalytical' idea of philosophy sketched in the two passages above, the second especially, would be further articulated by Cavell, but never rejected. Take, for instance, the idea that inner change is what is produced by both psychoanalysis and Wittgenstein's method. If we keep in mind Cavell's grammatical piece of wisdom (in the Wittgensteinian sense), according to which 'inner' does not only mean 'hidden', as in the sceptic's closet of consciousness, but also "*pervasive*, like atmosphere" (Cavell 1979: 99), which I take to mean all-embracing and characterizing the overall fabric of an individual, then we can see the idea coming back on many occasions. In "Aesthetic Problems of Modern Philosophy" (1965), such inner change is described in terms of a "revolution", a simultaneous reconception of the subject and its world (Cavell 2002: 79–80). In *The Claim of Reason*, Cavell describes the transformation he has in mind with the concept of "rebirth" (Cavell 1979: 125); in *Conditions Handsome and Unhandsome* and other texts of the same period, he stresses notions such as "transfiguration and conversion" (Cavell 1990: 36). Their conscious political, spiritual and religious overtones notwithstanding, these terms are appropriately read through the above-mentioned psychoanalytical lenses.

Further instances of continuity can be recalled. The conception of ordinary language philosophy as aiming at self-knowledge is at the core of the crucial section of *The Claim of Reason* devoted to "projective imagination" (Cavell 1979: 145–54). Moreover, on several other occasions, in the same book, Cavell renews the attempt, made in the previous quote when referring to "self-impositions" and the "refusal to know" as the enemies of self-knowledge, to translate into philosophical terms the Freudian notions of resistance, repression and defence mechanism: Freud and Wittgenstein are mentioned as examples of "serious criticism" of human conduct, aware of how "tenacious" a point of view (a Wittgensteinian "picture") can be and how much more than logical coherence can be at stake in it (166). Later in the book, some version of the psychoanalytical concept of rationalization seems to be at play when Cavell argues that Wittgenstein's method is able to problematize "the justifications and explanations" we give ourselves, our ways of "trying to intellectualize our life" and our "critical super-egos" (175).

In the two texts by Cavell directly devoted to Freud, this stance on philosophy and psychoanalysis is reprised with some significant additions. Both in his "Psychoanalysis and Cinema", delivered in 1985 and republished in 1996 in *Contesting Tears*, and in a lecture on Freud delivered throughout the 1990s and published in his *Cities of Words* (2004), Cavell asserts that psychoanalysis should be understood as Freud's "fulfilment" of and "inheritance" from philosophy. Now the scope of the parallel goes well beyond Wittgenstein. “Psychoanalysis and Cinema” sees Freud as working within and making “concrete” the German-speaking line of philosophy initiated by Kant and running through figures like Fichte, Schelling, Hegel, Schopenhauer and Nietzsche. Cavell goes as far as to say that Freud should be considered a third way, alternative to those of Heidegger and Wittgenstein, of inheriting classical German philosophy and its focus, starting with Kant, on the conditions of possibility of human experience (Cavell 1996: 95–7). In *Cities of Words*, given the perfectionist framework to which I will soon return, the parallel extends even further, as Freudian analysis is linked to Socratic maieutic practice, Plato’s allegory of the cave and the Emersonian notion of self-reliance.

In both texts, Cavell addresses the reasons adduced by Freud for his distrust towards philosophy. It is as if, having ignored or dodged the reservations on psychoanalysis expressed by Wittgenstein in his *Lectures and Conversations* (to which I will briefly return later), he were now in the business of smoothing Freud’s own symmetrical doubts. Cavell suggests that Freud’s repeated gesture of distinguishing himself from philosophy and the frequent accusations levelled at philosophers of unabashedly ignoring the unconscious are not only inconclusive (did Nietzsche, for example, really ignore the unconscious?), but suspicious. By ironically using Freud’s own argumentation against himself (“If he had to deny it [a closeness of psychoanalysis to philosophy] so firmly, there must be strong reason to affirm it”; Cavell 2004: 282), Cavell argues that his “competition” with philosophy is indeed ambiguous: rather than simply a wish to “replace” or do away with philosophy, it could be interpreted as the not-so-veiled proposal to translate philosophy into psychoanalysis. Conversely, this would seem to entail the intention of “conceiving psychoanalysis as philosophy” (Cavell 1996: 92; 2004: 290).

## 3. *Psychoanalysis, perfectionism and education*

If we shift our attention from the methodological and meta-philosophical dimension recalled so far to Cavell’s moral reflections, we find a further articulation of this psychoanalytically inclined insistence on inner change and self-knowledge.

In “Part Three” of *The Claim of Reason*, Cavell had already argued that, “because the self is not obvious to the self”, the rationality of morality should be seen as lying “in following the methods which lead [..] to a knowledge and definition of ourselves” (Cavell 1979: 312). The nexus between morality and self-knowledge is further articulated in Cavell’s later books on Emersonian perfectionism, where Freud once more plays a prominent role. Not only does the father of psychoanalysis figure in the little ‘canon’ of perfectionist works and authors sketched early on in *Conditions Handsome and Unhandsome* (Cavell 1990: 5) and reprised with some changes in the structure of *Cities of Words*, more importantly, the relationship between psychoanalyst and patient is taken both as isomorphic to (in *Conditions Handsome and Unhandsome*) and a paradigmatic example of (in *Cities of Words*) a perfectionist moral intercourse.

Cavell describes perfectionism as the "dimension or tradition of moral life" in which the focus is on "some idea of being true to oneself", "becoming intelligible to oneself" and gradually dissipating a "sense of obscurity" (Cavell 1990: 1–2, xxxi). Within this outline of perfectionism, Cavell stresses that a significant other, mostly likened to the Aristotelian friend, has the pedagogic role of catalysing perfectionist moral change, spurring and guiding it. Unsurprisingly, the psychoanalyst is among the figures used to characterize such an educational friendship. In the context of a discussion of Emerson's notion of genius in "Self-Reliance", Cavell suggests that our relationship with a perfectionist 'friend' – in the specific case, a book able to interpret us, to philosophically call us into question – can be seen as taking the form of what Freud calls "transference" (57). *Conditions Handsome and Unhandsome* only makes this connection in passing, almost metaphorically. The Freudian notion of transference, however, had been already referred to by Cavell a few years earlier, in the context of an attempt to justify his idea that reading certain texts can have a therapeutic effect on the reader (Cavell 1984: 52). Furthermore, the parallel is repeated in *Contesting Tears* (Cavell 1996: 113) and systematically articulated in the chapter on Freud in *Cities of Words* (Cavell 2004: 295). A further, related interaction between perfectionism and psychoanalysis identified in *Cities of Words* concerns the concept of education. Given the inherently pedagogical dimension of perfectionism, Cavell finds it interesting that Freud repeatedly described psychoanalysis as "re-education" (290), "a second education [*Nach-erziehung*] of the adult, as a corrective to his education as a child" (Freud 1926).

Additional parallels between perfectionism and psychoanalysis can be identified if we do not limit ourselves to what Cavell states explicitly. Take for instance Cavell's insistence that perfectionism is characterized by "a double picture, or picture of doubleness" of the self (Cavell 1990: xxi–xxiii). Here Cavell is offering a characterization of the Emersonian and Nietzschean idea of the self as always becoming, as split between what it is (the "attained") and what it could become (the "next"). However, from a different but related perspective, Freud's self is also becoming and split, it is equally "double", or even threefold or more. Consider, also, Cavell's idea that the distinctively Emersonian, namely democratic, trait of his perfectionism lies in its envisioning no final state of virtue that is identical and normative for everybody, or path "plottable from outside the journey" (xxxiv). Compare this to when Freud, commenting on the power given to analysts by the transference mechanism, warns them against the temptation to try and mould patients according to their own ideal. This, Freud adds in quite an 'Emersonian' remark, would be to repeat the errors of parents who "crushed their child's independence"; on the contrary, for "all his attempts at improving and educating the patient the analyst must respect his individuality" (Freud 1940: 52).

All this said, it is also hard not to sense a psychoanalytical undertone, or at least an implicit connection with psychoanalysis, in the scenes of "instruction" and the attention to childhood at the centre of the chapter on Wittgenstein in *Conditions Handome and Unhandsome* (Cavell 1990: 64–100); or in Cavell's idea, put forth as early as *The Claim of Reason*, that philosophy can be conceived of as the "education of grownups" (Cavell 1979: 125). Indeed, this Cavellian expression could be a further and more appropriate English translation of Freud's notion of *Nach-erziehung.*

## 4. *A psychoanalytical reframing of scepticism*

A further presence of a form of psychoanalytical thinking, this time mostly implicit, can be identified if we take a chronological step backwards and reflect on the reframing of the problem of scepticism in philosophy launched by Cavell in the two closing essays of *Must We Mean* and in *The Claim of Reason*. I will try to make the psychoanalytical undertones of Cavell's treatment of scepticism apparent through a brief contrast with that of his favourite ordinary language philosophers, Austin and Wittgenstein.

For the Austin of "Other Minds", philosophers asking questions like "how do I know that someone has feelings at all?" are being intellectually wilful (Austin mentions "the wile of the metaphysician") or hazy (they are "barking [their] way up the wrong tree"; Austin 1961: 55, 84). What we need to do in response is to spell out for them the everyday, specific circumstances in which it makes sense to raise doubts about such matters and show how they differ from the circumstances in which their sceptical question was asked (see also Cavell 1979: 49–64). Wittgenstein's stance is perhaps more nuanced, or more avowedly so. Besides passages which support an understanding of the battle against scepticism as "a battle against the bewitchment of our intelligence by means of language" (Wittgenstein 1997: I. § 109), thus as a mainly intellectual matter, others focus on attitudes and resistances ("What has to be overcome is not difficulty of the intellect but of the will"; Wittgenstein 2005: 300e). However, what Wittgenstein's critical method ultimately does is point out the "images", "analogies" and "misunderstandings concerning the use of words" (Wittgenstein 1997: I. § 90) that lead us astray. Even if we accept the therapeutic inflection of readings of the *Investigations* such as Cavell's (and others': see Baker 2004), it remains true that, once these images, analogies and misunderstandings are identified, the research seems to have reached its goal: Wittgenstein embarks on no further inquiries into the supposed psychological or existential causes behind their formation. This last level of investigation, on the contrary, is exactly what Cavell can be said to be attempting.

Cavell's strategy is a complex mix of Wittgensteinian exegesis, reflection on human finitude inspired by existentialism and the kind of psychoanalytical philosophizing I have characterized above. For him, rather than idly attempting to refute scepticism (Cavell 1979: 37–48), the critic should try to assimilate the seed of truth that it contains and identify the actual human experiences that scepticism is at the same time expressing and falsifying through their intellectualization. Indeed, we are not far from Freud's stance on dream-work, fantasies, slips and symptoms more generally, or from his notion of "working through" (*Durcharbeiten*), to which Norris has already directed his attention (Norris 2017: 68).

The seed of truth is that our relationship with the world and others, in Cavell's famous formulations, is not one of knowledge and certainty, but of "acceptance" and "acknowledgment" (Cavell 2002: 298; 1979, 45–7). For Cavell, this truth is contained in and expressed by sceptical thinking, albeit, one might say, unwillingly or unconsciously. In fact, scepticism's own self-representation falsifies this aspect of our existential condition (the reference to Heidegger's concept of '*existentiale*' is explicit: Cavell 2002*:* 243) and covers it with something else. Rather than deriving from our human finitude and separateness, scepticism interprets it in intellectual terms, as a failure of our knowledge. Cavell – in yet another move possibly inspired by Freud – understands this shift of attention on the sceptic's part as a rationalization of a wish for something else: for example, to escape the responsibility of maintaining the forms of life we share (Cavell 1979: 109);  to avoid the burden of having to respond to the pain of others (342); to repress awareness of the contradictions of one's own position towards others' humanity (the slave-owner: 372–8); to find refuge in a fantasy of privacy that, by thwarting any possibility of others knowing us, falsely suggests that we cannot fail to know ourselves (109, 351); etc.

On Cavell's account, scepticism is – again, quite psychoanalytically – a denial, an avoidance of something which, in a sense, we still cannot fail to know: that we know each other well enough, that the real challenge is to acknowledge each other, to recognize ourselves and let ourselves be recognized (Cavell 2002*:* 252). Hence, rather than refutation, the therapy of scepticism entails the pointing out (on the part of the critic/therapist) and acknowledgement (on the part of the sceptic/patient) of what had hitherto been an object of avoidance (of Freudian 'repression').

## 5. *Freud in Cavell's critical writings*

*Disowning Knowledge* (1987) furthers the inquiry into the violence and blindness (towards oneself and others) that derives from attempts to escape the existential frailty of the human condition. That Cavell analysed this condition not only in Wittgensteinian and Heideggerian, but also in growingly explicit psychoanalytical terms, is made especially clear by the essay on *Hamlet*.

The necessity to accept human finitude is described in this chapter in terms of the concepts of 'individuation' and 'separation' that can be seen as reminiscent of Carl Gustav Jung and Melanie Klein (Cavell 2003*:* 188–9). It is also explicitly connected, through Laplanche and Pontalis' reinterpretation (1968), to Freud's concept of the primal scene as discussed in the famous case study of the Wolf Man. Hamlet is interpreted by Cavell as harbouring fantasies about the sexual relationship between his mother and father that are attempts to make sense, as any child grappling with the primal scene wants to do, of his own origin as a finite and separate being (Cavell 2003: 184–8; as shown by Alfano 2018, separation and dependency, with regard to the child's relationship with the mother in particular, are topics that Cavell probably derived specifically from Melanie Klein and Donald Winnicott). The whole essay on *Hamlet* is indeed built in Freudian terms, with a brilliantly counterintuitive interpretation of Hamlet's play within a play based on the mechanisms of inversion, displacement and condensation that Freud found typical of dream-work. Furthermore, the essay gives a 'Freudian' name – "deferred representation", in mimicry of Freud's notion of "deferred action" – for the dramaturgical structure Cavell identified in *Othello* and finds again in *Hamlet* (Cavell 2003: 132–3, 189–91).

A look at the index of names in *Disowning Knowledge* reveals that Freud is the most referenced author, significantly more so than Wittgenstein even. The 23 occurrences – covering topics which range from the death drive to the Oedipus complex, from incest to hysteria and the relation between jokes and the unconscious – do not need to be recalled here. Other references to psychoanalysis are unaccredited, for instance, a possible implicit use of Jung's concept of shadow in the interpretation of a line uttered by the Fool in *King Lear* (283).

There is also another line of psychoanalytic thinking running through Cavell's book. It can be found in each of the essays therein, but with unrivalled clarity in the ones on *King Lear* and *Othello*. Cavell's psychological character analysis often entails an appeal to the reader to try to make sense of behaviour and evaluations on the part of Shakespeare's protagonists which seem utterly irrational and impossible to understand: how can Lear prefer Goneril and Regan's affected declarations of love over Cordelia's stern sincerity (57)? How can Othello put his trust in Iago, the epitome of dissimulation, rather than in the loving Desdemona (133)?

Cavell's answer – as Freudian as the question itself – is that the two characters unconsciously want to be deceived, because what the deceiver allows them to believe (in Othello's case, that Desdemona is unfaithful; in Lear's, that his favourite daughter does not love him) is – however terrible – still preferable to some scenario that they unconsciously fear even more (that Othello took away Desdemona's purity along with her virginity; that a crownless Lear has nothing which can ensure him or pay back Cordelia's love: 57–61, 133).

A brief quotation from Freud's above-mentioned case history of the Wolf Man will make the parallel clear. Explaining that his patient preferred to believe the horrid thought that his relief at his sister's death was due to economic reasons (he could inherit the entire family estate) rather than deeper feelings of competition for the love of their father, Freud remarks:

> now I am the only child and my father must love me and me alone [...] while the thought in itself was entirely capable of becoming conscious, its homosexual background was so unbearable that it was easier to disguise it as filthy greed, for this no doubt came as a great relief (Freud 2002: 281).

Both Freud and Cavell explain a seemingly inexplicable preference for a doomed condition on the basis of an attempt to avoid something that the unconscious deems even more frightening.

I will not dwell at length on *Pursuits of Happiness*, Cavell's book on what he calls the film comedy of remarriage. Suffice it to say that a sentence from Freud's *Three Essays on the Theory of Sexuality* ("The finding of an object is in fact the refinding of it") is not only mentioned in the epigraph, but can be seen as expressing the core intuition behind the idea of remarriage, with its perfectionist insistence on the transformation of incestuous intimacy into socially sanctioned erotic union. Moreover, a few pages later, it is suggested that the comedies of remarriage are moved by a fundamental awareness of the conflict between "eros and civilization" (the implicit reference is to Marcuse's text of the same name and thus to Freud's *Civilization and Its Discontents*; Cavell 1981: 43, 64–5; see also 89–90).

I shall rather focus on *Contesting Tears* (1996), which deals with psychoanalysis in an even more overt and extensive manner. In Cavell's essays on the cinematic melodrama of the unknown woman, with their interest, along with scepticism and perfectionism, in topics such as sex, love, gender, homosexuality and the subjugation of women, Freud is once again an overarching presence. Even more than in *Pursuits of Happiness* and *Disowning Knowledge*, certainly more than in *Must We Mean* and *The Claim of Reason*, here Cavell appears willing to engage with the 'technical' details of psychoanalytical theory, in a direct albeit not exclusive address to an audience of psychoanalysts and psychiatrists

– this was literally true of the spoken version of chapters 2 and 3 of the book, delivered in 1985 and 1984 respectively at the Washington School of Psychiatry and the Columbia Psychoanalytic Center in New York (Cavell 1996: xi).

The range of Cavell's use of psychoanalytical concepts is, again, broad: it spans from the 'local' Freudian interpretation of the hat symbolism in *Now, Voyager* in chapter 3 (119), to the more substantial hypothesis, sparked by a contrast in chapter 4 between Freud and Lacan's understandings of the dissolution of the Oedipus complex, that the threat of castration could offer a developmental explanation of a male wish to control "the woman's voice" (179–90), or to the equally substantial refutation of attempts, typical of some Lacanian school and feminist film theory, to use a Freudian and Marxian concept of fetishism to characterize the overall workings of cinema and thus demonstrate its inherent patriarchality (207–10, 219).

The broadest and most systematic treatment of psychoanalysis in the book, however, can be found in its second chapter, entitled "Psychoanalysis and Cinema". Some of its ground has already been covered during my preceding discussion of the relation between psychoanalysis and Cavell's therapeutic idea of philosophy. What I left out, however, was Cavell's articulation of the relationship between psychoanalysis and scepticism, here given its most explicit treatment of all Cavell's oeuvre.

Cavell's complex and certainly idiosyncratic take on the topic can be summarized as follows: if scepticism can be understood, as Cavell proposes, as the human subject's metaphysical repression of its own intimacy with itself and with the other, two crucial aspects of psychoanalysis make it uniquely fit to address scepticism and even offer a therapy for it. The first aspect is the Freudian understanding of the mind as sunk into the unconscious, or the unconscious as the starting point of any therapeutic enterprise. The second is Freud's systematic study, first sparked by the interest in hysteria, of the relation between somatic symptoms (the body) and their repressed or forgotten psychological causes (the soul), to put it differently, his understanding of the body as inherently "expressive of mind" (105). Freud, with his psychodynamic explanations of organic pathologies and notions such as that of 'somatic compliance', was all too aware of the truth expressed by Wittgenstein when he wrote that "The human body is the best picture of the human soul" (Wittgenstein 1997*:* II. § IV) – a sentence that for Cavell is both anti-behaviourist and deflationary of other-mind scepticism (Cavell 1996: 104).

A sentence from this chapter can be used here to sum up many of the ideas touched upon in the first half of this article: "The advent of psychoanalysis", writes Cavell, "is the place, perhaps the last, in which the human psyche as such, the idea that there is a life of the mind […] receives its proof" (94).

## 6. *Some sources on psychoanalysis*

Cavell's work of the 1980s and 90s shows a substantial engagement with the debate inside and on psychoanalysis. References range from Jacques Lacan and the object relations theories of Melanie Klein and Donald Winnicott, to Jacques Derrida and film and literary critics influenced by psychoanalysis (often through post-structuralism) and engaged in feminism and queer and gender studies, such as Shoshana Felman, Linda Williams, Eve Kosofsky Sedgwick and Janet Adelman. If we slide back to a previous phase of Cavell's writing, however, the situation changes significantly. Explicit references to psychoanalysis in Cavell's books and essays of the 1960s and 70s are restricted to the sole Freud, an indication that at that time, with French post-structuralism's penetration into the literature departments of North American universities still absent or moving its first steps, Cavell's interest in psychoanalysis was being pursued 'privately', with very few interlocutors in the academic world in which he was immersed.

Cavell would have agreed with psychoanalyst and philosopher Marcia Cavell, his wife until 1961, when she wrote in 2006 that "neither in Great Britain nor in the United States has philosophy been much affected by psychoanalysis" (M. Cavell 2006: 5). Obviously there are exceptions, for example, Richard Wollheim and Stuart Hampshire (see for instance Wollheim *et al.* 1982), or the Richard Rorty of *Contingency, Irony, and Solidarity* (1989). But most of them, at least within the analytic style or tradition of philosophy, came quite late – again, in the 1980s and 90s, thus more than two decades after Cavell's first philosophical attempt to tackle Wittgenstein and Freud. Most importantly, they were mainly concerned, as argued by Levine, with whether psychoanalysis should be considered scientific, a topic that was indeed far from Cavell's interest (see for example Grünbaum 1993; Macmillan 1997*;* and the literature review in Levine 2000: 6–7). Cavell was quite explicit on his sense of lacking interlocutors: "Most philosophers in my tradition, I believe, relate to psychoanalysis, if at all, with suspicion, habitually asking whether psychoanalysis deserves the title of a science. I am not here interested in that question" (Cavell 2004: 286).

The only work dealing with Freud and coming from the analytic tradition that Cavell actually mentions is John Wisdom's *Philosophy and Psychoanalysis*, published in 1953 and collecting essays first published between 1933 and 1948. Wisdom highlights some similitudes between psychoanalysis and aspects of the idea of philosophy he was championing under the influence of Wittgenstein: both philosophy and psychoanalysis try to bring out models and fantasies that unconsciously dominate our thought, both use paradoxical sentences to make us see things that were already in plain sight in new ways, and both can free us from "mental cramps". Wisdom even draws a parallel between the neuroses and psychoses dealt with by the psychoanalyst and the metaphysical fixations of the other-mind and external-world sceptic (Wisdom 1969: 169–81, 248–82). It is thus reasonable to think that Cavell drew some inspiration from Wisdom's ideas for his therapeutic interpretation of Wittgenstein. However, Wisdom's parallel seems of a more limited scope than Cavell's. Its stress falls more on the theoretical possibility of seeing things differently than on what most interests Cavell, namely self-knowledge and personal 'inner' change. This probably explains why Cavell's references to Wisdom are so cursory: the essays from *Philosophy and Psychoanalysis* are mentioned by Cavell but three times, always in footnotes, and only in his very early essays on Austin (1958) and Wittgenstein (1962); even more significantly, they are never referred to in the actual passages in which Cavell articulates his parallel between Wittgenstein's and Freud's methods (Cavell 2002: 18–19, 37, 54).

It is also significant that, despite all his insistence on Wittgenstein, Cavell never discusses, or even mentions, his remarks on Freud recorded in *Lectures and Conversations*, published in 1967. We can surmise that, even if he had read them, he would not exactly have known what to make of their ambivalent stance. That Wittgenstein famously defined himself to Rush Rees as a "Freudian" would confirm the parallel put forth by Cavell in 1962. However, when Wittgenstein accuses psychoanalysis of the same tendency towards unwarranted generalization that the *Investigation*s describe as a philosophical malady, he can be seen as situating himself in the exact same strain of reflection on the epistemic status of psychoanalysis that, as I remarked, Cavell found alien to his interests.

This lack of interlocutors can also be partially explained by the success of behaviourism in American academia. As argued by Mahoney, the 1960s were "expansive and exciting" years for behaviourism, when "the ruling authority of psychoanalysis" was a primary polemical target (Mahoney 1984: 303–4). Harvard University, where Cavell was a PhD candidate between 1951 and 1961 and taught from 1963 onwards, was no different: B. F. Skinner, a pivotal figure of behaviourism, taught there, and there Cavell also came into personal contact with philosophers influenced by behaviourism such as W. V. O. Quine, who was a faculty member, and Gilbert Ryle, who visited from Oxford and whose *The Concept of the Mind* was a common topic of discussion (Cavell 2010: 247–8, 281, 290). Cavell also recounts that in 1948, as a student at UCLA, he had taken some psychology courses, to his dismay finding them to be dominated by experimentalists and entirely hostile to Freud (242).

In fact, in the account offered in his autobiography of his early years as a student and young teacher in UCLA, Berkeley and Harvard between the late 1940s and 60s, only two names of colleagues and friends interested in psychoanalysis stand out. The first is that of Kurt Rudolph Fischer, a Jewish-Austrian 'continental' philosopher who taught in Berkeley before 1967 and later published (mostly in German) on topics such as existentialism, Nietzsche and psychoanalysis (*Little Did I Know*, 343–52; also see *Contesting Tears*, 225). The second is Michael Fried, indeed not a philosopher but an art critic and historian: Cavell declares a debt both to his work on artistic modernism (Cavell 2010: 406–7) and his use of psychoanalysis in criticism (Cavell 1996*:* 223).

What is left is Cavell's engagement with psychoanalysis *outside* academia in the 1960s and 70s. He underwent analysis personally twice in his life: the first course started in Berkeley around 1960, the second in Boston in 1976 when he was grappling with the writing of *The Claim of Reason*, whose manuscript even became material for the sessions (Cavell 2010*:* 108–9). Somewhere towards the end of the 1970s Cavell also began training as a therapist at the Boston Psychoanalytic Institute. It is perhaps revealing of the distrust between philosophy and psychoanalysis at that time that restrictions imposed on him by senior members of the institute finally induced him to drop out of the programme. It is equally telling of Cavell's intellectual temperament that this institutional dead-end prompted him to pursue, as he writes, the "therapeutic registers in his writing" and the "therapeutic impulse" in himself with even more determination (Cavell 2010: 512–4).

## 7. *On the possible influence of Lionel Trilling*

One last step back will take us to the time of Cavell's first discovery of Freud. It was autumn 1947 and the 21-year-old Jew from Atlanta had barely enrolled on the graduate programme in music composition at the Julliard School of New York, when the deep vocational crisis began that would reroute him to philosophy. The few months spent in New York preparing the application for Julliard and then mostly avoiding its classes were in reality mainly dedicated, Cavell writes, to "reading whatever it was that people called philosophy", which he had heard "had something to do with examining one's life" (Cavell 2004: 282). But what 'philosophy' was Cavell actually reading at the time? It was first of all Freud, which Cavell dived into "ten to twelve hours a day", starting from the *Introductory Lectures on Psychoanalysis* (Cavell 2010: 185). Besides Freud, his main reading material was *Partisan Review*, the major journal of the so-called Anti-Stalinist Left of New York (see the interview with Cavell in Borradori 1994: 118–36). This simultaneity offers us a trail, thus far overlooked by scholars, that I would like to follow in this closing section.

A brief detour is needed here. By 'Anti-Stalinist Left', alternatively referred to as the 'New York Intellectuals', intellectual historians mean a group of scholars, critics and literary authors publishing, roughly between the late 1930s and early 80s and reaching their apogee in the 1940s and 50s, in cultural journals such as *Partisan Review*, *Commentary* and *Dissent* (see Bloom 1986). Their ranks included the Columbia University professor and literary critic Lionel Trilling, philosophers such as the pragmatist Sidney Hook and the German-born Hannah Arendt, art critics such as Clement Greenberg and Meyer Schapiro, as well as novelists (for example Saul Bellow and Norman Mailer), film critics (Robert Warshow) and sociologists (Nathan Glazer). Mostly American Jews born, like Cavell, of Eastern European immigrants, they were characterized by an intellectual brand that can be synthesized here under three aspirations that find parallels in Cavell's mature work: 1) reconciling the acquisitions of modernism in the arts and literature with progressive politics; 2) assimilating European high culture into the American experience and its democracy, as well as finding and legitimizing its American equivalents; and 3) putting the critical intellect to work outside the disciplinary and institutional boundaries of academia, which for them meant writing not only as scholars but also as public intellectuals.

Cavell testifies to having discovered the New York Intellectuals in 1947 and having been a regular reader of their journals between 1948 and 1951, when he was a student at UCLA (see Cavell's afterword in Warshow 2001: 290). Many traces in his books and interviews, however, tell us that Cavell kept reading them for his entire intellectual life. Clement Greenberg's reflections on modernism in the visual arts, through the mediation of Michael Fried, had a strong influence on Cavell's writing on the same topic between 1965 and 1971 (see Cavell 2002: 68–90, 167–96, and especially 171, 201; 1979: 108–18, especially 113). Robert Warshow's critical essays on Hollywood films and American pop culture can be seen as an inspiration for much of what Cavell wrote in *The World Viewed* (1971). In addition, they were explicitly drawn upon in "Film and the University", the appendix to *Pursuits of Happiness* (1981), and celebrated in Cavell's afterword to the reissue of the collection of Warshow's essays *The Immediate Experience* (2001). Lionel Trilling's literary and cultural criticism too is praised persistently. In 1980, five years after Trilling's death, Cavell participated in the *Lionel Trilling seminars*, a permanent seminar at Columbia University, and described Trilling as someone "whose work has been so nourishing to me, from the time I began searching for my way into the world of mind" (Cavell 1984: 188). In 1981, one of the essays of *Pursuits of Happiness* made reference to "The Fate of Pleasure", one of Trilling's most important contributions on the subject of modern literature (Cavell 1981: 154). In his interview with Borradori in the early 1990s, Cavell described his reaction to Trilling (and Warshow) when he was a young man as an "ecstatic experience" (Borradori 1994: 122). Finally, Cavell's autobiography is replete with anecdotes and passages that confirm his prolonged appreciation of the Anti-Stalinist Left as a group (Cavell 2010: 11, 158, 185, 231, 242, 252, 299, 406–7).

Freud and psychoanalysis were a frequent subject for the New York Intellectuals. *Partisan Review* editor William Barret, for example, in the very year 1947 authored an imaginary dialogue between Freud and Heidegger on *angst* and authenticity (Barrett 1947). However, I would like to focus on the figure in the group most deeply and persistently engaged in psychoanalysis, namely Lionel Trilling, formulating the hypothesis that Cavell's youthful interest in Freud was sustained and prolonged by the steady flow of works that Trilling devoted to the psychoanalyst: for example, "Freud and Literature" (1940, revised version in September 1947), "Art and Neurosis" (1945), "Neurosis and the Health of the Artist" (December 1947), "Two Analyses of Sigmund Freud" (December 1947), "Freud's Last Book" (1949), *Freud and the Crisis of Our Culture* (1955) and "A Review of the Correspondence Between Sigmund Freud and C.G. Jung" (1974), as well as the numerous passages on Freud in *Sincerity and Authenticity* (1972), the transcription of Trilling's Norton Lectures at Harvard in 1970.

Cavell never references these texts and, although he was teaching at Harvard at the time of Trilling's lectures, we have no way of judging whether he actually followed them. Therefore, besides my previous considerations on Cavell's appreciation of Trilling and the New York Intellectuals in general, my hypothesis will be based only on thematic correspondences between Cavell's treatment of Freud and that of Trilling, as well as the indirect but significant clue offered by the chronological correspondence of the issue or reissue, in autumn 1947, of three of Trilling's above-mentioned essays, with Cavell's own discovery of Freud in the same months. I will highlight some of such thematic correspondences, focusing especially on essays published by Trilling before 1962.

Trilling, who in 1955 was the first literary critic – indeed the first layman – ever asked to give the annual Freud Anniversary Lecture at the New York Psychoanalytic Society (the previous year had been the turn of Anna Freud), repeatedly insisted on the inherent similarities between psychoanalysis and literature. Despite appreciatively mentioning aspects of Ernest Jones' analysis of *Hamlet* and Freud's analysis of *King Lear*, Trilling finds himself dissatisfied, like Cavell some decades later, with the overt orthodoxy of psychoanalytical literary criticism (Trilling 2008: 39, 46–52; Cavell 1996: 91). Indeed, Trilling does not connect psychoanalysis to literary criticism, but to literature itself, the output of novelists, playwrights, and poets.

The introductory paragraph of "Freud and Literature" (1940/1947) opens with the words: "The Freudian psychology is the only systematic account of the human mind which […] deserves to stand beside the chaotic mass of psychological insights which literature has accumulated through the centuries". The same paragraph closes with a reminder of Freud's statement that it was not he, but the poets who discovered the unconscious, while all he did was find "the scientific method by which the unconscious can be studied" (Trilling 2008: 34). Some 60 years later, Cavell would insist on exactly the same idea: after characterizing Freud's achievement as an "unsurpassed horizon of knowledge about the human mind", Cavell recalls that Freud "likes to insist that his insights into the human mind have been anticipated by the creative writers of our civilitazion" and that he only "systematized the culture's power of insight into a new science" (Cavell 2004: 286–7).

In "Freud and Literature", Trilling states that the greatest merit of psychoanalysis lies in having made "poetry indigenous to the very constitution of the mind" and that "psychoanalysis is one of the culminations of the Romanticist literature of the nineteenth century" (Trilling 2008: 35, 52–3). Cavell, in a passage of *In Quest of the Ordinary,* would remark in passing that there is "a sense in which [psychoanalysis] was preceded by romanticism" (Cavell 1988: 48). Even on the subject of Freud's positivism, Trilling and Cavell are close. Trilling insists on the ambivalence of its effects: "From his rationalistic positivism", he writes, "comes much of Freud's strength and what weaknesses he has", the weakness lying mainly in Freud's tendency to understand art only as a substitute gratification for repressed drives (Trilling 2008: 40–6). Cavell is less dismissive, but shares Trilling's idea of a tension, within Freud's work, between the scientific and the 'literary' mind. In the chapter on Freud in *Cities of Words* (2004), it is almost as if Cavell were trying to defend Freud from Trilling's critique, while at the same time maintaining Trilling's view on the limitations of a positivistic understanding of psychoanalysis: he states that the reason why his chapter focuses on Freud's text about Jensen's *Gradiva* is that in it Freud most explicitly insists on his break from "advanced Western thought, as represented in philosophy and established in science", and "his continuity with the high literary tradition of Western culture" (Cavell 2004: 283).

Also close to what Cavell would write decades later is the idea, put forth by Trilling in *Freud and the Crisis of Our Culture* (1955) and later expanded in *Sincerity and Authenticity* (1970), that psychoanalysis and literature share an understanding of the self which is typically 'modern'. For Trilling, Freudian psychoanalysis – especially *Civilization and its Discontents* – epitomizes an understanding of human subjectivity first born out of Shakespeare's theatre and later developed by the novel: the modern self is understood as always in development, as always caught in a complex dialectic, an antagonism even, between the inner and the outer, individual and society, social masks and what lies behind them (Trilling 1955: 33–58). This same dialectic would be described by Cavell in terms of his modified notions of scepticism and perfectionism. The similarities are significant: not only would Cavell go along with the – certainly widespread – narrative that traces back the "modern self" to Shakespeare (Cavell 2003: 3), he would also, as I recalled above, draw an idiosyncratic connection (but perhaps less so in light of its possible source in Trilling) between the history of modern scepticism and psychoanalysis (Cavell 1996: 104–5).

One further aspect of Trilling's assessment of the common genius of literature and psychoanalysis finds resonance in Cavell's work. For Trilling, literature and psychoanalysis are at one in that they share a moral and epistemic commitment to the other. They presuppose and at the same time instil a capacity to see the inner life of other human beings, to accept it and take it into account, recognizing its reality both in its separateness and its relation to us. In his 1955 lecture, Trilling describes the capacity of the literary author and the psychoanalyst to "imagine the selfhood of others", to come to the "realization of the selfhood of others in pain", to affect a "willing suspension of disbelief in the selfhood of someone else", and declares it "the essence of moral life" (Trilling 1955: 18–19). From the closing essays of *Must We Mean What We Say* onwards, Cavell would explore this same capacity, a cornerstone of our moral life for him too, in terms of his philosophy of acknowledgement. Among the cultural forces able to foster the human capacity to acknowledge the subjectivity of other human beings, Cavell would count not only psychoanalysis and literature, but also film and – most importantly – philosophy itself, at least in the therapeutic conception of it that he elaborates by joining together aspects of the work of Wittgenstein and Freud.

This quick comparison does not cover the whole range of either Trilling's or Cavell's ideas on Freud. Certainly, it is not able to demonstrate a direct influence, especially given the already mentioned absence of explicit references as well as the conspicuous time lag that occurred between the publication of Trilling's ideas on Freud and Cavell's alleged reworking of them.

Should the hypothesis of an influence be found worthy of consideration, the lack of references and belated reception can perhaps be interpreted in the following two non-mutually exclusive ways. Cavell explicitly stated that his "lifelong quarrel with the profession of philosophy" as it stood in the American academia of the time certainly entailed a wish to open it to new possibilities (continental philosophy, psychoanalysis and literature), but never to simply do away with its epistemic and communicative paradigms, insofar as he recognized them to be a "genuine present of philosophy" (Cavell 1984: 31–2). Thus it can be surmised that the relaxed, non-academic cultural criticism of Trilling (and other New York Intellectuals) was too far removed from the philosophical academia of Cavell's time to be put to work in the writing through which Cavell was seeking to find a distinctly personal, but also legitimate and institutionally sanctioned place in the philosophical world. Cavell almost states something of the sort to explain why, despite the great influence the Anti-Stalinist left had exercised on him, at a certain point he had to leave it behind in favour of his professional philosophical work (Borradori 1994: 122).

It is also possible that, having been interiorized by Cavell in his early years, Trilling's ideas on Freud worked as a silent co-incentive to reflect on psychoanalysis in a time when the only philosophical avenue at Cavell's disposal was the therapeutic inflection in the likes of Wisdom and Wittgenstein. Then, by the time of Cavell's more mature reflection, when his securer position in the philosophical world and the encounter with post-structuralism and the new currents of literary and film criticism had permitted wider philosophical engagement with psychoanalysis, Cavell worked some versions of Trilling's ideas into his writing while no longer fully aware of their exact provenance. Reception can work this way, especially in case of the cross-disciplinary dissemination of the work of culturally influential intellectual figures such as Freud, whose pervasive presence in American culture until the 1980s, outside academia at least, is unquestionable (Hale 1995), and Trilling, who, according to a survey circa 1970, ranked among the ten most influential intellectuals in the United States (Rodden 1999: xxxiv).

If, on the contrary, the evidence backing the hypothesis of Trilling's influence on Cavell is found to be thin, his treatment of Freud as sketched out above can at least be considered a representative example of the kind of work on psychoanalysis in American literary circles to which Cavell might have been exposed in the decades of his education and early philosophical production.

Raffaele Ariano
Facoltà di Filosofia, Università Vita-Salute San Raffaele di Milano
ariano.raffaele@hsr.it

## *References*

Alfano, Chiara, 2018, "Toward an Ordinary Language Psychoanalysis: On Skepticism and Infancy", in *New Literary History* 49, 1: 23-45.

Assif, Adeena, 2020, "'The Dialogue of the Mind with Itself': Freud, Cavell, and Company", in *Common Knowledge* 26, 1: 12-38.

Austin, John Langshaw, 1961, "Other Minds", in *Philosophical Papers*, Clarendon Press, Oxford: 44-84.

Baker, Gordon, 2004, *Wittgenstein's Method: Neglected Aspects*, Blackwell, Oxford.

Barrett, William, 1947, "Dialogue on Anxiety", in *Partisan Review* 14, 2: 151-159.

Bloom, Alexander, 1986, *Prodigal Sons: The New York Intellectuals & Their World*, Oxford University Press, Oxford.

Borradori, Giovanna, 1994, *The American Philosopher: Conversations with Quine, Davidson, Putnam, Nozick, Danto, Rorty, Cavell, MacIntyre, and Kuhn*, The University of Chicago Press, Chicago.

Cavell, Marcia, 2006, *Becoming a Subject: Reflections in Philosophy and Psychoanalysis*, Clarendon Press, Oxford.

Cavell, Stanley, 2002, *Must We Mean What We Say: A Book of Essays* [1969], Cambridge University Press, Cambridge.

—, 1979, *The World Viewed: Reflections on the Ontology of film* [1971], Harvard University Press, Cambridge MA.

—, 1979, *The Claim of Reason: Wittgenstein, Skepticism, Morality, and Tragedy*, Oxford University Press, Oxford.

—, 1981, *Pursuits of Happiness: The Hollywood Comedy of Remarriage*, Harvard University Press, Cambridge MA.

—, 1984, *Themes Out of School: Effect and Causes*, The University of Chicago Press, Chicago.

—, 2003, *Disowning Knowledge: In Seven Plays of Shakespeare* [1987], Cambridge University Press, Cambridge.

—, 1988, *In Quest of the Ordinary: Lines of Skepticism and Romanticism*, University of Chicago Press, Chicago.

—, 1990, *Conditions Handsome and Unhandsome: The Constitution of Emersonian Perfectionism,* The University of Chicago Press, Chicago.

—, 1996, *Contesting Tears: The Hollywood Melodrama of the Unknown Woman*, The University of Chicago Press, Chicago.

—, 2004, *Cities of Words: Pedagogical Letters on a Register of the Moral Life*, Harvard University Press, Cambridge MA.

—, 2005, *Philosophy the Day After Tomorrow,* Harvard University Press, Cambridge MA.

—, 2010, *Little Did I Know: Excerpts from Memory*, Stanford University Press, Stanford.

Conant, James, 1989, “Interview with Cavell”, in *The Senses of Stanley Cavell*, edited by Richard Fleming and Michael Payne, Bucknell University Press, Lewisburg: 22-31.

Eldridge, Richard, 2011, “Criticism and the Risk of the Self: Stanley Cavell’s Modernism and Elizabeth Bishop’s”, in *Stanley Cavell: Philosophy, Literature and Criticism*, edited by James Loxley and Andrew Taylor, Manchester University Press, Manchester: 92-105.

Freud, Sigmund, 2002, *The Wolfman and Other Cases,* Penguin Books, London.

—, 1926, “Sigmund Freud on Psychoanalysis”, Encyclopædia Britannica, 03/07/2002, URL = https://www.britannica.com/topic/Sigmund-Freud-on-psychoanalysis-1983319.

—, 1940, “An Outline of Psycho-Analysis”, in *International Journal of Psycho-Analysis*, 21: 27-84.

Gould, Timothy, 1998, *Hearing Things: Voice and Method in the Writing of Stanley Cavell*, The University of Chicago Press, Chicago.

Grünbaum, Adolf, 1993, *Validation in the Clinical Theory of Psychoanalysis: A Study* *in the Philosophy of Psychoanalysis*, International Universities Press, Madison, CT.

Hale, Nathan G., 1995, *The Rise and Crisis of Psychoanalysis in the United States: Freud and the Americans, 1917-1985*. Oxford University Press, Oxford.

Laplanche, Jean and Pontalis Jean-Bertrand, 1968, "Fantasme oridinaire, fantasmes des origins, origine du fantasme", in *International Journal of Psychoanalysis*.

Levine, Michael, 2000, *The Analytic Freud: Philosophy and Psychoanalysis*. Routledge, London.

Mahoney Michael. J., 1984, "Psychoanalysis and Behaviorism", in *Psychoanalytic Therapy and Behavior Therapy*, edited by H. Arkowitz, S. B. Messer. Springer, Boston: 303-325.

Macmillan, Malcolm, 1997, *Freud Evaluated*, MIT Press, Cambridge MA.

Mulhall, Stephen, 1994, *Stanley Cavell: Philosophy's Recounting of the Ordinary*, Oxford University Press, Oxford.

Norris, Andrew, 2017, *Becoming Who We Are: Politics and Practical Philosophy in the Work of Stanley Cavell*, Oxford, University Press, Oxford

Rodden, John, 1999, *Lionel Trilling and the Critics: Opposing Selves*, University of Nebraska Press, Lincoln.

Rorty, Richard, 1989, *Contingency, Irony, and Solidarity*, Cambridge University Press, Cambridge.

Trilling, Lionel, 1945, "Art and Neurosis", in *Partisan Review*, Winter 1945: 41-48.

—, 1947, "Neurosis and the Health of the Artist", in *New Leader* 30(12).

—, 1947, "Two Analyses of Sigmund Freud", in *New York Times Book Review*, 14 December: 4.

—, 1949, "Freud's Last Book", in *New York Times Book Review*, 27 February: 1-17.

—, 1955, *Freud and the Crisis of Our Culture*, The Beacon Press, Boston.

—, 1963, "The Fate of Pleasure", in *Partisan Review*, Summer 1963: 167-191.

—, 1974, "A Review of the Correspondence Between Sigmund Freud and C.G. Jung", in *The New York Times Book Review*, 21 April.

—, 1972, *Sincerity and Authenticity: The Charles Eliot Norton Lectures, 1969-1970*, Harvard University Press, Cambridge MA.

—, 2008, "Freud and Literature", in *The Liberal Imagination* [1950]. New York Review of Books, New York: 34-57 [originally published in *The Kenyon Review*, Spring 1940, and in revised form in *Horizon*, September 1947].

Warshow, Robert, 2001, *The Immediate Experience: Movies, Comics, Theatre and Other Aspects of Popular Culture* [1946], Harvard University Press, Cambridge MA.

Wisdom, John, 1969, *Philosophy and Psycho-analysis* [1953], University of California Press, Berkeley.

Wittgenstein, Ludwig, 1997, *Philosophical Investigations* [1953], Blackwell, Oxford.

—, 1967, *Lectures and Conversations on Aesthetics, Psychology, and Religious Belief*, Basil Blackwell, Oxford.

—, 2005, *The Big Typescript: German-English Scholar's Edition*, Blackwell, Oxford.

Wollheim, Richard, James Hopkins, 1982, eds., *Philosophical Essays on Freud*, Cambridge University Press, Cambridge MA.

# Some arguments against the possibility of an infinite past

Luca Bellotti

*Abstract*: In this brief note we discuss some arguments against the purely conceptual possibility of an infinite past, arguing that they are ungrounded and showing how some points of the contemporary debate can be found in some mid-thirteenth-century controversies on the topic.

*Keywords:* infinite past, eternity of the world, philosophy of time

We will consider some classical arguments proposed by various scholars, including in particular G.J. Whitrow (1966, 1980), P. Huby (1971), W.L. Craig (2000), D.A. Conway (1974), which should demonstrate the purely conceptual impossibility of an infinite past. We believe that these arguments are ungrounded, either because they invalidly draw conclusions from true assumptions, or because they are based on false assumptions. Starting from the critical discussion of these arguments carried out by Q. Smith (1987) and R. Sorabji (2006), now classical in the vast literature on the topic, we will see how the arguments (some of which have, for better or worse, an undeniably Zenonian flavour) are articulated and why they are untenable. We will also show how some fundamental aspects of the contemporary discussion can be found, in all their precision, in the mid-thirteenth-century controversy between Bonaventure and Thomas Aquinas on the problem of the eternity of the world, as well as in the treatise *De aeternitate mundi* by Boethius of Dacia (*circa* 1270).

The arguments we will examine should show, perhaps unsurprisingly, that there is no reason, particularly in the light of the mathematics of infinity of the last 150 years (purely set-theoretic notions will prove to be sufficient below - without denying the possible relevance of mereological, topological, metrical, or measure-theoretic ones for the problem), to deny the purely conceptual possibility of an infinite past. Distinct, and certainly relevant, is the problem of the physical possibility of an infinite past; however, it seems that this problem is on a decidedly different level from the one we want to discuss here, a level in which

Submitted: 11 June 2022
ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i1.418
Accepted: 19 January 2023

considerations based purely on the analysis of concepts are insufficient. What we can conclude in the present context is that the question of the infinity of the past does not seem apt to be decided in the negative with purely *a priori* arguments.

Our main motivation is mixed, both theoretical and comparative (though not strictly historical). Although some more recent literature in the philosophy of time on the question of a possibly infinite past shows interests which are admittedly different from ours, the persistence of arguments like those discussed below in recent debates, with proposals which do not withstand mathematical scrutiny and should, in our opinion, be simply ruled out, shows perhaps the necessity of recalling again some fundamental points (not yet taken as uncontroversial) before any further refinement.

We will consider and re-evaluate the six classical fundamental arguments against the conceptual possibility of an infinite past which were collected and systematized by Smith (1987); these arguments are substantially similar to those examined by Sorabji (2006, Part III, Ch. 4, especially pp. 219-224), but the two discussions are independent. We will see how the two authors critically analyze these arguments, we will compare their solutions, which in more than one case diverge (or in any case start from different points of view), and we will see to what extent they are tenable. Of course, we will refer to further, more recent contributions to the debate when necessary.

1. (Whitrow 1966). If the series of past events is infinite, it must constitute an actual infinity, since the events really happened; but an actual infinity of past events is impossible: there would be events of the past separated from the present by an infinity of intermediate events.

Following Smith (1987), we immediately notice that the argument equivocates about 'actual'. Initially, 'actual' is opposed to 'potential' in the sense of the relationship between the concepts of act and power; subsequently 'actual' refers to the infinity of a series of events such that some of them are separated from the present by an infinite number of intermediate events. But it is well possible that there is an infinite series of events that actually happened, such that each is separated from the present by a finite number of intermediate events. A model of such a series is simply the set of negative integers in their usual order.

2. (Whitrow 1966, Huby 1971, Craig 2000). Recall that $\aleph_0$ is the smallest infinite cardinal number, i.e. the cardinality of the set of natural numbers, as well as of any countable set. The argument is as follows: (1) $\aleph_0$ events happened before the present; (2) Events divided from the present by $\aleph_0$ events occurred; (3) From an event divided from the present by $\aleph_0$ events, this could not have been reached. It is believed that (1) implies (2), which in turn implies (3), whence the absurdity that the present cannot be reached.

Smith (*op. cit.*) notes the incorrectness of the inference from (1) to (2), which is evident once again considering the model of negative integers in their natural order: they constitute a set of cardinality $\aleph_0$ in which any element is separated from zero ('the present') by a finite number of elements. Note that the order of the elements is important: e.g., if we order the negative integers in such a way that all the even ones precede all the odd ones, we will certainly have elements that are separated from others by $\aleph_0$ elements; but this is not the relevant order for the argument. The fact that we deal with the same set is irrelevant: the order properties of a set are in general completely independent from those of an extensionally identical set ordered in a different way, since different orders can be defined on the same set; if two sets are placed in one-to-one correspondence, the order properties of a certain element of the first set are in general independent of the order properties of the corresponding element of the second set.

Sorabji's formulation (*op. cit.*) of the argument we are discussing (further examined and criticized, more recently, by Puryear 2014; see also Morriston 2022) is substantially analogous to Smith's formulation: if an infinity of days had passed before the present day, the latter would never have been able to occur. Sorabji replies that this would be true if there were a first day followed by an infinity of days before reaching today; but those who maintain the possibility of an eternal world *a parte ante* just do not admit the existence of a first day.

It is interesting to note that an argument against the existence of the world *ab aeterno* that falls within this setting can be found in Bonaventure (*Commentarius in quatuor libros Sententiarum Petri Lombardi*, II, dist. 1, pars 1, art. 1, q . 2, pp. 12-17) and is contested by Thomas Aquinas (*Summa contra gentiles*, II, 38, arg. 4; see also *Summa theologiae*, I, 46, 2 and *Scriptum super libros Sententiarum* II, d. 1, q. 1, a. 5; the tract *De aeternitate mundi contra murmurantes* of 1270 is of course also important on the subject). Bonaventure's argument is substantially based on the classic Aristotelian principle (see e.g. Aristotle, *De caelo*, I, 4, 272a3) *impossibile est infinita pertransiri*, which in our case would imply, if the past were infinite, the unreachability of the present. The use of the Aristotelian idea that an actual infinity cannot be traversed, in order to demonstrate the impossibility of an infinite past, actually dates back to Johannes Philoponus, in particular to his *De aeternitate mundi contra Proclum* of 529 AD (see Sorabji, *loc. cit.*, for references). Now, Aquinas accepts Bonaventure's two assumptions: (1) the eternity of the world implies that the present day has been preceded by an infinite number of days; (2) infinity cannot be crossed. However, the conclusion is not what Bonaventure would like: every given day in the past is in fact separated from the present by a finite number of intermediate days. Aquinas's argument is this: if the world is eternal, the past days can be taken either simultaneously, or in succession; if they are taken simultaneously, there is no question of ‘crossing’, since a starting point is missing; if instead they are taken in succession, we can designate one of the past days as the starting point, but in this case the days that must be crossed are finite in number. What therefore divides Bonaventure and Aquinas is a point which, as we have seen above, remains fundamental even in contemporary discussions on the possibility of an infinite past: Bonaventure believes that an infinite series of past days must contain days that are separated from the present by an infinite number of intermediate days; Thomas believes, on the contrary, that an infinity of past days can be real and yet each past day remains separated from the present by a finite number of days, however large it may be.

We must remark, at this point, that the post-Cantorian mathematics of infinity, in its less controversial aspects from a foundational point of view, aspects which are currently widely accepted, cannot avoid to agree with Aquinas in the present controversy. If the pure conceivability of an infinite past is at stake, then there seems to be no reasonable doubt that contemporary set theory offers models such that an infinite past is simply a priori not impossible.

Sorabji (*op. cit.*) attributes the argument just discussed to Bonaventure in the following form: if we think ‘backwards’, starting from the present, we will never find a year at an infinite distance from the present; then the past years are finite in number. It is a question, Sorabji observes, of *ignoratio elenchi*: in fact, no one claims that there are years in the past that are infinitely distant from the present; we have a set of years that are all finitely distant from the present, and nevertheless this set is infinite. This is exactly the argument used by Aquinas against Bonaventure. Sorabji, however, thinks that Aquinas has only partially grasped the truth, in his objection to Bonaventure: Aquinas would have correctly seen that the distance between the present and any past year is in any case finite, but he would have incorrectly deduced that in a universe without a beginning no infinity of years would be gone through. We do not understand what is wrong with this deduction, once we accept the premise that any crossing requires two extremes to be fixed, one initial and one final, a premise in fact assumed by Aquinas, that seems entirely reasonable.

Still in this order of ideas, Sorabji presents the following argument (reported in Sorabji 2006, p. 221, attributing it to P. Huby). An infinity of future years from the present will always remain potential and will never be completed. Why shouldn’t we say the same of an infinity of past years? The answer starts from the lack of analogy between past and future, consisting in the fact that the past does not start now, although certainly our thoughts on the past do. But then, when does the past begin? The answer, consistently with what we have been saying so far, should be clear: the past, under the hypothesis we are considering, does not begin. Not that past and future are inherently asymmetrical;

it is their being ‘crossed’ that presents a crucial difference: while the crossing of future years starting from the present is still something that has two extremes, two boundaries, the ‘crossing’ of the past we are now dealing with has only one boundary, which is the one *a parte post*, and no boundary *a parte ante*. Thus we speak of ‘crossing’ in a metaphorical sense, since, as we have seen, a real crossing requires two boundaries. Arguably, it is already the very notion of ‘crossing’ that can be considered no more than a metaphor; but at least in this context its features are sufficiently clear, intuitively, to allow the refutation of the argument proposed.

Having replied along these lines, Sorabji (*loc. cit.*) makes a remark that seems rather strange: a set of future years starting from the present would become infinite in actuality only if it reached a year infinitely distant from the present; but the same cannot be said of the past. This seems to contradict Aquinas’s argument that the actual infinity of a set does not require that there are elements of the set ‘at infinity’, an argument which, as we have seen, is difficult to refute, given that we have the succession of natural numbers as a simple example. Sorabji’s claim remains all the more strange, since the preceding argument seems correct and does not depend on it in any way, since it is based on the difference, which certainly exists, between going through the past in its totality and going through the future in the sense of pushing forward into the future indefinitely.

3. (Conway 1974, Craig 2000). The set of past events is never complete, but new events are always being added to it; therefore there cannot exist in the past an actual or complete infinity of events. A model could be a library of $\aleph_0$ volumes in which each volume is marked with a natural number: it would be impossible to add a volume to this library. Two assumptions seem to be present here: the first, that nothing can be added to an actually infinite set; the second, that if all the negative integers have been assigned to past events then no new events can be added to the latter.

To the first assumption we can answer that to a set of cardinality $\aleph_0$ we can add not only any finite number of elements, as Smith (*op. cit.*) recognizes, but also any finite number of disjoint sets, each of cardinality $\aleph_0$, without altering its cardinality. Assuming the axiom of countable choice, moreover, we have that even a countable union of countable sets remains countable. The second assumption is answered with the example of the famous so-called ‘Hilbert’s hotel’: it is a hotel with $\aleph_0$ rooms; even assuming that they are all already occupied, a new guest can be accommodated by moving the guest from the first room to the second, the guest from the second to the third, and so on, and assigning the first room, thus free, to the newcomer. In our case, every time a ‘new’ event becomes part of the set of past events, we can simply reassign the negative integers, ‘scaling’ them by one, as in the case of Hilbert’s hotel rooms. Basically, it is a question of taking seriously the fact (which Dedekind even took as *defining* the notion of infinity) that a countably infinite set can be placed in one-to-one correspondence with an infinite proper subset of itself.

Dealing with Hilbert's hotel, Sorabji (*op. cit.*) simply emphasizes that this example is in no way a symptom of the absurdity of the notion of actual infinity, as Huby and Craig would like, but only an application of a true assertion about infinite sets, perhaps counter-intuitive at first sight, but certainly justifiable in the light of post-Cantorian mathematics. Sorabji also briefly discusses a very simple formulation of the argument that the past, if infinite, cannot be completed or 'accomplished': infinity, by definition, cannot come to an end, so it cannot be completed or 'accomplished' in any way. To this, he correctly replies that an infinite series may well have an end: in our case we consider the infinite set of past years, which ends in the present, and therefore has an end.

4. (Craig 2000, Whitrow 1966). Tristram Shandy's paradox would demonstrate the impossibility of an infinite past. This is a 'paradox' highlighted by Russell in the *Principles of mathematics* (1903, §340). Tristram Shandy is the well-known character of Laurence Sterne, who writes his autobiography so slowly that it takes him a year to describe the first day of his life. The argument is as follows: at every moment in the past Tristram Shandy was writing his autobiography, regularly taking a year to describe a day; therefore the distance between a past day and the time in which it will be described grows with time; therefore there is no day at a finite distance from any previous day in which all the previous days have already been described; now, the present day is at a finite distance from any past day; conclusion: in the present day not all past days have been described, and the autobiography is incomplete. However, if in relation to the present day there are an infinite number of past days and an infinite number of past days described, then in relation to (and with respect to) any present there are no days not described; but this contradicts the conclusion just obtained.

Smith's discussion of Tristram Shandy's paradox (*op. cit.*; see also Eells 1988) is fundamentally correct, but in our opinion it contains an example that is not entirely relevant. Smith asserts that what in the previous argument does not work is the transition from 'the number of past days described equals the number of past days' to 'there are no past days not described': the past days described constitute a proper subset of the set of past days, yet the two sets have the same cardinality. Smith suggests considering the set of even numbers and the set of natural numbers as a model, but this does not seem relevant here. Instead, we must take the set of positive integers multiples of 365 (ignoring leap years for simplicity) and the set of all positive integers: in fact, if one takes a countably infinite set of days, the first day corresponds to one year, therefore to 365 days; the second to two years, or 2 times 365 days, etc. The two sets can be placed in one-to-one correspondence and therefore have the same cardinality; however one is a proper subset of the other. Smith correctly asserts that at no point in the past, and in no present, does Tristram Shandy complete his autobiography; however, in an infinite time in the direction of the future the autobiography will be completed, since, given $\aleph_0$ days, for each n, the n-th day will be described at the (n times 365)-th day, and the days needed to complete the work will never be missing.

Russell (1903, §340) already made a similar remark, and presented the argument in the following form: (1) Tristram Shandy writes down the events of a day in a year; (2) The series of days and years does not have an end; (3) The events of the n-th day are written down in the n-th year; (4) Each assigned day is the n-th, for a suitable value of n; (5) Therefore each assigned day will have its own description; (6) Therefore no part of the biography will remain to be written; (7) Since there is a one-to-one correlation between the instants of happening and the instants of writing, and since the former constitute a proper part of the latter, the whole and the part have (in this case) the same number of elements.

Let us now see how the Tristram Shandy paradox is dealt with by Sorabji (*op. cit.*). He disputes the claim that the paradox can hold up as an argument against the possibility of an infinite past. Infinite time allows Tristram Shandy to describe an infinity of days, but not all; ‘infinitely many’ does not imply ‘all’; it follows, according to Sorabji, that Russell is wrong when he asserts that no part of the biography will remain unwritten; this holds, in particular, if we assume that Tristram Shandy’s life did not begin; nor will there come a day in which all days have been recorded.

While we do agree with Sorabji’s conclusion, we do not accept his argument. Indeed, he wants to refute the use of the paradox made by those who deny the possibility of an infinite past, yet he denies that the days are sooner or later all recorded. But this is precisely what cannot be denied: as we have seen, Russell is right in asserting that no part of the biography will remain unwritten; there are ‘enough’ years to ‘cover’ every day. It is clear that this does not mean that sooner or later there will be a year in which the work is completed: it is only in the infinite (past, present and future) totality of $\aleph_0$ years that the work will be completed. Even in the hypothesis, which Smith adopts from the start, that Tristram Shandy has lived eternally in the past, it is not clear why the lack of a beginning should change something: taken any day in the past, Tristram Shandy will describe it, sooner or later, taking one year; perhaps he has already described it at the present time, perhaps he has not yet, but this is not relevant with respect to the infinite totality of years that will in any case be needed to complete the autobiography (see, for further discussion, the exchange between Oderberg 2002 and Oppy 2002).

5. (Craig 2000). We can introduce infinite classes by means of the property satisfied by their members, without the need for 'successive synthesis' (in Kantian terms); but the events of the past are essentially given just in succession, so they cannot be actually infinite in number.

Here Smith's objection (*op. cit.*) is very simple: the events of the past (discretely understood, as always in this discussion), if they are infinitely many, are given simultaneously in thought, and this does not prevent the fact that 'in reality' they are given in their normal succession, one by one. Under the same scope falls the 'Kantian' argument (on which see also Puryear 2014 and Morriston 2022) considered by Sorabji (*op. cit.*), echoing the thesis of Kant's first antinomy: the universe must have had a beginning since an infinite series can never be completed by means of 'successive synthesis'. It is clear, Sorabji replies, that this does not exclude an infinity of years, but rather a way of 'reaching it'; no one argues that at a certain point the number of past years becomes, from finite that it was, infinite: it has, so to speak, always been infinite.

6. (Conway 1974, Whitrow 1980). Since it is admitted that in reality events are given in succession, how is it possible that in reality they form an infinite collection? Furthermore, it is not clear how it is possible to conceive someone who writes all negative integers from eternity (in the past) to end with the number -1; counting, by following the descending succession of negative integers, is certainly possible, but it is an inverse process with respect to the succession of events from the past to the present.

Smith objects to this argument in a rather articulate way (*op. cit.*). First of all, a discrete succession of events in time cannot form an infinite set in a finite time, but can do so in an infinite time; so the succession of negative integers has not actually been written, but could be written in an infinite time interval. We add that something stronger is valid: if we are willing to admit (just in the present connection) the continuity of the set of events, there is no reason why an infinite number of events cannot happen in a finite time. For example, a point that moves from the origin of the real line in the positive direction can of course travel the interval from 0 to 1 in a finite time while moving at finite speed. If we identify an event with a position of the point on the line, there are as many events as there are real numbers between 0 and 1, that is to say as many as there are real numbers themselves: these events all do occur, and in a finite time.

Furthermore, the fact that the counting processes with which we are familiar always have a beginning does not imply that one cannot imagine counting processes that do not have this property. If a counting process is simply, as Smith proposes, a synthetic series of counting acts, then nothing prohibits thinking of a one-to-one correspondence between past events and counting acts, such that at present the series of such acts comes to an end. Therefore, if it is true that our own counting when referred to the past goes in the opposite direction with respect to the occurrence of events, we can well conceive a being who in every moment of the past was counting precisely in the order in which past events occurred.

Sorabji addresses the problems related to the counting of past years by first discussing the following objection (*op. cit.*): if the universe did not begin, the counting of years (assuming it has always been such as to assign greater numbers to successive years) should have already reached infinity at any time, however remote, in the past; but how can one conceive of completing a count of this type? Sorabji's answer is that there is a crucial difference between counting and crossing: the need to take a starting number in the case of counting. The absence of a starting point in the sequence of the past years results in the difficulty of imagining in a simple way any count (in a proper sense) of the years in the past: a count in fact always seems to require a first element. One could counter-object (the objection is reported by Sorabji, *op. cit.*, p. 219 and credited to N. Kretzmann) that a count can in fact be imagined, provided that it is 'backwards', i.e. such that one descends from numbers larger in modulus to numbers smaller in modulus, up to zero; and yet, if we are not prepared to say that whoever reaches zero in this counting has concluded to count infinity, we should not even be willing to admit that they have crossed an infinity of years. Sorabji invites, on the other hand, to imagine a beginningless measuring device embedded in a beginningless universe, such as to count how many years remain before a particularly important event, which will correspond to the year zero. It is certainly possible to imagine such a device, and therefore a sort of 'backward' counting. Note how this last counter-objection is similar to Smith's considerations above: in both cases it is admitted that the concept of 'counting' can be extended, without losing its essential properties, to include a sort of 'backward' counting (certainly different from any count we are used to). It is also quite curious that in the course of the same argument, as we have just seen, Sorabji at first asserts that what differentiates counting from crossing is the fact that the former must have a starting point, but then he concedes without problems that one can imagine a 'reversed' form of counting that has no such property. In our opinion, it is the first statement that should be given a provisional value, and then should be discarded: on second thoughts, Sorabji himself recognizes that the idea that counting necessarily presupposes a first element proves too restrictive.

We conclude by observing that a further demonstration of the pervasiveness of arguments based strictly on mathematical infinity in the discussion on the eternity of the world, already in the debate in the thirteenth century, is found in the short treatise by Boethius of Dacia *De aeternitate mundi* (ed. *Opera*, 1976).

Here we find at least three arguments that can be traced back to the patterns of reasoning on infinity that we have identified above. The first two are among the arguments against the eternity of the world (*op. cit.*, pp. 337f., arguments 6 and 10), and are given as follows.

> (1) If something can be added to A, say B, then something can be greater than A; to all the time that preceded the present, one can add more time; therefore there may be something greater than all the time that preceded the present; but nothing can be greater than infinity; therefore all the time that preceded the present is not infinite, and therefore neither are motion nor the world.
>
> (2) If the world were eternal, then it would have passed through an infinite motion and an infinite time, since if the world were eternal the time that preceded the present moment would be infinite; but that the infinite is crossed and taken as something determinate (in the text: *pertransitum et acceptum*) is impossible; therefore the world is not eternal.

The first of these arguments is basically on the line of the third of the arguments refuted above, the one concerning the library of $\aleph_0$ volumes to which, nevertheless, new volumes can always be added without altering the total number. Boethius correctly states that to all the time preceding the present, one can add still more time: the incorrect assumption is that this addition in itself determines an increase in the total number of temporal units, which instead, as we know, remain countably many. We could of course add, after Cantor, that it is not true either that there are no infinities 'greater' (in a precise sense) than countable infinity (which is the only infinity, of course, to which Boethius of Dacia could implicitly refer).

The second argument, on the other hand, is a classic example of application of the principle *impossibile est infinita pertransiri*, which we have already found in Bonaventure, discussing the second class of objections above, and falls under the counter-objections relevant to this principle, among which the distinction between the existence of an infinity of days, each finitely distant from the present, and the existence of a day infinitely distant from the present, remains fundamental.

Another argument in which, albeit not exclusively, considerations of a purely mathematical nature on infinity appear is the second of the series of arguments aimed at demonstrating the reality (not only the possibility, as we have been doing here) of the eternity of the world (Boethius of Dacia, *op. cit.*, p. 341), which is duly answered in the final part of the treatise (*ibid.*, p. 360). In the second part of the argument, in order to show that there was no eternity before the existence of the world, it is asserted that what is preceded by an eternal duration would never come into being; to this Boethius replies that, for example, what is done today, and which was not there before, has an eternal duration behind it (that is, eternity itself, which has always been), and yet it undeniably comes to being. Now, this answer is too 'ostensive', so to speak, not to make one suspect an *ignoratio elenchi*; however, the argumentative technique at work here is none other than the one seen above (second class of objections) in the argument, discussed by Sorabji, similar but not identical to that of the 'non-traversability' of infinity: the crucial point is that there is no starting point in the past from which it would be necessary to cross an infinite number of temporal units to reach the present; to speak of the eternity of the world means precisely to deny the existence of such starting point.

Finally, it is interesting that in Boethius we explicitly find (*op. cit.*, pp. 353f.) the denial of the possibility for the mathematician (whether, according to the subdivision of the *Quadrivium*, arithmetician or geometer or astronomer or musician) to decide, starting from the principles of his science, in one sense or another the hypothesis that the world is eternal. The arguments we have reconstructed seem to some extent to support Boethius, at least on this point, and at least as regards the hypothesis that the world is not eternal: we have seen that the *a priori* arguments of mathematical nature aimed at proving this are not correct.

A fundamental problem, which in our opinion remains and which we have not addressed here as it is not directly relevant, is whether we can separate the aspect of pure conceivability in an abstract sense, in matters concerning time, from considerations of a different type, for example phenomenological (in a general sense), or cosmological, or, more generally, from considerations that philosophically take into account the concepts and results concerning the problem of time that have emerged in the last century in the physical sciences, mainly in the theory of relativity.

## *Acknowledgments*

I wish to thank Cristina D'Ancona, Danilo Manca and two anonymous referees.

Luca Bellotti
Dipartimento di Civiltà e Forme del Sapere, Università di Pisa
luca.bellotti@unipi.it

## *References*

Aristotle, 2005, *De caelo*, Oxford University Press, Oxford (First ed. 1936).

Boethius of Dacia, 1976, *De Aeternitate Mundi*, in *Opera*, vol. VI, part II, Copenhagen.

Bonaventure, 1938, *Commentarius in quatuor libros Sententiarum Petri Lombardi*, Ad Claras Aquas, Quaracchi.

Conway, D. A., 1974, "Possibility and Infinite Time: A Logical Paradox in St. Thomas' Third Way," in *International Philosophical Quarterly* 14: 201-208.

Craig, W. L., 2000, *The Kalām Cosmological Argument*, Eugene, New York.

Eells, E., 1988, "Quentin Smith on Infinity and the Past," in *Philosophy of Science,* 55: 453-455.

Huby, P. M., 1971, "Kant or Cantor? That the Universe, if Real, Must Be Finite in Both Space and Time," in *Philosophy* 46: 121-132.

Morriston, D., 2022, "Infinity, Time and Successive Addition," in *Australasian Journal of Philosophy* 100: 70-85.

Oderberg, D. S., 2002, "The Tristram Shandy Paradox: a Reply to Graham Oppy," in *Philosophia Christi* 4: 351-360.

Oppy, G., 2002, "The Tristram Shandy Paradox: a Reply to David Oderberg," in *Philosophia Christi* 4: 335-350.

Puryear, S., 2014, "Finitism and the Beginning of the Universe", in *Australasian Journal of Philosophy* 92: 619-629.

Russell, B., 1903, *The Principles of Mathematics*, Routledge, London.

Smith, Q., 1987, "Infinity and the Past," in *Philosophy of Science* 54: 63-75.

Sorabji, R., 2006, *Time, Creation and the Continuum*, University of Chicago Press, Chicago.

Thomas Aquinas, 1964, *Summa contra gentiles*; *Summa theologiae*; *Scriptum super libros Sententiarum*; *De aeternitate mundi contra murmurantes*, ed. Leonina, Roma.

Whitrow, G. T., 1966, "Time and the Universe," in J. T. Fraser, ed., *The Voices of Time*, Brazillers, New York 1966, 564-581.

—, 1980, *The Natural Philosophy of Time*, Oxford University Press, Oxford.

# Why authenticity precedes autonomy

Nikos Erinakis

*Abstract*: Most thinkers either identify authenticity with autonomy or take the one to be a core condition for the other. In this paper, I discuss what I believe that authenticity is not. My aim is to distinguish the two notions in regard to their very essence, function and role in our everyday life, while I argue that the conditions of the prominent conceptions of authenticity that relate it to autonomy are unconvincing. I investigate the weaknesses of both the higher-order endorsement models and the externalist historical models by maintaining that none of activity, wholeheartedness, reflection, and rationality is either necessary or sufficient for authenticity. Since manipulation in regard to higher-order desires may take place, one can meet any of these conditions while at the same time being inauthentic. Given this, it has been argued that although these conditions are perhaps insufficient for authenticity, they are still necessary. However, I argue that they are also unnecessary — that is, authenticity comes before activity, wholeheartedness, reflection and rationality, and not vice versa.

*Keywords:* authenticity, autonomy, identification, reflection, reasons, attitudes.

## 1. *Introduction*

In this paper, I elaborate on the weaknesses of the higher-order endorsement models and the externalist historical models of authenticity by concentrating on the reasons why I believe activity, wholeheartedness, rational and mere reflection, and both reflective and unreflective reasons are inadequate to operate as either necessary or sufficient conditions for authenticity. Since manipulation in regard to higher-order desires may take place, one can meet any of these conditions while at the same time being inauthentic with respect to an attitude. Given this, it has been argued that those conditions may not be sufficient for authenticity, but that they still are necessary. In contrast to the majority of the prominent autonomy and authenticity thinkers, I argue that they are not necessary either. This should create a basis upon which I maintain that when distinguishing which attitudes and creations are authentic, we should not only trust rationality and reflective thinking, but also other capacities of ours, like imagination, intuition, inclinations and drives, as long as they are creative.

ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i1.376

Submitted: 21 August 2021
Accepted: 17 October 2022

I claim that taking a step back and rationally reflecting on what is one's own cannot ensure that what one settles on is truly one's own authentic creation. The processes of rationality and all kinds of reasoning and reflection must also be authentic if they are to be adequate tools for distinguishing what is authentic from what is not. They need to have been formulated and developed creatively — not solely rationally — in order to be one's own and not simply externally generated. Given this, I argue that authenticity should come first in order to ensure a development of an authentic process of reflection and reasoning and not as a result of them.

Many thinkers understand authenticity in terms of the simple idea that what is authentic is whatever is one's own, with the question of what it is for something to be one's own either neglected or misconstrued as a question about autonomy. I aim at showing that a broader understanding of authenticity is required and that autonomy and authenticity are not only not coextensive but also potentially contradicting and conflicting. What is important regarding the quest for authenticity is to determine in which ways one's creations are one's own. Hence, there are two central questions that need to be answered: *What* it means for a creation to be one's own, and *how* it comes to be one's own.

In regard to the dominant contemporary autonomy and authenticity conceptions, there are two ways in which authenticity conditions are generally introduced. The first is that we seek conditions based on which we can distinguish authentic from inauthentic features of the self. The second is that we seek conditions that present the tools based on which the agent is able to formulate and develop authentic features. While studying various scholars that refer to higher-order endorsement and historical models, we may notice that Harry Frankfurt's (1988) conception of autonomy is equated with authenticity, Gerald Dworkin's (1988) with authenticity and independence, John Christman's (1991) with authenticity and competence and Alfred Mele's (1993) with self-control and authenticity. More precisely, it seems to me that the prominent theories of autonomy can be divided into two categories. In the one, autonomy is equated with authenticity, i.e. they conceive authenticity as both necessary and sufficient for autonomy, and in the other autonomy consists of authenticity plus some other element, i.e. they conceive authenticity as necessary but insufficient for autonomy. Accounts of the former kind have been developed by Frankfurt and Christman, while accounts of the latter kind have been developed by Dworkin and Mele. Frankfurt's and Dworkin's models are often considered as almost the same because of their hierarchical nature. However, in my opinion, Frankfurt's and Dworkin's conceptions of autonomy, despite their similarities, are importantly distinct, since the former can be equated with authenticity while the latter requires independence too, thus, they should not be conflated into one model.

Furthermore, even though Christman seems to distinguish authenticity from competence, he does not, as his competency condition is absorbed into the one of authenticity, with the result that he equates autonomy with authenticity too.

Thus, most thinkers who develop conceptions of autonomy seem to take for granted that authenticity is, if not autonomy itself, at least a core condition for autonomy, or in other words, that it is the first and basic step for autonomy to obtain. I believe that this is the source of several critical misunderstandings, beginning with the negligence of the importance of authenticity as a fundamentally separate concept. Only if authenticity is understood in its own terms can the various different dimensions of it be revealed.

## 2. *Activity*

Many theorists argue that authenticity and activity are directly connected, and more precisely that in order for a person to be authentic with respect to a certain desire one necessarily needs to be active towards it. The connection between activity and authenticity in the sense of ownership of attitudes is evident both in Frankfurt (1988, 1999, 2002a, 2002b) and Richard Moran (2002), who claim that what is required for a desire to be authentic is for the agent to be active with respect to it.

Frankfurt is rather clear about his view of what activity is. In order for one to be active with respect to a desire, one must identify with that desire. In other words, we are active towards only those passions that are genuinely internal to us, i.e. our own. For him, ownership of higher-order attitudes, identification with those attitudes and activity with respect to them all amount to the same thing. In his own words:

> Now a person is active with respect to his own desires when he identifies himself with them, and he is active with respect to what he does when what he does is the outcome of his identification of himself with the desire that moves him in doing it. Without such identification the person is a passive bystander to his desires and to what he does. (Frankfurt, 1988: 54)

Furthermore, he also writes: ‘The attempt to explicate being active in terms of endorsement is inevitably circular, accordingly, since asserting that a person endorses something necessarily presupposes that he is active.’ (Frankfurt, 2002b: 220) This suggests that we are active towards those desires that are truly our own, ‘which express our nature most fully and most authentically,’ (Frankfurt, 2002b: 224) or in other words that are in such a degree our own that ‘do not accommodate themselves to our thinking. Rather, our thinking accommodates itself to them.’ (Frankfurt, 2002b: 224) However, it also suggests that not only are identification and ownership a presupposition for activity, but that activity is also a presupposition of identification and ownership.  Identifying with a desire means being active towards it and being active towards a desire is necessary and sufficient for being able to identify with it. In this sense, authenticity cannot exist without activity and vice versa. Following from this, in his theory, authenticity is equated with identification, which is equated with ownership, and identification presupposes activity, while activity presupposes identification too. Thus, Frankfurt equates authenticity with activity or, at least, activity, in his view, can be considered a both necessary and sufficient condition for authenticity.

In *Contours of Agency*, Frankfurt's 'Reply' to Moran includes a number of interesting points. He writes: 'In his [Moran's] view identifying with something like a thought or a desire consists in "assuming some kind of active stance toward it".' (Frankfurt, 2002b: 218) For Moran, Frankfurt's grouping of the internal/external and active/passive distinctions makes sense for sensations and bodily movements but not for attitudes and mental states. In order to support the distinction between attitudes and sensations in terms of a person's responsibility towards them, Moran refers to the connection of it with activity, which for him presupposes identification. He attempts the same equation between the agent's ownership of beliefs and attitudes and her activity towards them. In other words, one is active with respect to an attitude if this attitude is one's own and in this sense one has endorsed and identified with it. Hence, in Moran's view too, activity is equated with authenticity.

Activity, however, cannot operate as a sufficient condition for authenticity, since a person, even when she is active with respect to an attitude, could have been manipulated into being active or into wanting to be active towards it.[1] Even if the person identifies with a desire based on higher-order reflection, her second-order desires may be a product of external manipulation. Consider the case of a person who is hypnotized by agents of the secret service of a country in order to murder the prime minister and to confess afterwards that he had personal or ideological reasons to do so. This person will certainly believe that his self is both active towards his second order desires and, since he identifies with those, active towards his first order desires too. In reality though, he has been manipulated into believing this and committing a crime, which he did not authentically desire to commit in the first place. Thus, one may be active towards a desire, while inauthentic with respect to it. Moreover, this same argument may just as easily be made against all of the other internalist conditions with which I deal in this paper, i.e. wholeheartedness, all kinds of reflection, and having any kind of subjective reasons for desiring or doing something.

[1] This is discussed in depth in Mele's *Autonomous Agents* (2005) and Christman's 'Autonomy and Personal History' (1991) and *The Politics of Persons* (2009).

This said, I shall argue that activity, besides not being a sufficient condition for authenticity, is not a necessary condition for it either. The distinction between authenticity and activity should be clear. If a person is active that does not mean in any sense that she is necessarily authentic, i.e. it is possible for a person to be authentic but passive. It is often thought that when a person experiences a strong emotion that overwhelms her, she is passive towards it, since she can do nothing to control it. Even so, she might be completely authentic with respect to it since it may arise from her internally generated attitudes.

Consider the following example:

> *Unfaithfulness*. A person meets someone and they both experience an extreme sexual connection between them. They authentically desire to sleep with each other. However, both of them are in strong relationships and they know that besides the sexual connection they share nothing else, while each of them has countless things in common with their current partner. Despite that, they go on and spend the night together. A common friend tells on them and they both end up divorced from their partners and unable to see each other again because of guilt or because they do not fit at all in everyday life.

The desire that these two persons experienced was so strong that they both felt passive with respect to it, and they could do nothing to control or change it. If they had been able to reflect properly (either rationally or not) on this desire they would have probably avoided having sex, and they would probably be better off afterwards. However, this does not change the fact that what both authentically desired at that moment was to sleep with each other. They may be considered passive with respect to this desire that surpasses any form of their rational resistance and gets control of them, but that does not mean that they are not also authentic with respect to it. In other words, this might have just been a strongly authentic desire that rendered them passive.

However, in many cases the question of passivity and activity might be more complex than it looks. In this sense, it would be better to speak of cases where the agent *experiences* something as active or passive and not necessarily *is* active or passive, since in reality one may be active in both cases. Attitudes, which are generally considered passive, may be actually active in cases that are direct responses of the person towards the stimuli that caused them. For instance, even inertia may be an active response in many instances. Nevertheless, when one is either active or passive, or even when one experiences an attitude as being passive towards it, while in reality one may be, in a different sense, active, one can be authentic with respect to it.

Authenticity and activity should come apart as notions. Authenticity does not require activity in order to obtain, i.e. activity is neither necessary nor sufficient for authenticity.

## 3. *Wholeheartedness*

In Frankfurt's view, identification with a desire requires a certain sort of stability or equilibrium with respect to one's attitude towards it; this is the role of wholeheartedness. For him, wholeheartedness means having a higher-order desire without reservation or other conflicting higher-order desires. Authenticity with respect to, or identification with, a desire is a matter of being reflectively satisfied with it, and this in turn is a matter of being wholehearted with respect to it. He writes: 'Now I will try to develop a more fully articulated understanding of what it is to be wholehearted, by construing it as tantamount to the enjoyment of a kind of self-satisfaction.' (Frankfurt, 1999: 102) and 'Identification is constituted neatly by an endorsing higher-order desire with which the person is satisfied.' (Frankfurt, 1999: 105) Thus, for Frankfurt wholeheartedness is both a necessary and a sufficient condition for self-ownership of the attitudes, i.e. for authenticity. However, I shall argue that it is neither sufficient nor necessary.

Frankfurt conceives ambivalence as a volitional division in the self that keeps an agent from settling upon or from tolerating any coherent affective or motivational identity. A person is ambivalent when she is moved by preferences regarding her desires that are incompatible. For Frankfurt, ambivalence is constituted by conflicting volitional movements which meet two conditions: Firstly, they are by their nature opposed and secondly, they are both wholly internal to a person's will rather than alien to him, i.e. she is not passive with respect to them. Conflicts involving first-order psychic elements alone do not pertain to the will; conflicts that pertain to the will arise out of a person's higher-order reflective attitudes. But even conflicts that do implicate a person's will are nonetheless distinct from ambivalence if some of the psychic forces they involve are exogenous — that is, if the person is not identified with them and they are, in that sense, external to her will. This leads Frankfurt to claim that if ambivalence is to be understood as an illness of the will, then for the will to be healthy it should be unified and wholehearted (Frankfurt, 1999: 100-1, 106-7).

In my view, wholeheartedness seems like an ideal that can be reached only in specific and rare cases. I can imagine how I could wholeheartedly decide with whom I generally want to spend the following years of my life, but in issues met in everyday life the state of wholeheartedness is not so clear. Most decisions we make are outcomes of conflict, but we rarely come out of this conflict with the feeling of wholeheartedness that Frankfurt describes. More than often we make a decision with some doubts or ambivalent thoughts about it. A part of ours might still want to decide to follow the other option. That is not to say, of course, that authentic decisions and actions cannot exist, but rather that wholeheartedness need not be a necessary condition for considering them such. I may authentically desire to cheat on my partner but that does not mean that I do it wholeheartedly, or I may have an authentic desire for self-harm but that does not mean that I harm myself wholeheartedly. A part of me might still want to do otherwise, even though doing otherwise might not be authentic. In this sense, wholeheartedness cannot operate as a sufficient condition for authenticity. Besides, the example of manipulation, mentioned in the previous section, stands here too. One may be manipulated in desiring wholeheartedly to act in a certain way. What remains, therefore, is to prove that it cannot operate as a necessary condition either.

Frankfurt explores the question of whether it is possible for a person to be satisfied with ambivalence. He takes for granted that we necessarily desire in a wholehearted way to be wholehearted: 'But no one can desire to be ambivalent for its own sake. It is a necessary truth about us, that we wholeheartedly desire to be wholehearted.' (Frankfurt, 1999: 106) However, I cannot see how this can be taken to be an axiom. There are people who prefer to be in a state of ambivalence, people who experience panic when they are with both legs on the one side of things. They may feel that by identifying themselves with only one desire out of two they become one-sided and they lose the complexity of their multisided nature. They may feel trapped by wholeheartedness, whereas their authentic state may be ambivalence and levitation between two or more equally authentic desires.

One may remain completely indecisive between two partners that one may have at a certain period of time. One may feel that choosing to be with only one of them would be inauthentic, since suppressing one's desire for the other partner would render one inauthentic with respect to this decision. In this case one may prefer the ambivalent state of being between both partners and not with each one exclusively. Thus, there may exist cases in which one may be authentic only when one levitates constantly between two different desires, whether these are irrelevant and unrelated to each other or they are conflicting.

At another point Frankfurt claims that the ambivalence of a person obstructs the way of a possible existence of a certain truth about this person; there exists neither truth nor lie about this person: 'This is why ambivalence, like self-deception, is an enemy of truth…[H]is ambivalence stands in the way of there being a certain truth about him at all. He is inclined in one direction, and he is inclined in a contrary direction as well; and his attitude toward these inclinations is unsettled. Thus, it is true of him neither that he prefers one of his alternatives, nor that he prefers the other, nor that he likes them equally.' (Frankfurt, 1999: 100) Could we, however, accept such an argument in this case? In my opinion, we cannot. The state of ambivalence may be part of the agent's authentic nature. Referring back to the discussion of the previous section, even if activity is lost because of the state of ambivalence, we may say that the agent is authentically passive, as long as the agent's authenticity is manifested more truly in a state of ambiguity.

Let us consider Agamemnon's case:

*Agamemnon's love*. Agamemnon needs to choose between sacrificing his daughter Iphigenia so that the Greek army can set out for Troy and win the war and keeping his daughter alive but losing the war. His parental love comes in clear contradiction with his desire to win.

Which of the two is Agamemnon's authentic desire? Perhaps both his love for his daughter and his desire to win the war are authentic desires but at the same time conflicting. However, he has to choose to act on only one of the two. If both desires are equally authentic, then are both potential decisions to be considered equally authentic too? For now, we may concentrate on the fact that whichever desire Agamemnon chooses to follow he is not going to be wholehearted with respect to it. However, that does not mean that he will not be authentic with respect to it either. Especially in the case that both conflicting desires are equally authentic, then whichever desire he decides to follow, his action will be just as authentic as the other. In this sense, wholeheartedness is not necessary for authenticity.

This said, two desires may be equally authentic. If these desires conflict, one may experience a state of pure ambivalence. This has both an important advantage and an important disadvantage. The advantage is that whichever desire one ends up following, one will be authentic with respect to it. The disadvantage is that one will have to sacrifice a part of oneself in following one of the desires and suppressing the other. This is evident in the case of Agamemnon. Each one of the available choices that he has leads him to an authentic path; however, he cannot move forward without making an unbearable sacrifice, and this is exactly what creates the essence of his tragedy, what makes him a tragic hero.

Nevertheless, Frankfurt might raise a certain objection to this. He might argue that one could be wholehearted with respect to both conflicting desires, i.e. be equally wholehearted in regard to each of them. What if Agamemnon was wholehearted with respect to both of his conflicting desires? But this is not a coherent possibility. Firstly, in order to be wholehearted, one's heart needs to be whole in regard to a certain attitude. Secondly, even if we do not take the word literally and we only refer to the abstract, metaphorical concept of wholeheartedness, I cannot see how one could desire absolutely one thing and at the same time desire absolutely another conflicting thing too. When conflicts exist; division takes place. This does not imply that because one cannot desire something in an absolute way, one cannot be authentic. As life goes on and one's inner nature expands, one may experience potentially more and more conflicts.

Regardless of this, authenticity may still obtain, even in respect to conflicting attitudes. Which one, however, is more authentic depends on its degree and not on whether it is endorsed absolutely by a person who identifies with it in an absolute wholehearted way. The self, even though in a certain sense it may seem unified macroscopically, experiences certain conflicts which can be compatible mainly with a fragmented conception of it. Authenticity, nonetheless, is not necessarily obstructed when in ambivalence or conflict. Besides, at times, a person's inner nature may be genuinely authentic when in ambivalence or conflict.

Based on the above, I argue that wholeheartedness is neither a necessary nor a sufficient condition for authenticity. A person can be authentic with respect to an attitude without, in any sense, being wholehearted towards it.

## 4. *Reflection*

As mentioned, the significant majority of accounts of autonomy and authenticity take rational reflection to be a necessary condition, except for Frankfurt's account in which reflection need not be rational. In the first subsection I deal with the condition of rational reflection[2], while in the second subsection I deal with Frankfurt's 'mere' reflection.

### 4.1. Rational reflection

Both in Alfred Mele's and John Christman's conceptions, rational reflection (either actual or hypothetical) is necessary for authenticity. Mele argues that in order for one to be authentic one's beliefs should be conducive to one's informed deliberation and that one should be a reliable deliberator (Mele, 1995: 187), while Christman devotes almost half of his conditions to the capacity of the agent to critically reflect (Christman, 2009: 155). The reason why most theorists tend to provide a condition of rational reflection for authenticity is because they believe that through this they avoid the danger of manipulation or other-directedness, which, as already mentioned, is evident in higher-order reflection theories. This, however, leads to a miscomprehension between the notions of activity, rational reflection and authenticity. In these thinkers' views, in order for one to be authentic one needs to be active, and in order for one to be active one necessarily needs to be able to rationally reflect. That is why they consider the capacity for rational reflection as at least a necessary condition for authenticity.

[2] I will be using the terms critical reflection and rational reflection interchangeably while referring to the same form of reflection based on the faculty of reasoning.

Turning now to Alfred Mele, in the first part of his book *Autonomous Agents* (1995) he discusses the notions of akrasia and self-control, arguing that self-control is the basis for autonomy.[3] He clarifies that self-control by itself cannot ensure autonomy, since the agent may be self-controlled, while, however, controlling herself in accordance with values and beliefs that are products of external manipulation. In the second part of his book he proposes the addition that must be made to self-control in order for autonomy to exist: authenticity. For Mele, in order for a pro-attitude to be possessed autonomously, it should be also possessed authentically.

Thus, it is clear that for Mele autonomy consists of self-control and authenticity. Even an ideally self-controlled person cannot be autonomous if the condition for authenticity is not met. For him, as with Dworkin, the capacity of one to reflect critically upon one's preferences and desires, and the ability either to identify with these or to change them in light of higher-order preferences and values, is necessary for autonomy. However, in order for autonomy to exist something more is required and this is where the historical aspect appears. Since for Mele autonomy is not simply an internalist matter, like it is for Frankfurt and Dworkin, the history of the individual and the formation of her characteristics play a significant role. This makes his conception an externalist one. As proven especially by his 2* condition (Mele, 1995: 171-2), he is interested in the history of the formation of each characteristic in order to distinguish whether it is a history which is authenticity-enabling or authenticity-blocking. In this sense, his conception of authenticity is clearly history-sensitive.

However, a number of thinkers acknowledge that rational reflection cannot be sufficient by itself as a sole condition for authenticity. Mele, while criticizing higher-order reflection theories, summarises the crucial weakness of rational reflection:

> Possession of a capacity for critical reflection is a plausible requirement for autonomy. But the problem of value engineering…suggests that even a robust and effectively exercised capacity of this kind is not sufficient for psychological autonomy…If the perspective from which an agent critically reflects upon his first order preferences and desires at a time is dominated by values produced by brainwashing and dominated in such a way as to dictate the results of his critical reflection it is difficult to view the reflection as autonomously conducted and the results as autonomously produced. (Mele, 1995: 147)

[3] The condition of self-control, which has been common to thinkers of freedom and autonomy, has its origins in Descartes's model of rational control and more importantly in Locke's rebuilding and redefinition of Descartes's theory of rational control of the self. Locke develops an idea of a process of self-remaking from which it is concluded that a person instead of blindly following the *telos* of nature may formulate one's own self.

Mele believes that in order to determine whether values and preferences are authentic we need to look to their history, and that it is therefore possible to solve these problems by supplementing a higher-order reflection theory with a historical condition. The problem, nevertheless, exists not only in the history of the formation of values and preferences, but also in the history of the formation of the processes of rationality and reflection themselves. Obviously there can be authentic preferences formulated and located through rationality and reflection, but it is inadequate to consider them the sole conditions for authenticity. In the same way as values, beliefs and desires may be manipulatively imposed on the agent, certain processes of reasoning or reflection may be manipulatively imposed on one too. Besides, this commonly occurs in societies during the upbringing in the early stages of persons' lives through various forms of social conditioning.

In other words, it is not only the material on which the agent reflects or reasons, i.e. values, beliefs etc., that may be manipulatively imposed, but also the process of rational reflection itself, the way in which the agent interprets, develops and uses those values and beliefs, that may be manipulatively imposed too. Having good reasons for desiring something does not mean that one authentically desires it, but more importantly, even if it did mean that, what the agent considers good or bad reasons for having a desire, i.e. one's way of reasoning, should be formulated authentically to begin with. Thinkers who develop historical conditions for authenticity, as Mele and Christman do, tend to neglect this latter aspect.

Furthermore, while concentrating on the relationship between authenticity and autonomy, Mele discusses the case of someone who voluntarily decides to be manipulated in order to promote her autonomy (e.g. she allows herself to be hypnotised in order to quit smoking). This is an interesting case which incorporates the crucial reason why the distinction between authenticity and autonomy is important. If one decided that a particular desire was inauthentic, then it would make sense to choose autonomously to reject it. But what if one's desire was authentic and one autonomously decided to reject it?

Based on Cal's case, an ex-smoker who is happy with her decision to quit smoking but sometimes still experiences a desire to smoke, Mele claims that even if the desires of an agent are not manifestations of her autonomy, the agent may be autonomous in continuing to have them. It would be interesting to consider Mele's argument in terms of authenticity in order to possibly stretch out a crucial difference between autonomy and authenticity. Think of a person who quit smoking last year but now desires to smoke a cigarette. Even though she has autonomously quit smoking for a year and she continues to rationally believe that she should not smoke, she may, while meeting Mele's requirements for authenticity, authentically desire to have a smoke. If she lights one up, she is authentically non-autonomous. In addition, based on Dworkin's theory, consider a person who experiences a first order desire to quit his job in order to travel with an old bike all the way through Pan-American Highway in Latin America. He experiences, however, a second-order desire that dictates him to keep his job in order to be able to retain his costly way of living. Although, he concludes after rational reflection that he should follow his second-order desire, he does not, and he embarks for Latin America. This person also is authentically non-autonomous.

Since most conceptions require the capacity for rational reflection in order for authenticity to obtain, it can be argued, based on their views, that emotions can compromise authenticity. However, there may be cases in which reasoning may compromise equally, or even more, the authenticity of emotions. For instance, in the case of Agamemnon, if, for the sake of this argument, we consider parental love a deeper emotion that originates before it is endorsed through reflective reasoning and the desire to win the war an outcome of rational reflection based on good reasons, we understand that, in some cases, rational thinking may compromise and constrain authentic desires through putting limits on the manifestations of our authentic attitudes. Given this, we could assume that sacrificing his daughter is a desire rational for him and the others, but completely inauthentic for him. In this sense we notice that through rational reflection authenticity is not guaranteed, since after serious and even independent rational reflection, one may decide to neglect one's authentic desire in order to follow an inauthentic desire, simply because one's reasoning and rational reflection dictate one to do so. What I am suggesting is that in the same way as autonomy theorists have argued that rationality should be the sole tool for determining the authentic attitudes of a person, the person's creative processes may be in turn the tool for determining her authentic processes of reasoning and reflection. Besides, as I shall argue elsewhere, it is my view that creative attitudes are the ones that create the reasons on which authentic reasoning should be based and not vice versa.

As mentioned, many thinkers claim that for one to be authentic with respect to a desire, one must critically reflect on it. This presupposes that an agent must have good reasons in order to identify or endorse a desire, and that one is capable of discovering or developing these good reasons through rational reflection. However, Frankfurt disagrees with this. His notion of reflection, which I discuss in more detail in the next subsection, does not involve rationality. He writes:

> Identification and wholeheartedness are volitional states that necessarily create reasons but that do not otherwise depend upon them. We can identify with various psychic elements, and we can be wholehearted in various thoughts and attitudes, without having any reasons for doing so. On the other hand, it is in virtue of these states of our wills that certain things count for us as reasons. (Frankfurt, 2002b: 218)

Take, for example, the passivity, or potential inauthenticity, of an akratic or mentally ill person. Moran (2002: 192-3) claims that what characterizes her is the absence of rational endorsement, which for Frankfurt is different from mere approval. For Moran an unwilling narcotics addict is passive towards her desire for the drug because she does not endorse that desire rationally. He claims that since a person's intentional attitudes are supported by reasons, one identifies more with them than with one's sensations, as the former reflect more accurately who we are than the latter. For Frankfurt (2002b: 219), on the other hand, whether the endorsement is rational or not does not make a difference in rendering the addict active towards the desire.

Taking Frankfurt's argument one step further, a person may identify with certain desires without having any good reasons, and be completely foolish but still authentic with respect to them. In other words, these desires may be completely irrational but still authentic. On the other hand, a command or an other-directed desire that you take to be rational need not be authentic; this only means that you have reflected on it and it seems to make sense to you. Perhaps you may rationally agree with it and you may be able to understand that it might be authentic to you, but this alone is not adequate. Considering something rational while reflecting on it and deciding to incorporate it, even through identification, cannot adequately prove that you are authentic with respect to it.

In addition, Frankfurt talks about desires that are so deeply rooted in us that we cannot avoid or reject them. I do not agree with Frankfurt that such desires are necessarily authentic, since as Mele and others have pointed out, those desires might be a product of manipulation. I do agree with Frankfurt though that truly authentic desires determine our thinking whereas our thinking and/or reasoning in many cases is unable to determine them, i.e. it is authenticity that creates reasons and not vice versa. These desires are not simply as Frankfurt claims 'stronger than we are' (Frankfurt, 2002b: 224), they might be exactly what we are. They might be stronger than our reasoning and rational reflection, but this is perhaps why they constitute and manifest what we are more faithfully. They reach aspects of us that lie beyond reasons. The fact that one locates certain reasons for a desire is neither necessary nor sufficient for it being actually authentic; on the contrary, the fact that one experiences a desire as authentic is a strong reason by itself to accept it as such and this can itself generate reasons.

In order to shed more light on this argument, we could refer to one of Frankfurt's examples, in which a mother believes that what would be rationally best would be to give up her child for adoption, but she finds that she cannot go through with it (Frankfurt, 2002a: 149-151, 160-1). For Gary Watson this is a kind of defeat, since he claims that: '[T]he second outcome [i.e. to give her child away] leaves her with a kind of volitional or authorial integrity that is not achieved in the other case' (Watson, 2002: 150), while for Frankfurt it may be a liberation (in any case, more information about the mother is required in order to reach a sound conclusion). It seems to me that even if the mother rationally decided to give her child away, this would mean that she would have decided to act inauthentically, i.e. to overcome her authentic desire and act without its influence on her; in other words, to impose on herself a rational necessity in order to overcome her authentic one. The mother, after rationally reflecting, might have more than good reasons to give her child away, but that does not mean that it would be authentic of her to do so. Given this, the mother might act completely irrationally, both in the sense of acting against her best judgment based on good reasons and, as I shall argue, of acting against other unreflective reasons that she may have, and still be authentic. We do not always agree with or find rational our authentic desires, and we do not always identify with them, but this does not mean that they are not authentic.

In this sense, rationality and reasoning may be inadequate to help us in distinguishing our authentic desires from our inauthentic ones. The concept of the rational agent cannot represent the whole nature of a person and it seems wrong to base our conception of authenticity on an agential idea that excludes other fundamental aspects of our inner nature. The equation of human nature with rationality is a distorted, one-sided ideal that constricts and confines both the actuality and the potentiality of human nature. For reasons already mentioned, like manipulation through implantation of second order desires, I consider self-reflection inadequate too. Thus, the solution lies in understanding how these desires can be authentic without necessarily invoking our ability to critically reflect or our taking ourselves to have good reasons for having them.

Rational reflection is neither necessary nor sufficient for authenticity. One can be absolutely authentic without the use of rational reflection or without even the hypothetical capacity for it. However, that does not mean that I agree with Frankfurt's conception, since, as I argue in the next subsection, reflection of any kind is not necessary for authenticity either.

### 4.2. Mere reflection

Frankfurt takes reflection to be a condition for authenticity, but he does not require this reflection to be rational. Having good reasons for identifying with an attitude through reflection may not be involved at all in his view. However, his notion of reflection experiences an unavoidable flaw. The common counter-argument to Frankfurt's conception of higher-order reflection is the historical objection to which I referred in the second section. Mele (2005) and Christman (2009) have developed their objection by proving the possibility of manipulation of one's higher-order desires. One cannot be considered authentic based solely on one's processes of reflection and endorsement. This alone is enough to prove that reflection, even without the rational/critical aspect, cannot operate as a sufficient condition for authenticity.

What is more, Frankfurt argumentation is not adequately convincing in considering that we can conclude whether a desire is internal or external only through the processes of reflection and identification. I argue that one can be absolutely authentic without the use of any kind of reflection. Consider the following example:

> *In search of the authentic foot.* A dancer or actress who self-choreographs her kinesiology for an art performance is looking for which of her two legs she should use as the centre of expression of her movements. The obvious answer that she should use her good foot, admittedly does not involve the artistically meaningful dimension that she seeks. Thus, a colleague of hers, as she tries different versions — which she films and does not want to interrupt them — approaches her and, without her knowing, suddenly pushes her. Instinctively she puts one foot in front of her, not the good one, in order to avoid the fall, but at the same time not to spoil the attempt of a choreographic ensemble of that version. In this way, she realizes that the answer to her dilemma has been revealed.

She could not have figured out which leg she would like to use as the canter of her kinesiology only through rational and/or mere reflection. The reason her colleague pushed her without warning was because, in order to find it, she had to trust her instinct without further thought. Of course, the reflection was useful later, since based on this she could decide which foot to use in order to better express the artistic meaning of her performance. But in order to detect it, she first needed the help of her instinctive reaction. Obviously, finding one's authentic foot is a physical characteristic of the body and thus significantly different from attitudes. However, I use this example as an analogy in order to argue that the same also stands for attitudes and decisions. Consider another example:

> *Ionesco's Bérenger.* Bérenger is the central character in Ionesco's *Rhinoceros*. In the play the inhabitants of a small, provincial French town turn into rhinoceroses; ultimately the only human who does not succumb to this mass metamorphosis is the central character, Bérenger. The play is often read as a metaphor and criticism of the sudden upsurge of Fascism and Nazism.

Bérenger, before being able to rationalize why he feels the need to go against the 'Rhinoceritidis', experiences that need as an intuitive reaction. He says: 'Now I 'll never become a rhinoceros, never, never! I 've gone past changing. I want to, I really do, but I can't, I just can't…People who try to hang on to their individuality always come to a bad end! Oh well, too bad! I 'll take on the whole of them! I 'll put up a fight against the lot of them the whole lot of them! I'm the last man left, and I'm staying that way until the end. I'm not capitulating!' (Ionesco, 1960: 107) For the time being, a deeper intuitive reaction is revealing to him his authentic desire and guides him in remaining authentic. Bérenger experiences, in the form of a feeling instead of a reflective conclusion, the need to resist. He does not raise any rational or intellectual arguments against the '*Rhinoceritidis*', he simply experiences a strong need for resistance against it and a robust feeling that he would be alienated were he to succumb to it.

According to this, one could argue that Bérenger could be considered a wanton in Frankfurt's sense. Frankfurt defines a wanton as an agent with no second-order volitions who does not care what she wills (Frankfurt, 1988: 16-7). An individual who is a wanton may have rational faculties of a higher order, but she is not concerned with the desirability of her desires, or with what her will ought to be. Frankfurt claims that a wanton's identity is her first-order desires. However, why can there not be cases in which those first-order desires are authentic? Since a wanton's identity is her first-order desires, then if those are authentic, she is authentic too. Besides, a first-order desire might be much more authentic than one's reflective desire to be a person that would desire and will something different. Furthermore, in Frankfurt's view, a wanton has no stake in the conflict between two desires and, as the one desire prevails and the other is left unsatisfied, the wanton is neither a winner nor a loser. But, what Frankfurt has not taken into account is that if the wanton is authentic in the state of ambivalence, i.e. authentically desires to experience ambivalence, then she can be satisfied by remaining in such a state.

Imagine an authentic wanton; for instance, a child dancing freely. Bérenger does resist the transformation and he clearly chooses between becoming a rhinoceros or not. He may not have or acknowledge good reasons for doing so, like the child who dances freely, since his feeling of resistance to this transformation operates as a reason itself. Thus, Bérenger, despite of whether he is a wanton or not, even if he had been 'trapped' in a state of ambivalence, he would have had equal chances to be authentic.

That form of resistance is an outcome of authenticity coming from an intuitive feeling as opposed to a more rational way of reflective thinking (which from time to time and from society to society may be conceived differently). Even if at a first glance that non-rational 'inner voice' might seem completely irrational, it still remains authentic. That inner voice may be understood as a strong, almost robust inclination that has been formed not necessarily by rational reflection but by emotions or an intuitive feeling that the agent has not rationalized yet. This may seem to be in line with Frankfurt's point. However, as I shall argue, this by itself is not adequate for authenticity. Bérenger's example constitutes a case in which a person may act in the eyes of the others, or even in the eyes of himself, completely unreflectively but completely authentically too. His desire to remain as he is and not to succumb is both unreflective and authentic.

Following from the above, one might be authentic with respect to a desire not only despite a lack of rational endorsement, but also despite a lack of any kind of endorsement or reflection. For example, recall the *Unfaithfulness* example mentioned in Section 2, where two people experience a strong connection and authentically desire to sleep with each other. Whether they do so or not, this was an authentic desire, whereas the one produced by reflection might be inauthentic and other-directed. I do not intend to suggest that first-order desires are necessarily more authentic than second-order desires. My aim is simply to claim that there are equal possibilities of first-order and second-order desires being authentic or inauthentic. In this sense, reflection in general is not only an insufficient condition for authenticity, but also an unnecessary one.

## 5. *Unreflective reasons*

I have argued that reflection is neither a sufficient nor a necessary condition of authenticity. However, another line of argumentation has also been suggested. Nomy Arpaly (2003) argues that one can base one's attitudes and decisions on good reasons that one has not reflected on. A possible extension of Arpaly's view might hold that one can be authentic with respect to an attitude only if one has good reasons for it — even if one has not reflected on these reasons. I shall argue that good reasons of any kind, even unreflective, are neither necessary nor sufficient for authenticity.

More precisely, Arpaly's account implies that in cases that one may act without an articulated reason in mind, one should not come to the conclusion that one is acting irrationally but rather consider the possibility that one is acting on good reasons which one simply has not yet articulated. In the same sense, when one tends to act against one's 'considered judgment' — the judgment one makes on the basis of the reasons one can articulate — one should not automatically conclude that acting on this inclination would be irrational, but rather one should consider also the possibility that one is acting on good reasons which one may not have articulated yet. Let us consider Huckleberry Finn's (Twain, 2008) case:

*Huckleberry Finn*. Finn saves his friend Jim, an escaped slave, by not turning him in to the authorities, even though this was illegal. Arpaly concludes that Finn is praiseworthy because he is responsive to the right reasons. Even though he cannot correctly represent those reasons as moral reasons, and he himself does not understand the nature of his actions, Arpaly suggests that he is right with respect to them.

Finn, however, may have not acted on the basis of a reason. Finn may have acted in the way he did out of an attitude, which is not necessarily based on other kinds of beliefs but mostly on intuitive feelings like empathy and sympathy for a fellow human being and, in this case, a friend. However, one could argue that those feelings of empathy and sympathy are responsive to moral reasons to begin with. Given that, an agent that acts based on other beliefs that may not be rational in any sense, reflective or unreflective, may nevertheless still do so authentically. If we assume, for the sake of the argument, that even if there were no good reasons, even unreflective, for saving his friend, i.e. that for Finn neither acting on moral reasons nor saving his friend was important for him, this would not prove that Finn did not save him authentically. It may be important for moral reasons to base the moral worth of actions on having good reasons for such actions, but in relation to authenticity having reasons of any kind is not relevant. Arpaly's theory is fruitful in the sense that she proves the non-importance of deliberation or reflection in actually acting rationally or being self-controlled. However, in terms of authenticity one more step is required in arguing that being rational in any sense and having good reasons for a decision or action is not necessary for acting authentically either.

In my view, in order for an attitude to be authentic, the reasons for it not only should not necessarily be known, but also they should not necessarily be good, and, in fact, they should not necessarily exist at all. What I discussed in the previous section stands for Arpaly's theory too. Attitudes that are authentic to a person may be the source of unreflective reasons and not vice versa or they may operate as reasons themselves and the authenticity of the former should not be based on the latter. Following from this, reasons of any kind are not necessary for authenticity. They might of course obtain and they might often be in line with the person's authentic attitude, but it is not they that constitute an attitude authentic.

For instance, in Frankfurt's case of the mother and the adoption, she has explicit reasons for wanting to give away the child, while she has inchoate reasons for wanting to keep it. None of these reasons, however, are adequate to render her attitude to give her child away or to keep it authentic. The feeling or intuition that creates the attitude of the mother to keep her child need not be based on any kind of reason, reflective or unreflective, in order for her to be authentic with respect to it. In further support of this, let us consider one more example:

> *Authentically self-destructive person.* Her reasons may not be good even for her, they may not make any sense even through the prism of her strong depression, but she continues to act in a self-destructive way that leads her to suicide.

The desire of this person to kill herself, even though she may not have any reason to do so, may still be more authentic than rationally deciding to avoid it. Even in the case that she considers all the good reasons not to act in such a way, they are still not strong enough to overcome her desire to harm herself. Committing suicide in her situation may be something completely irrational. This, however, does not prove that it is also something inauthentic. Irrational or non-rational persons can be authentic and in some occasions they can be even more authentic than rational persons.

## *6. Conclusion*

I have argued that the prominent contemporary autonomy conceptions can be divided into three categories, those which consider authenticity as i) necessary and sufficient for autonomy, ii) necessary but insufficient for autonomy, and iii) neither necessary nor sufficient for autonomy. Therefore, the line between where authenticity ends and autonomy begins and more importantly where the two overlap (if they actually do) is hard to be distinguished based on them. In addition, we have highlighted that many thinkers take for granted that authenticity should be based on rationality and self-reflection, i.e. on the exact same elements that autonomy is based on too. Given this, the occasions when authenticity comes into direct conflict with autonomy tend to be neglected and unexplored. If a more enriched and inclusive account of authenticity is proposed, not based only on the same features as autonomy, then authenticity could not simply be the basis of autonomy. Identification should not be misunderstood as either authenticity or autonomy per se. In terms of authenticity, there are cases that the person might not be able to identify with a desire of hers but still this desire to be authentic of hers. In this sense, Frankfurt's and Christman's theories of autonomy, even though they are equated with their understanding of authenticity, remain theories closer to the essence of autonomy than to authenticity. Moreover, I understand Dworkin's and Mele's theories as mainly theories of autonomy, which misuse the nature and role of authenticity in regard to autonomy. This said, the theories to which I referred are theories of autonomy that are identified with or based on authenticity.

Even if most thinkers tend to identify authenticity with autonomy or, at least, consider the one a core condition for the other, it is my view that for a person to be authentic with respect to an attitude, not only rationality and good reasons but also activity, wholeheartedness, reflection and unreflective reasons are neither necessary nor sufficient. Following from this, since the aforementioned conditions traditionally describe autonomy, we should distinguish the different nature and roles that authenticity and autonomy have in our everyday life.

Frankfurt's theory has critical flaws, since it does not take into account the personal history and development of the person. On the other hand, theories which incorporate the personal history of the agent are restricted to conditions founded solely on rationality, rendering them weak, inadequate and unrealistic. Nevertheless, the historical aspect is required for an adequate conception of authenticity and it should be retained, but without the necessity of the rational or any other kind of reflection, since, as I have claimed, reflection in any form cannot guarantee authenticity. This said, in short, the historical condition required for authenticity needs to be based on an enriched conception of creativity that I shall develop in a following article and it is developmental, externalist, non-intellectualist, non-rationalist and content-neutral.

More precisely, in contrast to the majority of prominent theorists of autonomy and authenticity, who base their conceptions of authenticity on rationality, I shall base mine on creativity, while I also explore other relevant notions, such as novelty, originality, and imagination. Furthermore, while all theories of authenticity require the existence of a true self or at least some kind of self, I shall put forward a conception that is not a 'self-expression' view of authenticity; that is, the theory proposed will not require a substantial theory of the self. Creativity has been widely understood as the production of something that is original and valuable in some way. My aim is to develop a conception of creativity designed specifically to help us understand authenticity. I shall focus on what a creative process is, and understand it in terms of a psychological conception of novelty and of sensitivity in regard to the intrinsic value of the creative outcome.

It would be, however, a critical miscomprehension of my theory to construe it as individualistic and lacking social/relational elements. I am not denying the importance of social interrelations with other persons and social entities in the formulation of authentic creations. On the contrary, the account proposed involves both social and asocial aspects. Besides, there cannot exist ex-nihilo creations, i.e. outcomes of parthenogenesis. Whereas manipulation, oppression and coercion bypass creativity and authenticity, more voluntary forms of influence enhance them. One is endlessly creating one's inner nature, not through an inward self-directed direction, but in a constant creative feedback with one's social reality. Both individual and social life can be radically transformed through creativity, and in this sense creativity and authenticity are capable of potentially playing a crucial transformative role in both an individual and a collective level.

Against the simplification of founding authenticity solely on reflective rationality, my aim is to grasp a more complete image of our nature. In my view, creativity is a more wholly human capacity than mere rationality and in this respect is more appropriate to operate as a core condition of authenticity. Hence, based on the above, I shall argue for a new view of authenticity and its relation to autonomy. The motivation behind the view I am considering is to pull apart authenticity from autonomy, reflective rationality and the self, which I believe seriously restrict it, and to direct it towards imaginativeness and creativity, where it may be more at home.

Nikos Erinakis
Department of Philosophy, University of Athens
nikos.erinakis@gmail.com

## *References*

Arpaly, N., 2003, *Unprincipled Virtue: An Inquiry into Moral Agency*, Oxford University Press, New York.

Christman, J., 1991, "Autonomy and Personal History," in *Canadian Journal of Philosophy*, 21: 1-24.

—, 1993, "Defending Historical Autonomy: A Reply to Professor Mele," in *Canadian Journal of Philosophy*, 23: 281-289.

—, 2005, "Procedural Autonomy and Liberal Legitimacy," in: Taylor, J. S., ed., *Personal Autonomy. New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy*, Cambridge University Press, Cambridge, 277–298.

—, 2009, *The Politics of Persons: Individual Autonomy and Socio-historical Selves*, Cambridge University Press, Cambridge.

Dworkin, G., 1988, *The Theory and Practice of Autonomy*, Cambridge University Press, Cambridge.

Frankfurt, H., 1988, *The importance of what we care about: philosophical essays*, Cambridge University Press, Cambridge and New York.

—, 1999, *Necessity, Volition, and Love*, Cambridge University Press, Cambridge.

—, 2002a, "Reply to Gary Watson," in Buss, Sarah, and Lee Overton, eds., *Contours of Agency: Essays on Themes from Harry Frankfurt*, MIT, Cambridge MA, 160-164.

—, 2002b, "Reply to Richard Moran," in Buss, Sarah, and Lee Overton, eds., *Contours of Agency: Essays on Themes from Harry Frankfurt*, MIT, Cambridge MA, 218-225.

Ionesco, E., 1960, *Rhinoceros and Other Plays*, trans. by Derek Prouse, Grove Press, New York.

Mele, A., 1993, "History and Personal Autonomy," in *Canadian Journal of Philosophy*, 23: 271-280.

Mele, A., 1995, *Autonomous Agents: From Self-Control to Autonomy.* Oxford University Press, Oxford.

Mele, A., 2002, "Autonomy, Self-Control and Weakness of Will," in Robert Kane, ed., *The Oxford Handbook of Free Will*, Oxford University Press, Oxford.

Moran, R., 2002, "Frankfurt on Identification: Ambiguities of Activity in Mental Life," in Sarah Buss and Lee Overton, eds., *Contours of Agency: Essays on Themes from Harry Frankfurt*, MIT, Cambridge, MA, 189-217.

Twain, M., 2008, *Adventures of Huckleberry Finn*, Oxford University Press, New York.

# Focus

# 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)

Submitted: 14 November 2022
ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i1.442 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.

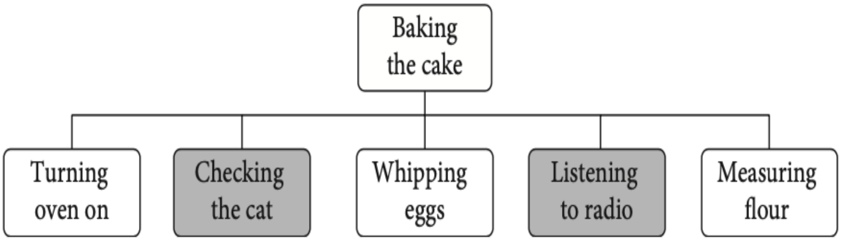
Finally, given the nature of the attitudes at the center of contemporary theories, they evaluate the rationality of an agent at a specific point in time. If the only attitudes relevant to the evaluation of actions are the ones the agent has at the time of the action, we have a time-slice theory of rationality. 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

[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.

marathons. As you go out for your first training run while singing, you realize 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. <u>Preference Relative to an End</u>
> 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

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

done under the guise of the good; the structure is inherited from the nature of the good you are pursuing. 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.

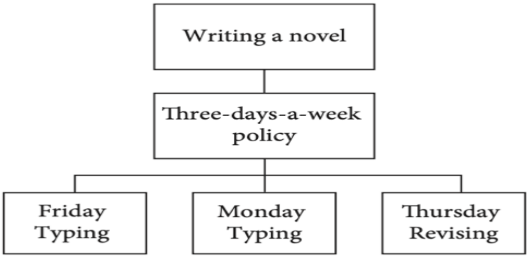
means to retrieving the Holy Grail and the abandonment of the end were concomitant. Thus she is always in compliance with the principle of instrumental reasoning; after all, reason does not tell us never to abandon our ends. In fact, Shifter might do this consistently: conscious of her cowardice, she 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.

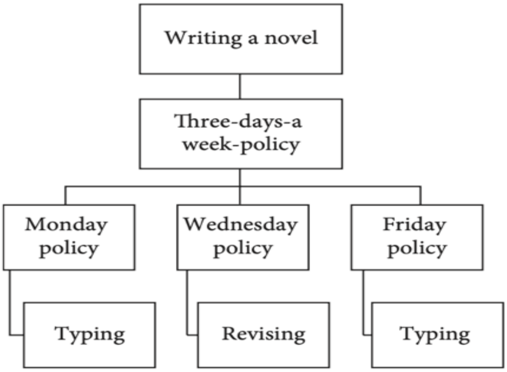
Extended agency gives rise to a problem of managing the pursuit of our ends through long periods of time, when at each particular time we might prefer not to take means to this end. As mentioned above, I am rationally permitted throughout this interval to act against my preferences so as to take the 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 pursue 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:



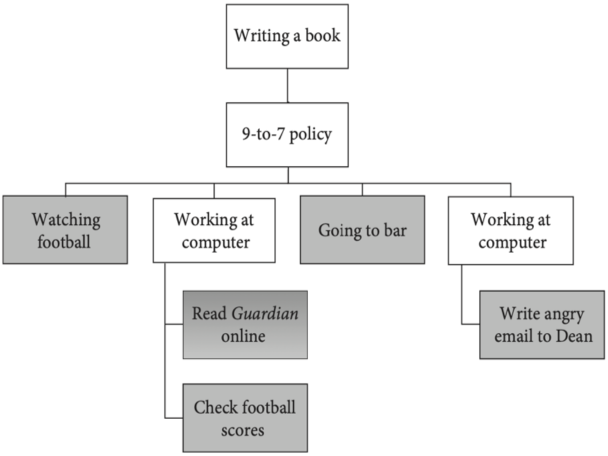
More flexible intermediate policies

[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.

# Rational powers and inaction

Sarah K. Paul

Abstract: This discussion of Sergio Tenenbaum's excellent book, *Rational Powers in Action*, focuses on two noteworthy aspects of the big picture. First, questions are raised about Tenenbaum's methodology of giving primacy to cases in which the agent has all the requisite background knowledge, including knowledge of a means that will be sufficient for achieving her end, and no significant false beliefs. Second, the implications of Tenenbaum's views concerning the rational constraints on revising our ends are examined.

*Keywords:* Sergio Tenenbaum, instrumental rationality, trying

*Rational Powers in Action* is a brilliant book. It is an extensive, resourceful, enjoyably-written articulation and defense of a genuinely new theory of instrumental rationality. It seeks to overthrow the tyranny of orthodox decision theory, understood as a theory of instrumental rationality, but it does so from within a profound grasp of that tradition. Further, the book takes aim at the relatively widespread view that "future-directed intentions" are attitudes governed by distinctive rational norms of non-reconsideration and persistence. Those who are inclined to continue holding these views – like myself, in the latter case – will have to contend going forward with Tenenbaum's powerful arguments against them.

In this response, I want to focus on two aspects of the big picture that I find especially interesting, at the unfortunate expense of leaving many of the central arguments untouched. First, I will discuss Tenenbaum's methodology of giving primacy to cases in which the agent has all the requisite background knowledge, including knowledge of a means that will be sufficient for achieving her end, and no significant false beliefs. Second, I will turn to the claims that the view makes about the rational constraints on revising our ends.

ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i1.443

Submitted: 14 November 2022
Accepted: 18 November 2022

## 1. *Uncertainty, error, and trying*

I’d like to start by bringing out an aspect of Tenenbaum’s approach that is not fully committed to, or explicitly defended at length, but that I think goes deep into the foundations. One major aim of the theory, as Tenenbaum characterizes it, is to vindicate the idea that practical reason extends all the way to intentional action – to what is real, and nothing short of that. Much of the book is devoted to arguing against the idea that the inputs into a theory of instrumental rationality must be mental attitudes or events like preferences, desires, intentions, or choices, understood as phenomena that are metaphysically distinct and separable from action. The central thesis is that “instrumental rationality is rationality *in action*” (2020: viii). Further, Tenenbaum argues that the principles of *ETR* Derivation, *ETR* Coherence, and *ETR* Exercise are the only basic principles that govern the exercise of our instrumentally rational powers (see Tenenbaum’s précis in this journal for statements of these principles).

This means that whenever an agent is legitimately required by the principles of instrumental rationality to take means to her extended ends, and to ensure that her ends are consistent with one another, there must be relevant intentional actions going on. The theory risks extensional inadequacy if there are good reasons to doubt that whenever the agent has an extended end that is a source of instrumental pressures, there is a corresponding intentional action occurring. The *ETR* addresses this worry by employing a quite broad conception of intentional action, and by emphasizing the indeterminacy that is present in nearly every end we pursue. Tenenbaum argues that most extended actions are “gappy,” in the sense that they are compatible with substantial periods of inactivity (2020: 70). We can be getting in shape, for example, without actively doing anything to contribute to that end for an extensive amount of time. Indeed, on his view, intending to do something in the future is simply an instance of intentional action in which there is a gap in the beginning, unpreceded by any active part. If I now (in winter) intend to get in shape next summer, I already count as pursuing the end of getting in shape, though all the active parts of my action have yet to occur. And (luckily for us), the end of getting in shape is indeterminate in the sense that there is quite a bit of vagueness as to what counts as succeeding or exactly when I must act to bring about success. The structure of the pursuit does not require me to do much of anything at any particular moment; I simply need to do enough exercising over time to count as being sufficiently in shape, by my own lights, at some indeterminate point in time.

Further, the principles govern actions that are in progress. And goal-directed actions in progress are subject to the so-called ‘imperfective paradox’: one can *be doing* something that one never ends up successfully having done. I can currently be getting in shape without ever ending up in shape. These features of the intentional pursuit of indeterminate ends, characterized in the progressive, collectively serve to break down the barrier that intuitively exists between having an end and actually acting in pursuit of it.

At the same time, we might worry that this way of thinking about intentional action raises a new threat of unreality, insofar as tangible progress toward one's goal is rarely required. This suggests that our powers of instrumental rationality might often fall short of leading us to actually achieve our ends. When we are operating with false beliefs, or are uncertain about how to realize our ends, the reality of effectiveness threatens to remain largely in our minds. So we might ask: just how real is the rational meant to be, according to the *ETR*? Where does Tenenbaum's view stand on the question of whether it is necessarily a defect in one's instrumental rationality to fall short of achieving one's ends?

It seems to me that the book is ambivalent about this question. On one hand, a striking feature of Tenenbaum's approach is that for most of the book, he formulates the central *ETR* Derivation principle in terms of knowledge: he assumes that the instrumentally rational agent has knowledge of some sufficient and contributory means to her ends, and no false beliefs that will interfere with her effectiveness. This choice puts the focus on the kind of case in which the agent knows just what she needs to do in order to, say, become a profitable stand-up comedian, rather than on the case in which she is uncertain about what it will take, or in which she falsely believes that her innate talent for improvisation will suffice. The assumption does much to exclude the possibility of massive failure, since it follows that the conclusion of instrumental reasoning just is the intentional pursuit of means known to be (jointly) sufficient or contributory to success. The implication is that we only exercise our powers of instrumental rationality without defect in those cases where we know how to achieve our ends and are therefore in a position to be genuinely effective.

That said, Tenenbaum gestures in the final chapter at the possibility of giving this assumption up and reformulating the Derivation principle in terms of belief rather than knowledge. The instrumentally rational agent would then be understood as deriving means to her ends by way of beliefs that are potentially false, and thus failing to be truly effective. At the same time, Tenenbaum indicates a preference to hold onto the knowledge version, thereby understanding instrumental rationality in terms of actual effectiveness. Compare a similar claim he has defended elsewhere concerning deontological theories of morality: the deontic status of an act does not depend on the agent's epistemic states (Tenenbaum 2017). When it comes to morality, we might think, we are required to keep our promises, not merely to do what we believe would amount to keeping our promises, or what would be most likely to amount to such. Likewise, the idea would be that we are instrumentally required to take the actual means to our ends, not to do what we believe would be effective, or what would likely be effective. An agent who falsely believes there is water in his glass is failing to be instrumentally rational when he takes a drink of petrol, since this action will in fact do nothing to further his end of quenching his thirst. The power of instrumentally rational agency is the power to get things done; thus, the power is not exercised in the same way in the case of knowledge and in the case of error.

This is a fascinating conception of instrumental rationality, but also radical and in some ways counterintuitive. The thirsty agent *does* seem to be instrumentally rational in taking a drink; his practical reasoning strikes many of us as impeccable, structurally speaking, though his beliefs happen to be inaccurate. Is the *ETR* committed to this "factive" view about instrumental rationality? Tenenbaum claims not, stating that we could simply revise the minor premise of the Derivation principle to refer to the agent's beliefs rather than her knowledge. However, I want to suggest that such a revision would not in fact sit easily with other aspects of the view. To deal with the problem of false beliefs in this way would be at least potentially at odds with the way the *ETR* approaches the problem of uncertainty.

We lack knowledge of the minor premise of the Derivation Principle not only when we have false beliefs, but also when we are uncertain about how to achieve our ends. This is a relatively common situation to be in, especially with respect to high-level ends that are difficult to achieve – competitive careers, advanced degrees, long-term relationships, health, wealth, and happiness, among others. We strive to achieve these ends, but we often do not know of any means that it will suffice. And in response to such uncertainty, we sometimes formulate our intentions as disjunctive or conditional on whether some currently unknown circumstance will obtain, committing ourselves only to keeping certain options open until we figure out more specifically what we want to do. We intend things like "to pursue a PhD if we are admitted to a good program with full funding," and if not, "to either enroll in law school or go backpacking in Europe."

To address this challenge to the *ETR*, Tenenbaum denies that we *can* pursue ends if we are uncertain about how to achieve them. Rather, he argues, risk and uncertainty change the nature of the actions available to us. "If I realize that none of the means available to me can ensure that I earn a million dollars," he writes, then 'becoming a millionaire' is not a possible intentional action for me" (2020: 205). Rather, one must adopt the related end of 'trying to become a millionaire', which is a different action that involves distinct sufficient and contributory means. This resourceful move allows Tenenbaum to keep the basic structure of the view in place, since an agent who lacks knowledge of a sufficient means of E-ing may yet have knowledge of a means that is sufficient for trying to E.

However, what are the grounds for thinking that uncertainty about whether we can E prevents us from even having that end? Can't I have the end of writing a successful book even if I am uncertain about whether I can do it? (Of course, I know in some sense what it is one does in order to write a book, but I am very uncertain about whether a successful book will result if *I* take those means). The obvious thing to say in defense of this claim is that intentionally E-ing requires "practical knowledge" that one is E-ing – a claim often attributed to G.E.M. Anscombe. If one does not know how to write a good book, it follows that one could not have practical knowledge of writing it, and therefore that one could not be writing a good book intentionally. But Tenenbaum attempts to stay neutral about this Anscombean idea for the purposes of his book. And more importantly, endorsing that idea would be in tension with the possibility of revising the *ETR* to allow for instrumental reasoning to proceed by way of false beliefs. After all, the agent acting in light of false beliefs would presumably lack practical knowledge as well, at least under some descriptions that are essential for understanding what is rational about her action. The agent drinking petrol does not know he is quenching his thirst (because he isn't), and so he could not be manifesting his instrumentally rational powers in pursuit of that end.

Perhaps there is an independent motivation for the idea that trying to E is a substantively different action from doing E, one that makes no appeal to controversial claims about practical knowledge. It is true that we often talk this way (though it is not clear that Tenenbaum would want to say that we should be guided by common parlance in every case, as I'll explain in a moment). But talk can be superficial, and the important question is whether 'trying' really has an internal structure that will yield plausible results about what is instrumentally required of an agent who is trying. Note that many of the high-level ends that play an important constraining role on the *ETR* view will presumably be cases of trying. For example, the solution to the problem of the self-torturer appeals to the end of "living a relatively pain-free life." Tenenbaum also talks about the end of living "a good and happy life," understood as the joint realization of the totality of our other ends. These kinds of high-level ends will be implicated at almost all moments, and do important work by issuing permissions that allow us to violate our Pareto preferences. But surely most of us do not know of any means that is sufficient to prevent chronic, debilitating pain or deep and persistent unhappiness. We're simply trying to avoid these things. So it seems important to understand exactly what the theory says when it comes to trying.

Now, 'trying' is a very slippery concept. There is an anemic sense of trying in which it is enough to lift a finger, which means that an agent who is trying in this sense incurs almost no instrumental obligations. Tenenbaum sets this notion aside and focuses instead on a more substantive reading, which he glosses as “doing my best to succeed under the circumstances” (2020: 214). According to the *ETR*, then, instrumental rationality in pursuit of the end of trying to E will be a matter of pursuing some means or set of means known to be sufficient for trying, understood as doing one’s best under the circumstances. To understand this, we therefore need to have some grasp of what the success conditions for “doing one’s best” are.

I’m not convinced that there is a determinate standard here that is internal to the structure of the activity of ‘doing one’s best’, as opposed to the context-dependent, external standards we might use to praise or blame the agent’s efforts. The agent himself will not think of his aim as ‘doing his best,’ or conceive of the standards of success as something other than achieving his end. Indeed, if he does not achieve his end, he will take himself to have failed in his pursuit. And he will not reason about how to do his best, under that description; this sounds like what you should do if you are trying to *appear* to have done your best, to escape censure. Rather, a rational agent who is really trying to accomplish the end will take whatever acceptable means are available to achieve the end, not merely those that will suffice for having done his best. And he will rule out any other pursuits that would cause him to fail at the end he is trying to achieve, not merely those that would cause him to fail to try. Staying out all night at a party with friends is not obviously incompatible with *trying* to complete a marathon the next day, but an instrumentally rational agent will rule this out as being incompatible (let’s suppose) with *succeeding* at running the marathon.

The point is that the standards a rational agent holds himself to when he is really trying seem to derive from the end itself, and not some lesser measure of success. This makes it difficult to see why we should suppose that uncertainty necessarily renders the pursuit of that end unavailable to the agent. To be sure, the more anemic sense of trying does seem to have a different internal structure and generate few if any instrumental requirements. But the existence of the other, more committal form of trying is enough to cast doubt on the strategy of handling cases of uncertainty in the way Tenenbaum does.

We might try falling back on the idea that ordinary language encourages us to describe our actions in terms of trying when we are uncertain of success. But this would put the ETR in a difficult position with respect to other pursuits that do not fit well with ordinary language. Consider the sorts of logically complex intentions mentioned earlier, with a disjunctive or conditional structure: intending to do X if C, or to do either X or Y depending on how certain future events unfold. Such commitments are undoubtedly subject to demands of instrumental rationality; at the least, we are irrational if we do not act so as to preserve the possibility of X-ing or Y-ing should the relevant circumstances arise. Common parlance does not support the idea that there is an ongoing action to do the needed work, however. If my intention is 'to walk to the library if Ivy is there' or 'to walk to either the library or the store', it is quite a stretch to say that I *am now* doing those things – especially if I haven't moved from my couch because I don't know yet whether Ivy is at the library. These kinds of cases suggest that Tenenbaum should not wish to put too much weight on the surface grammar of act-descriptions.

To take stock: what I have been trying to illustrate in this section is that difficult questions arise when we consider agency in the face of uncertainty, and I worry that the book treats these difficulties too lightly. Tenenbaum wants to avoid committing to the more radical interpretation of the view, according to which our instrumentally rational powers are only fully exercised without defect when we know how to bring about our ends and are thus able to be effective. But it is not so straightforward to simply reformulate the view in terms of belief or credence rather than knowledge. If knowledge is not required in order to take means to our ends, then it is unclear why we should suppose that uncertainty changes the ends we can pursue, relegating us to trying rather than doing. There are good reasons to doubt that there is always a deep distinction here from the perspective of our instrumental obligations, and the fact that we draw this distinction in ordinary language carries little weight once we notice that the ETR will need to depart from ordinary parlance in characterizing some of our more logically complex ends. The approach of treating cases of uncertainty and error as substantively different from cases of knowledge therefore seems unmotivated, in the absence of a more explicit commitment and full-throated defense of the idea that instrumental rationality should be understood in a factive way.

## 2. *Virtues, vices, and patterns of end-revision*

Let me now turn to a different aspect of Tenenbaum's account. First, a brief comment on Tenenbaum's treatment of the role of future-directed intentions and policies in the framework of the *ETR*. Philosophers have generally treated policies and future-directed intentions – intentions to perform an action that will begin at a later time – as attitudes of some sort. And many have thought they are the kind of thing to which norms or principles of instrumental rationality apply. For instance, some have argued that norms of structural rationality govern the coherence and persistence of our future-directed intentions over time. Perhaps we ought not to reconsider our intentions without good reason, for example, on pain of exhibiting a form of incoherence over time that will make us vulnerable to temptation and otherwise prevent us from being effective.

These claims pose a challenge to the *ETR*. In response, Tenenbaum argues that we can understand policies and future-directed intentions as extended actions rather than attitudes, at least with respect to their internal structure. A policy of calling one's mother once a week is not relevantly different, he argues, from intentionally pursuing the end of calling her once a week (2020: 126). And as we saw earlier, he denies that future-directed intentions are fundamentally different in kind from other instances of extended action; on his view, they are simply actions in which there is a gap in the beginning, unpreceded by any active part. If we accept these conclusions, then policies and future-directed intentions turn out to be the kind of thing – extended action – to which principles of instrumental rationality can apply. That said, Tenenbaum argues extensively against the existence of non-derivative requirements enjoining intention stability or forbidding reconsideration in any particular instance. On his view, an agent can be perfectly instrumentally rational from the extended perspective, executing their intentions and policies through their actions in the knowledge that the overall pattern will suffice, without obeying any strict requirement never to reconsider or shuffle their intentions arbitrarily. They simply have to avoid doing these things too much.

This sounds eminently reasonable. But *how* do we avoid doing such things too much? Tenenbaum likes to quote Leonard Cohen lyrics to demonstrate the possibility and appeal of having a policy of faithfulness "give or take a night or two" (2020: 133). The problem is that like the lover to whom Cohen's song "Everybody Knows" was addressed, people often end up taking a lot more than a couple of nights. Tenenbaum grants that there is a place in our theory of instrumental rationality for such things as resoluteness, constancy, and self-control, but he categorizes these as instrumental virtues rather than a matter of adhering to certain principles. I'll admit to having the kind of philosophical constitution that is frustrated by talk of powers and "dispositions of the will." These sound to me like names for certain patterns of behavior, when what I want to understand is the mechanism behind those patterns. Attempting to conform to a principle is one possible mechanism for achieving an acceptable pattern, and even if the content of the principle is unjustifiably strict, the *acceptance* of that principle by the agent might be justified by appeal to its results. Viewed this way, it might be true as Tenenbaum argues that *if* we non-accidentally end up satisfying our goals and policies, we cannot be deemed instrumentally irrational for all the reconsidering, procrastinating, self-indulging, and vacillating we did along the way. And yet the best way to ensure that we non-accidentally succeed in satisfying our goals and policies might be for us to view any such lapses as problematic. In other words, the best mechanism might be overkill.

At any rate, I want to raise a slightly different question about this part of the account. In the first part of the book, Tenenbaum defends an implication of the *ETR*, which is that there are no determinate rational restrictions on how one should revise one's ends when they come into conflict with one another. An agent in this situation can abandon either of the conflicting ends, adopt a higher-order end of giving priority to one or the other, or simply revise each of them to be more restricted so that they no longer conflict (i.e "do enough of each"). The *ETR* does not offer guidance on which way to go, and Tenenbaum claims that this is a virtue, since theories of rationality that tell us how to choose between our ends run afoul of what he calls the Toleration Constraint: a theory of instrumental rationality should avoid putting restrictions on the contents of the given attitudes, except as necessary for meeting the standards of success of these representations as defined by the theory (2020: 20).

But some instances or patterns of end-revision *are* intuitively problematic. For instance, when it comes to adjusting one's ends toward mutual compatibility, there is a difference between legitimately prudent satisficing and throwing your standards out the window. Sometimes there really is room to do well enough at everything you're committed to, but in other cases, you ought to give up at least one of your commitments rather than doing everything poorly. The distinction here belongs at least in part to instrumental and not merely substantive rationality, I think, in that the tendency to lower your standards too far is not really a way of effectively achieving all of your ends; it is more akin to akrasia. Tenenbaum himself brings up other problematic cases of end-revision in Chapter 7, where he discusses the idea of instrumental virtue and vice. He examines a case of a self-aware coward who always adjusts his ends so that he never finds himself in a position of continuing to have an end while chickening out about the means (2020: 177). Akrasia can take this form as well; when one notices that a judgment, intention or policy conflicts with the action one is really tempted to take right now, one might simply revise the pesky judgment or intention to eliminate the conflict. Inconstancy and irresoluteness can similarly occur without leading the agent to fail to take the necessary and sufficient means to any end she maintains throughout the relevant period. Thus, one of the central points of this chapter is that these problematic patterns of end-revision need not involve the failure to comply with any instrumental principle, and need not even involve *acting* irrationally. Rather, on Tenenbaum's view, they are defects in the agent's will.

I wonder whether this claim doesn't water down the initial thesis a fair bit, and put us in danger of running afoul of the Toleration Constraint. It turns out that many instances or patterns of end-revision in the face of conflict may be criticizable on broadly instrumental grounds even if they are permitted by the principles of *ETR*. And the objects of criticism are not extended actions, which means that instrumental rationality is not only a matter of "rationality in action;" it also includes dispositions of the will. Further, the *ETR* faces a challenge in explaining why some patterns of end-revision are instrumentally problematic if they never lead to a failure to take the means to one's ends. Intuitively, the *ETR* should want to explain the coward's pattern of behavior by attributing to him the high-level end of leading a danger-free life no matter what, leading him always to prioritize his own safety. But the Toleration Constraint advises us not to criticize him on those grounds.

Tenenbaum suggests instead that the instrumental defect lies in the fact that some dispositions of the will make some ends unavailable, no matter how good the agent represents them as being. We might think, however, that the ability to render some ends unavailable to ourselves is an instrumental *virtue*, insofar as things like cowardice, temptation, and fickleness incline us to take some ends to be good when they are not, and insofar as we can recognize about ourselves that this is so. The agent who is prone to temptation will be more effective at achieving her true ends if she can render the objects of temptation unavailable to her will at the key moments. Of course, the vicious agent renders the wrong ends unavailable to herself. So we would like some way of saying, without appealing to objective facts about which ends are legitimate, that some restrictions of the will are beneficial and some defective.

As I see it, this is a major motivation behind the idea that there is rational pressure to stick with a previous decision or conform to a policy, even if it conflicts with how one views things now. Theories of practical rationality that include norms of intention non-reconsideration or persistence are in a comparatively good position to explain how we can restrict our own wills over time without making substantive judgments about the legitimacy of any particular end. Tenenbaum critiques the way this basic thought has been developed in terms of strict principles or policies, and I think his points are well taken. But I am not yet sure how radically different his solutions are, insofar as they appeal to virtues of the will that are distinct from intentional action. Either way, it turns out that a fully instrumentally rational agent must do more than simply preserve means-end coherence and consistency somehow or other, with no constraints on how she adjusts her ends in order to do so. I should note that Tenenbaum sees his account of instrumental virtue and vice as being largely independent of the main *ETR* thesis. But it does seem to me that a theory of instrumental rationality should have something to say about why certain patterns of end-revision count as problematically inconstant, irresolute, akratic, or cowardly, and it looks as though this will require resources that go beyond the internal structure of intentional action.

## *Acknowledgements*

I am indebted to Luca Ferrero, Matthias Haase, and Sergio Tenenbaum for conversations that have greatly helped to shape the final version of these comments.

Sarah K. Paul
Philosophy Program, New York University Abu Dhabi
skp5@nyu.edu

## *References*

Tenenbaum, Sergio, 2020, *Rational Powers in Action: Instrumental Rationality and Extended Agency*, Oxford University Press, Oxford.

—, 2017, "Action, Deontology, and Risk: Against the Multiplicative Model," in *Ethics*, 127: 674-707.

# Instrumental rationality and proceeding acceptably over time

Chrisoula Andreou

*Abstract*: Theories of instrumental rationality often abstract away from the fact that actions are generally temporally extended and from crucial complications associated with this fact. Sergio Tenenbaum's *Rational Powers in Action* (2020) reveals and navigates these complications with great acuity, ultimately providing a powerful revisionary picture of instrumental rationality that highlights the extremely limited nature of the standard picture. Given that I share Tenenbaum's general concerns about the standard picture, my aim is to advance our general approach further by complicating and enriching debate regarding a picture of instrumental rationality that is accountable to the temporally extended nature of our actions and agency via the consideration of a few issues that merit further consideration and exploration. As I explain, despite stemming from or being associated with some important insights, some of the central ideas that Tenenbaum supports need to be qualified, modified, or reconsidered.

*Keywords:* cyclic preferences, incommensurability, instrumental rationality, satisficing, momentary versus extended actions, vague ends or projects

Theories of instrumental rationality provide, roughly speaking, evaluations and imperatives regarding choice or action that figure as relative to certain basic given attitudes or stances of the agent. Such theories often abstract away from the fact that actions are generally temporally extended and from crucial complications associated with this fact. Sergio Tenenbaum's *Rational Powers in Action* (2020) reveals and navigates these complications with great acuity, ultimately providing a powerful revisionary picture of instrumental rationality that highlights the extremely limited nature of the standard picture (which focuses on the selection of momentary acts, chosen and effected—in auspicious cases wherein they are not blocked—at a choice point).[1] Given that I share Tenenbaum's general concerns about the standard picture, this symposium paper will lack the drama of a piece aimed at devastating criticism. Instead, my aim is to continue to advance the project of complicating and enriching

[1] All page references to Tenenbaum's work will be to (Tenenbaum 2020).

ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i1.444

Submitted: 14 November 2022
Accepted: 18 November 2022

debate regarding a picture of instrumental rationality that is accountable to the temporally extended nature of our actions and agency by raising some issues that merit further consideration and exploration.

My focus will be on three central ideas that Tenenbaum supports. First, I will focus on the idea that, given how the pursuit of ends over time often works, "someone may be irrational over a period of time without there being any moment during that time at which they were irrational" (viii). Second, I will focus on the idea that an instrumentally rational agent will often have to seek "acceptable" realizations of her ends rather than maximizing, and not due to the agent's bounded rationality but due to the nature of the ends themselves.[2] Finally, I will focus on the idea that an agent may be rationally permitted to waver between options in a way that involves her incurring costs that she could have avoided had she resisted "brute shuffling," though not if the costs are devastating.[3] As will become apparent, I think that each of these ideas needs to be either qualified, modified, or reconsidered, despite stemming from or being associated with some important insights.

With respect to the first idea, consider Tenenbaum's example of an agent with the vague and indeterminate end of writing a book. Suppose, in particular, that you are the agent in question and that, as Tenenbaum explains, the following conditions hold:

> (i) [The project's] completion requires the successful execution of many momentary actions.
> (ii) For each momentary action in which you execute the project, failure to execute that action would not have prevented you from writing the book.
> (iii) On many occasions when you execute the project, there is something else that you would prefer to be doing, given how unlikely it is that executing the project at this time would make a difference to the success of your writing the book.
> (iv)Had you failed to execute the project every time you would have preferred to be doing something else, you would not have written the book.
> (v) You prefer executing the project at every momentary choice situation in which you could work on the project over not writing the book at all. (100-101)

For Tenenbaum, if, rather than succeeding, you failed to write the book as a result of having failed to execute the project every time you would have pre-

[2] Here and elsewhere, I use one of the singular personal (sometimes referred to as "preferred") pronouns "she," "he," or "they" rather than the unwieldy "she, or he, or they." I will not continue flagging instances in which "she, he, or they" is replaced with one of "she," "he," or "they."

[3] The phrase "brute shuffling" is borrowed from Michael Bratman (2012), who describes it as "lurching from one plan-like commitment to another incompatible commitment seen as equal or incomparable, in a way that involves abandoning one's prior intentions" (81).

ferred to be doing something else, you would count as irrational even though there is (Tenenbaum suggests) no particular moment at which you proceeded irrationally given that (by hypothesis) no particular momentary failure to pursue an end-directed action took you from being in a position to write the book to not being in a position to write the book.

Notably, Tenenbaum's view that some (rationally permissible) ends are indeterminate and vague is controversial, but I think he is right about this, and so I will accept this as common ground. Still, we should not jump to the conclusion that, in the contemplated case of failure, there is, other things equal, no moment at which you were (proceeding) irrational(ly). My reason for hesitation is based on the distinction between what is realized *in* a moment and what is being done *at* a moment (which I will briefly discuss here and which I say a great deal more about elsewhere).[4]

As Tenenbaum recognizes, doings are rarely momentary in the sense of being completed in a moment. Still, one can say of an agent engaged in the doing in progress of $\phi$-ing between $t_1$ and $t_n$, that the agent is, *at*, say, $t_k$, $\phi$-ing. For example, if the agent is making a cake between $t_1$ and $t_n$, then they are, *at*, say, $t_2$, when they turn on the oven, making a cake; importantly this holds even if they are interrupted and never complete the doing in progress of making a cake because they accidentally burn up the kitchen soon after turning on the oven. More generally, although a doing in progress *at* $t_x$ is not contained *in* $t_x$, and so we can say, to quote Michael Thompson (2008: 126), that the doing in progress "reach[es] beyond" $t_x$, its being in progress is not (to quote Thompson again) exposed to "simple disproof on the strength of what happens next" (2008: 126), since the doing in progress can be interrupted immediately after $t_x$.

To see that the distinction between what is realized *in* a moment and what is being done *at* that moment is potentially relevant in the above-mentioned failed book project case, consider the following: It may be that, although none of the agent's momentary doings—understood as doings completed in a moment—are irrational, the agent is, nonetheless, irrational *at* one or more of these moments because the agent is engaged in a doing in progress that reaches beyond her "momentary action" and is unacceptable relative to her end.[5] For instance, she may—given her dispositions and capacities, which, as Tenenbaum emphasizes, do not "easily show up in a snapshot of the agent's mind" (186)—be frittering away her life (which can be true at $t_x$ even if, unlike in the failed book project case of interest, her doing in progress of frittering away her life were, shortly after $t_x$, interrupted by, say, an unexpected transformation after a near death

[4] See, especially, (Andreou 2014), which I draw on in the next few paragraphs.

[5] For detailed discussion regarding relevantly similar cases, see (Andreou 2014).

experience prompting a life of great social and scholarly achievements). Perhaps there is invariably a problematic doing in progress in the cases of failure that are of interest.[6] Let me explain; and keep in mind that I am not assuming that there is anything inherently problematic about "frittering away" one's life (though there may be) but only that doing so is problematic if it is unacceptable relative to one or more of the agent's ends.

First note that I here allow, following Tenenbaum, that "one can be pursuing the end of ϕ-ing even while *at the same time* failing to take the necessary means to ϕ-ing, as long as pursuing an end extends through time" (128). As Tenenbaum explains, in such cases of failure, although one is failing to take the means to one's end, one is also doing certain things "that are intelligible only if taken as means to [one's] (failing) pursuit" (129). For example, in the book project case, one may be spending a great deal of time in front of one's computer with a Word document entitled "book" open (even if one also has several webpages open that one is browsing through). Note also that one need not be happy with the fact that one is, say, frittering away one's life to be accountable for this doing in progress. Relatedly, one can be accountable for this doing in progress just as one can be accountable for omissions like failing to take the necessary means to one's end; moreover, one can be failing to take the necessary means to one's end via the doing in progress of frittering away one's life.

Now, why not think that all cases of non-accidental failure (such as, for example, the case in which an agent, despite having the end of realizing long-term project P, has been frittering away her life, continues to fritter away her life, and ends her life having frittered it away) are ones in which, at least at some moments, there is a doing *in progress* that is incompatible with the agent's end, and that the agent is irrational *at* these moments, even if her momentary actions, which are contained *in* the relevant moments, are not irrational? This possibility should, I think, give us pause with respect to the suggestion that one may be irrational over a period of time without there being any moment during the relevant time frame *at* which one is irrational. Importantly, it can still be true that it is the irrationality of the doing in progress, which reaches beyond one's momentary actions, that explains one's irrationality at various moments during the relevant time frame rather than vice versa. And this point seems compatible with Tenenbaum's "non-supervenience thesis," according to which "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 $t_n$" (47), which seems quite right, even if we should resist or at least be skeptical about Tenenbaum's stronger suggestion/gloss that "an agent might be rational at each moment $t_x$ such that $t_x$ is within

[6] For in-depth discussion pertaining to this possibility, see (Andreou 2014).

the interval $t_{[1]}$ to $t_n$, and yet not be rational at interval $t_1$-$t_n$" (48).

Turn next to Tenenbaum's view that an instrumentally rational agent will often have to seek "acceptable" realizations of her ends rather than maximizing, and not due to bounded rationality but due to the nature of the ends themselves. Consider Warren Quinn's puzzle of the self-torturer,[7] which Tenenbaum describes (with some discretionary adjustments) as follows:

> A person has agreed to wear a device that delivers a constant but imperceptible electric shock. She, the self-torturer (ST), is then offered the following trade-off: she will receive a large sum of money—say, $100,000—if she agrees to raise the voltage on the device by a marginal, that is, imperceptible or nearly imperceptible, amount. She knows that she will be offered this same trade-off again each time she agrees to raise the voltage. It seems that, at each step of the way, the agent should and would raise the voltage; after all, each rise in voltage makes at most a marginal difference in pain, well worth a gain of $100,000. But in so doing, she would eventually find herself in unbearable pain, and would gladly return all of the money, even pay some in addition, to be restored to the initial setting, at which she was poor but pain-free. Thus the ST appears to face a dilemma: no matter which choice she makes—continue indefinitely or stop at some point—her action seems irrational, or leads quickly to a state of affairs that no rational agent would accept: If she continues indefinitely she continually loses money for no gain, while if she stops she fails to act on her preferences. (85)

As Tenenbaum emphasizes, "the self-torturer has … two fairly ordinary ends (roughly avoiding pain and making money) … [that] generate a very clear (though not well-behaved) preference ordering" (83-84). More specifically, the self-torturer's preferences over the options are cyclic in that, for each pair of *adjacent* settings, the self-torturer prefers to stop at the higher setting rather than the lower setting and yet there is a sufficiently high setting n (among many sufficiently high settings) which is such that the self-torturer prefers stopping at the initial setting at which the voltage is not raised at all over stopping at setting n. Though "perfectly innocent from the point of view of instrumental rationality," the self-torturer's ends make maximizing with respect to the preferences they generate impossible (100). For every setting, there is an alternative setting that the self-torturer prefers. And yet, as Quinn suggests, and as Tenenbaum and I accept as common ground between us, we, as theorists of instrumental rationality, are being "too easy on [ourselves]" and "too hard on the self-torturer" if we simply dismiss the self-torturer's preferences as irrational (Quinn 1993: 199). Instrumental rationality must, it seems, prompt the agent to stop at an acceptable stopping point. This is an intriguing and tricky idea. How shall we understand the notion of acceptability?

[7] See (Quinn 1993).

Insofar as some stopping points are supposed to be acceptable and some are not, even though maximization is out of the question, the standard of acceptability cannot be that an option is acceptable only if there is no higher-ranked option. Tenenbaum suggests that an acceptable option is one that is "good enough" or satisfactory, but it seems like, even when maximizing is not possible, settling for an option that is "good enough" (from the agent's perspective) is misguided if, for example, options that are great (from the agent's perspective) are available.[8] Suppose, for example, that the self-torturer stops at setting 0, which she deems satisfactory, even though things would be great (from her perspective) were she to stop at setting 20 instead. This seems irrational. Why endorse the agent's stopping at a setting that qualifies (for her) as "satisfactory" (all-things-considered) when a setting that qualifies (for her) as "great" (all-things-considered) is available?

Significantly, my reasoning here draws on the distinction between *categorial subjective appraisal responses* and *relational subjective appraisal responses*.[9] Although this is not the place to delve into the distinction, the basic idea is as follows:

> [Loosely speaking,] relational subjective appraisal responses rank options in relation to one another; it is these appraisal responses that are captured by the agent's preferences … By contrast, categorial subjective appraisal responses place options in categories, such as, for example, "great" or "terrible."[10] (Andreou, in press)

Like relational subjective appraisal responses, categorial subjective appraisal responses can vary from agent to agent. A's categorial subjective appraisal responses might categorize option *x*, say, eating these pickled tomatoes, as "great" while B's categorial subjective appraisal responses categorize the option as "terrible."

It might be suggested that, even if, relative to the agent's all-things-considered evaluations, the options fall along a spectrum of vaguely bounded evaluative categories like "terrible," "bad," "satisfactory," "good," and "great," the fact that there is always an option that is preferred over any option the agent considers implies that there will always be an option that falls into a higher evaluative category than any option the agent considers, and so, like maximizing relative to the agent's preferences, seeking to settle on an option in the highest evaluative category in play is also impossible. But this does not follow. All the options in the case of the self-torturer might fall within a finite spectrum of categories

[8] Here and in the next few paragraphs, I draw on (Andreou 2015).

[9] See (Andreou, in press) and (Andreou 2015); the latter uses slightly different terminology.

[10] I here loosely describe "the favoring of one option in a pair as the ranking of that option over the other—even when no ranking of all the options is to be had because the agent's preferences are cyclic. If my use of 'ranking' seems too loose, it can be eliminated by the reader via appropriate substitutions" (Andreou, in press).

with, say, low settings falling somewhere in the ballpark of bad and/or satisfactory, mid-range settings falling somewhere in the ballpark of satisfactory, good, or great, and higher settings "circling back" through satisfactory and bad to terrible. But then, even though maximizing relative to the agent's preferences remains out of the question, settling for a satisfactory option seems rash.

I propose that we (partially) characterize (rational) acceptability as follows: An option is acceptable only if there is no higher-ranked option or, if there are no maximal options (where a maximal option is such that there is no higher-ranked option), only if the option falls squarely within the highest (all-things-considered) evaluative category in play.[11] (I here restrict my attention to cases where there is a finite number of ordered categories in play, as, tangential complications aside, we can assume is the case in the self-torturer's predicament.) Where there are no maximal options, as in the case of the self-torturer, settling on an option that is acceptable according to the preceding characterization will necessarily involve satisficing in the sense of settling on an option which is such that a higher-ranked option is available. It need not, however, involve settling on an option that is "good enough" or "satisfactory" in an intuitive sense. For instance, where there are no maximal options and the evaluative categories in play are, say, just "bad" and "terrible," ending up with a bad option will qualify as (rationally) acceptable even though it falls short of ending up with a(n) (intuitively) satisfactory option. Relatedly, where the evaluative categories in play are, say, just "satisfactory" and "good," ending up with a satisfactory (as contrasted with good) option will *not* qualify as (rationally) acceptable. Although my suggested proposal for understanding acceptability glosses over a large number of complications (which I broach elsewhere),[12] it is, I hope somewhat illuminating with respect to the intriguing but tricky idea that an instrumentally rational agent will often have to seek "acceptable" realizations of her ends rather than maximizing, and not due to bounded rationality but due to the nature of the ends themselves.

Turn finally to the idea that an agent may be rationally permitted to waver between options in a way that involves him incurring costs that he could have avoided if he resisted "brute shuffling," though not if the result is "disastrous" (156). Consider Tenenbaum's $200 WASTED case:

> Larry is deciding between being a professional footballer or a stay-at-home dad. In order to become a professional footballer, he must buy a $200 ball and net set. If he wants to be a stay-at-home dad, he needs to buy the *How to Be a Stay-at-Home Dad* DVD for $200. Larry forms the intention to become a professional footballer, goes to the store, and buys the ball and net set. Ten minutes later he abandons his intention,

[11] See (Andreou 2015).

[12] See (Andreou 2015).

calls the Barcelona manager, and says that he no longer wishes to be on the team as he is now a stay-at-home dad. (153)

Suppose this is a case in which Larry finds being a professional footballer and being a stay-at-home dad incommensurable (with the options being unrankable for Larry as one better than the other or as exactly equally good). (Like some assumptions flagged above, the assumption that two options can be incommensurable is controversial but one that Tenenbaum and I accept as common ground.) Suppose, relatedly, "that a difference of $200 dollars in the cost of either alternative would not suddenly make one of the options better than the other" (153). As such, Larry's choosing to be a stay-at-home dad would be permissible even if being a stay-at-home dad cost $200 more than originally anticipated. Still, as Tenenbaum grants, Larry's two choices (in the passage quoted above) seem collectively "foolish"—"it seems that something went awry" (154). Despite this appearance, Tenenbaum suggests that, other things equal, Larry does not count as irrational. Tenenbaum does grant that insofar as "repeated changes of mind would lead [Larry] to an unacceptable actualization of his pursuit of enough financial resources," Larry would, in a case involving such repeated changes of mind, qualify as irrational (156). Here, again, we run into the notion of an acceptable—in the sense of satisfactory or good enough—option. But, again, why settle on such acceptability? What if repeated changes of mind lead Larry from a good financial state to one that is, though not disastrous, only merely satisfactory? Hasn't something gone awry? The answer, I contend, is yes. Although satisfactory, the result is rationally unacceptable for essentially the same reason that it would be if the result were disastrous (and Tenenbaum grants that the result would be rationally unacceptable then)—the reason is that the agent failed to settle on an option in the highest evaluative category in play. Tenenbaum might be willing to grant this, maintaining that, when such failure is at issue, other things are not equal; they count as equal only when the waste in financial resources is small. But, assuming now that satisfactoriness is not enough for rational acceptability in cases of incommensurability, one might wonder why one should count as acceptable a series of choices that realize, over time, a non-maximal option. Unlike in cases involving (rationally innocent) cyclic preferences, realizing (through one's choices over time) a maximal option seems like something that a rational agent can and should aspire to in cases of incommensurability (given that, as I think both Tenenbaum and I assume, temporally extended agents are accountable [other things equal] for how their choices over time add up, and, in particular, for avoiding self-defeating patterns of choice). But then something *has* gone awry when one has proceeded in a way that is wasteful, even when one's choices over time generate only a small, rather than disastrously large, waste of resources. The widely shared intuition that Larry has made a mistake should not, I think, be abandoned.

A reader less sympathetic than myself to Tenenbaum's general approach may balk at many of the assumptions that Tenenbaum and I both accept and that I incorporate into my discussion without defense. My aim has not been to defend our shared premises but, taking them as given, to advance our general approach further by complicating and enriching debate via the consideration of subtleties that merit further consideration and exploration. According to my reasoning, proceeding acceptably over time may well involve proceeding acceptably *at* each moment, even if, as Tenenbaum maintains, "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 $t_n$" (47). Moreover, proceeding acceptably over time, though it may, even apart from considerations of bounded rationality, often involve satisficing in the sense of settling on an option which is such that a higher-ranked option is available, this need not amount to settling on an option that is "good enough" or "satisfactory" in an intuitive sense. Instead, rationality may, if there are no maximal options, require one to settle on an option in the highest available evaluative category, which may be better or worse than "satisfactory." Finally, given the availability of one or more maximal options, as in cases of incommensurability, a rational agent can and should aspire to realize (through her choices over time) a maximal option, which requires avoiding "wasteful" instances of "brute shuffling," even if the result of such brute shuffling is "good enough."

Chrisoula Andreou
Department of Philosophy, University of Utah
c.andreou@utah.edu

## References

Andreou, C., 2014, "The Good, the Bad, and the Trivial," in *Philosophical Studies* 169: 209-225.

—, 2015, "The Real Puzzle of the Self-Torturer: Uncovering A New Dimension of Instrumental Rationality," in *Canadian Journal of Philosophy* 45: 562-575.

—, (in press), *Choosing Well*, Oxford University Press, New York.

Bratman, M., 2012, “Time, Rationality and Self-Governance,” in *Philosophical Issues* 22: 73-88.

Quinn, W., 1993, “The Puzzle of the Self-Torturer,” in *Morality and Action*, Cambridge University Press, Cambridge.

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

Thompson, M., 2008, “Naïve Action Theory,” in *Life and Action*, Harvard University Press, Cambridge MA, 83-146.

# The extended theory of instrumental rationality and means-ends coherence

John Brunero

*Abstract:* In *Rational Powers in Action*, Sergio Tenenbaum sets out a new theory of instrumental rationality that departs from standard discussions of means-ends coherence in the literature on structural rationality in at least two interesting ways: it takes intentional action (as opposed to intention) to be what puts in place the relevant instrumental requirements, and it applies to both necessary and non-necessary means. I consider these two developments in more detail. On the first, I argue that Tenenbaum's theory is too narrow since there could be instrumental irrationality with respect to an intention to φ even if one is not yet engaged in any relevant intentional action. On the second, I argue against Tenenbaum's claim that "*an agent is instrumentally irrational if she knowingly fails to pursue some sufficient means to an end she is pursuing.*"

*Keywords:* instrumental rationality,  means-ends coherence, intention, intentional action, trying

In his excellent book, *Rational Powers in Action: Instrumental Rationality and Extended Agency*, Sergio Tenenbaum lays out a highly ambitious, original, and powerful theory of instrumental rationality, which he calls the "extended theory of instrumental rationality" (abbreviated "ETR").[1] The five core components of that theory are stated in Chapter 2. The first is:

> *ETR BASIC*: The basic given attitude is intentional action, more specifically, the intentional pursuit of an end. (43)

Tenenbaum notes that any theory of instrumental rationality will specify some motivationally efficacious attitude (perhaps a desire, an intention, a preference, or something similar) as its "basic given attitude." That basic given attitude will then set a "basic standard of success" for the theory of instrumental rationality (11). For instance, if *desire* is the basic given attitude, then, roughly, an instrumentally rational agent will be one who satisfies her desires. The basic given attitude isn't *itself* up for rational assessment, at least insofar as the theory

[1] All in-text parenthetical page numbers are references to Tenenbaum (2020).

Submitted: 14 November 2022
ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i1.445
Accepted: 18 November 2022

of instrumental rationality goes. But it does set the standard by which we can say that someone is successful (or unsuccessful) with regard to the exercise of their instrumental rational powers. As is clear from *ETR BASIC*, Tenenbaum takes the basic given attitude to be *intentional action*.

The second and third components of the theory are its *principles of derivation* and *principles of coherence*:

*ETR DERIVATION*: An instrumentally rational agent derives means from ends according to the following principles of derivation:

*Principle of Instrumental Reasoning (Sufficient)*
Pursuing *A*
Pursuing $B_1$ & Pursuing $B_2$, …., & Pursuing $B_n$ is a (nontrivial) sufficient means to pursuing *A*
------------------------------------------
Pursuing $B_i$ (for any *i* between 1 and *n*) (while also pursuing $B_j$ for every *j* such 1≥ *j*≥ *n* and *j*≠*i*

*Principle of Instrumental Reasoning (Contributory)*
Pursuing *A*
Pursuing $B_1$ & Pursuing $B_2$, … , & Pursuing $B_n$ is a contributory means to pursuing *A*
------------------------------------------
Pursuing $B_i$ (for any *i* between 1 and *n*). (44)

These are principles of *reasoning* to sufficient and contributory means. But they do have, in Tenenbaum's view, some consequences for the evaluation of an agent's rationality or irrationality:

> But at the very minimum we can say the following: *an agent is instrumentally irrational if she knowingly fails to pursue some sufficient means to an end she is pursuing.* (47)

The principle of coherence prohibits one from holding ends one knows cannot be jointly realized:

(3) *ETR COHERENCE*: When an instrumentally rational agent realizes that her ends are incompatible (cannot be jointly realized), she abandons at least one of the ends from the smallest subset of her ends that cannot be jointly realized. (45)

For instance, if I realize that I cannot both swim in the race and watch the soccer match, which I know is on at the same time, I'll either give up the end of swimming in the race or the end of watching the soccer match.

If we look at *ETR DERIVATION*, we see that the "basic given attitude" of *intentional action* is both a *premise* ("Pursuing A") in the principles of instrumental reasoning and a *conclusion* ("Pursuing $B_i$"). The latter feature is noted in the fourth component of the *ETR*:

(4) *ETR EXERCISE*: The exercise of instrumentally rational agency is an intentional action.

The fifth and final component simply observes that the principles of derivation and coherence in (2) and (3), and any principles that can be derived from them, "exhaust the content of the principles of instrumental rationality" (47):

(5) *ETR COMPLETE*: No other basic principles govern the exercise of our instrumentally rational powers. (45)

These are the five central tenets of the *ETR*. Tenenbaum also lists out some "auxiliary hypotheses" (47) that are important to the arguments for the theory, but we'll focus here on the central tenets.

One noteworthy feature of theory is the way in which it departs from much of the discussion of "instrumental rationality" within the literature on structural rationality. Within that literature, there is a particular focus on a requirement of means-ends coherence, which is usually formulated along the following lines:

*Means-Ends Coherence:* Rationality requires that [if one intends to *X*, believes one will *X* only if one intends to *Y*, then one intends to *Y*].[2]

If I were to intend to swim in a race tomorrow, believe that to do so I must intend to register online, but not intend to register online, I would fail to do what rationality requires of me according to Means-Ends Coherence. The brackets indicate that the requirement is a "wide-scope" requirement in that "requires" has logical scope over a conditional.[3] What Means-Ends Coherence prohibits is a certain *combination* of attitudes (broadly understood to include both the attitudes one *has* and the attitudes one *lacks*): the combination of *intending to X*, *not intending to Y*, and *believing one must intend to Y in order to X*.

Means-Ends Coherence is not the only requirement of practical rationality, and, plausibly, it's not the only requirement of *instrumental* rationality. But it's often presented as a standard example of a structural requirement of rationality. In just looking at this formulation of the requirement, however, we can see two ways in which Tenenbaum's theory is different. First, whereas the requirement of Means-Ends Coherence is put in place by an *intention to X*, Tenenbaum's theory takes *intentional action* as the basic given attitude. Second, whereas Means-Ends Coherence is concerned exclusively with means believed to be *necessary* for an end, Tenenbaum's *ETR* extends to cover both *sufficient*

[2] This is the formulation I work with (but ultimately suggest would need some refinement) in Brunero (2020). For a small sample of other claims regarding the structural irrationality of means-ends incoherence, or formulations of the rational requirement prohibiting it, see Setiya (2007: 668), Bratman (2009: 413), Broome (2013: 159, 169), Kiesewetter (2017: 15, 46-47), Lord (2018: 21), and Worsnip (2021: 3).

[3] On wide-scope requirements, see Broome (2013: Ch. 8).

and *contributory* means. For many readers, I suspect this is a breath of fresh air. We've finally arrived at a theory of instrumental rationality that is sufficiently *practical* in that intentional *action* is both the "input" and "output" of the principles of instrumental reasoning, as sketched in *ETR DERIVATION.* And we've departed from what might seem like a peculiar philosophical obsession with *necessary* means, at the expense of consideration other instrumental relations.

I, too, welcome these developments. But I want to consider these two features of the *ETR* in more detail. In particular, in §1, I consider whether we should accept *ETR BASIC.* My central worry about *ETR BASIC*, very roughly, is that the focus on intentional action is too narrow, such that many of the central cases of instrumental irrationality, including cases that would be prohibited by Means-Ends Coherence, wouldn't be covered by the theory. In §2, I consider whether we should accept the verdicts about irrationality that Tenenbaum extracts from *ETR DERIVATION.* While I think it's not all that complicated to say what rationality requires when it comes to means believed to be necessary (here, I think something along the lines of Means-Ends Coherence is correct), matters become more complicated when we transition to thinking about means believed to be sufficient. In particular, I think there are counterexamples to Tenenbaum's claim that "*an agent is instrumentally irrational if she knowingly fails to pursue some sufficient means to an end she is pursuing*" (47) and that Tenenbaum's ingenious attempts to circumvent those counterexamples will cause further difficulties for the theory.

1. Tenenbaum tells us at the start of the book that "instrumental rationality is, roughly, a relation between intentional actions" (2). This is reflected in *ETR DERIVATION*, which has intentional actions in the role of both premise and conclusion. One way to challenge the thesis that instrumental rationality is a relation between intentional actions is to challenge the Aristotelian Thesis—that is, the thesis that intentional action is the conclusion of practical reasoning. Opponents of the Aristotelian Thesis will deny that practical reasoning concludes in an (intentional) action, and will instead insist that it concludes in an intention or a practical belief or judgment, and they would reject *ETR DERIVATION* on this basis.[4] But I'm going to set aside that debate here, and instead consider the role of intentional action as a "premise" in *ETR DERIVATION*,

[4] For defenses of the Aristotelian Thesis, see Clark (2001), Tenenbaum (2007), Dancy (2014, 2018), and Fernandez (2016). My own view (which owes much to Paul 2013) is that the Aristotelian thesis is mistaken, and practical reasoning concludes in either an intention or a practical judgment (see Brunero 2021). These complicated questions have been well explored by others, and would take us too far afield, so I'll leave them aside.

and as the attitude which sets the standard of (instrumental) rational success, according to *ETR BASIC*. I'll argue that the conception of instrumental rationality as a "relation between intentional actions" is too narrow, since one can be instrumentally irrational (or rational) with respect to a future-directed intention to φ, even if one hasn't yet engaged in any (non-mental) intentional action with respect to φ-ing.

It's clear that Tenenbaum wishes to contrast his theory with those theories which take some mental state to be the "basic given attitude." He writes:

> So, it's not an intention to write a book, or a preference for writing a book over not writing a book, that determines that my, say, writing Chapter 2 of the book is an exercise of my instrumentally rational powers. Rather, the basic given attitude in this case is my *writing a book* (intentionally), or my *intentional pursuit* of writing a book (or intentionally pursuing the end of writing a book). (44)

One question to raise here is whether it's possible to intend to write a book without having the "basic given attitude" specified by *ETR*—that is, without engaging in some relevant intentional action. (If it's not, the contrast Tenenbaum is drawing between *ETR* and other "mental state" theories becomes less interesting.) But it certainly *does* seem possible.[5] Suppose I'm deliberating about whether to swim in the race tomorrow, and I decide (thereby forming an intention) to swim in the race tomorrow. I'm certainly not now *swimming in the race*. (Doing so would be grounds for disqualification, since one isn't permitted to swim in the race in advance of the starter's whistle.) But nor does it seem true that I'm engaged in the *intentional pursuit* of swimming in the race (or intentionally pursuing the end of swimming in the race). At least on a fairly natural understanding of "pursue," to pursue an end would involve, perhaps among other things, the employment of measures directed toward the realization of that end. But I haven't yet undertaken any (non-mental) actions which facilitate my end of swimming in the race. All I've done is reach a decision to swim in the meet. Once I *start* employing those measures (e.g., researching directions to the meet, packing up my swim gear), it would make sense to say that I'm engaged in an intentional *pursuit* of swimming in the race (or intentionally pursuing the end of swimming in the race.). But, for now, I'm not yet pursuing any such thing.

Additionally, Tenenbaum tells us that intentional action "is an event or process in the external world" (11). And, in a passage contrasting mental actions with bodily actions, he writes: "For the purpose of proposing and evaluating a theory of instrumental rationality, we should think of intentional actions as primarily bodily actions" (15). But it certainly seems possible for me to form an intention to φ—perhaps I reach a decision to φ after deliberation—without yet

[5] For relevant discussion, see Davidson (1978) on "pure intending."

performing any bodily actions relevant to φ-ing. The "event or process in the external world" is yet to come.

If it is possible to intend to φ without yet engaging in the intentional pursuit of φ-ing, this raises a concern about Tenenbaum's theory of instrumental rationality. Suppose I've formed an intention to swim in the race, but haven't yet taken those measures which would license our saying that I'm intentionally pursuing the end of swimming in the race. Intuitively, even at this early stage, there could be instrumental irrationality. If I intended to swim, but didn't intend to register, believing this to be necessary, I would be convicted of irrationality under Means-Ends Coherence. But if Tenenbaum's theory gets a grip only later on—once the measures needed for an intentional pursuit are undertaken—it's unclear how it can deliver this verdict.

There are some subtleties about time and rationality that I'm passing over here. First, we need to accommodate the phenomena of "rational delay."[6] The updating of attitudes is a process which takes time; it can't be done instantaneously. And so we might want to allow a "grace period" of sorts, giving the person (who intends to swim and believes intending to register is necessary) some time to form the intention to register. (It's doubtful we'll be able to specify the length of the grace period with any precision; we can only say that excessive slowness is not allowed.) Second, we need to accommodate the phenomena of "rational self-trust."[7] It may be that there's no irrationality in failing to intend to register if one rationally trusts that one will form the intention at some later point, before it's too late. Such temporal subtleties will be relevant to the project of arriving at a more precise formulation of Means-Ends Coherence. But they need not concern us here. Let's just work with an example which will allow us to set them aside. First, let's assume that I've intended to swim in the race, and believed intending to register is necessary, for quite some time. Maybe others have even pointed out to me that I have these two attitudes and they've given me plenty of time to reflect on that fact and update my attitudes, but I haven't yet done so. Issues of rational delay do not come into play here. Second, let's assume that it's obvious to all involved that a decision on registering is necessary immediately—perhaps the online registration window is about to close—and so considerations of rational self-trust won't come into play. Since I must decide now, it's not an option to trust myself to form the intention later on. But, importantly, neither of these assumptions will involve my taking measures to promote my swimming in the race. We can still have a case in which I intend to swim in the race tomorrow (and irrationally don't intend to register) but I'm not yet

[6] See Podgorski (2017).

[7] See Setiya (2007: 668).

intentionally pursuing swimming in the race. And the worry is that Tenenbaum might not have the resources to allow that the norms of instrumentality rationality get a grip this early on.

One available reply to this worry comes out of Tenenbaum's discussion of what he calls "gappy actions." Tenenbaum observes that many actions are such that we can be in the process of performing them, while not at that very moment taking steps that facilitate or promote the performance of that action (70-76). He gives the example of baking a cake. In the course of performing this action, I may engage in several other actions:

Turning the oven on
*Checking the cat*
Whipping eggs
*Listening to the radio*
Measuring flour (130; see Fig. 5.1)

The italicized actions are the "gaps" in my gappy action of baking the cake, since they are neither instrumental nor constitutive means to baking the cake. But once one allows for the possibility of gappy actions, there's no reason to disallow the "gap" from appearing in the *initial stages* of the gappy action. Perhaps we should think of my intentional pursuit of swimming in the race as a gappy action with an *initial gap*, and allow that the action begins at the moment I intend to swim in the race, but the instrumental (or constitutive) means are taken later on. Tenenbaum suggests a view along these lines in Chapter 5:

> As I see the need to paint the fence, I could get an early start by painting the first yard, the first foot, the first inch, or just by forming the intention to paint it in the near future. Forming the intention is just the limit case of early engagement in the pursuit of certain means to an end, not any different than engaging in a gappy action, except that the relevant gap is prior to the fully active parts of the action. (124)

So, with respect to our example, we could allow that one is engaging in the intentional pursuit of swimming in the race even *before* one takes any instrumental measures that promote or constitute swimming in the race. Let's call this the "initial gap strategy."

The initial gap strategy goes some ways toward solving our difficulty. But it doesn't seem to go far enough. Suppose that I initially intend to swim in the race, but I don't *ever* take any instrumental or constitutive means to doing so. In this case, it's hard to see how we can say that there's an initial gap, since there's no other surrounding actions to give definition to that gap—that is, there's nothing parallel to turning the oven on, whipping the eggs, and measuring the flour in the earlier example, which are the instrumental or constitutive means, and which set the boundaries of the gaps. The "gap" seems to no longer exist, much like the donut hole that disappears after the donut is consumed. More importantly, it doesn't seem like there's any extended gappy action of *intentionally pursuing swimming in the race* in cases in which *no* instrumental or constitutive means are taken. But this is worrisome since such cases could very well be cases in which one is instrumentally irrational. Our central example of means-ends incoherence—in which one intends to swim in the race, believes one must intend to register, but doesn't intend to register—could be a case in which no instrumental or constitutive means to swimming are ever undertaken. This case seems to me (and to many others writing about structural irrationality) to be a case of instrumental irrationality. But it's not clear to me how Tenenbaum's theory can deliver that result.

So far, I've argued that the *ETR* is too narrow: there are central cases of instrumental irrationality that would be prohibited by Means-Ends Coherence, but wouldn't be prohibited by the *ETR*. These cases involve agents who have intended to do something without yet engaging in any intentional *action* or *pursuit*. However, it's worth considering ways to extend the extended theory to cover such cases. We could revise our conception of the basic given attitude by first allowing that there could be *more than one* basic given attitude, and then state that both future-directed intentions and intentional actions count as basic given attitudes for the purposes of the theory:

> *ETR BASIC EXTENDED*: The basic given attitudes are intentional action, more specifically, the intentional pursuit of an end, *and future-directed intentions*.

The revision would have the advantage of improving extensional adequacy, in that the theory could now in principle address those cases I'm concerned about. And it seems to be a modest revision in that it wouldn't require too much tinkering with the other components of Tenenbaum's view. What changes would we need to make? If the basic given attitude is supposed to specify the *premise*s in the principles of reasoning, we may need to make the necessary changes to the two principles of reasoning in *ETR DERIVATION*. Additionally, Tenenbaum holds that the principle of coherence is to some extent a consequence of the principles of derivation (see p. 18), so we may also have to allow that *ETR COHERENCE* applies both to the intentional pursuit of ends and to future-directed intentions as well. But this should also be seen as a welcome development, since it's already widely thought that there's a rational prohibition on inconsistent future-directed intentions.[8] In short, it seems like extending the *ETR* in this proposed way would have many benefits and few costs.

[8] As Bratman observes, there's a requirement that our intentions and beliefs fit into a "consistent conception of the future." See Bratman (1981: 259).

2. As I noted earlier, Means-Ends Coherence applies only to means believed to be necessary. It would be a mistake to formulate a coherence requirement along the same lines applicable to sufficient means. Consider:

> *Mistaken Means-Ends Coherence*: Rationality requires that [if one intends to *X*, and believes that *Y*-ing is sufficient for *X*-ing, then one intends to *Y*].

Suppose I intend to donate money to some particular charitable organization, and I know there are two sufficient means to making the donation: mailing a check, and depositing an envelope with cash in the donation box. Suppose I intend to mail a check, and I don't intend to deposit the envelope. There's no irrationality here whatsoever. Yet I would be in violation of Mistaken Means-Ends Coherence: I intend to make a donation, believe depositing the envelope would suffice, but don't intend to deposit the envelope. This shows that Mistaken Means-Ends Coherence is, as its name indicates, mistaken.

Of course, this is no challenge to Tenenbaum's theory, since he doesn't endorse this view. In his view, rationality would only require, at a minimum, that one take *some* sufficient means. More precisely, his view is:

> But at the very minimum we can say the following: *an agent is instrumentally irrational if she knowingly fails to pursue some sufficient means to an end she is pursuing.* (47)[9]

When I intend to make a donation, and decide upon writing a check instead of depositing the envelope, I'm still pursuing some sufficient means, and so I don't run afoul of Tenenbaum's requirement.[10]

[9] One of the most interesting features of Tenenbaum's view, which I'm neglecting here since I won't have space to discuss it adequately, is (putting it very roughly) his suggestion that we move away from discussions of principles and rules of rationality to discussion of rational powers and virtues. As Keshav Singh (forthcoming) puts it, in a very insightful critical notice of Tenenbaum's book and my own, Tenenbaum offers us a "power-centric" rather than a "principle-centric" approach to instrumental rationality (whereas my own approach is, as Singh notes, firmly within the "principle-centric" tradition.) But, as Singh observes, Tenenbaum's criticism of the "principle-centric" approach involves pointing out how such principles won't deliver everything we want out of a theory of rationality, and we need to talk about rational virtues as well. But that doesn't mean that Tenenbaum rejects the enterprise of specifying principles (which is well-illustrated by his statement of a principle here, and also the two principles in *ETR DERIVATION*). And it's worth investigating whether the principle quoted here is correct.

[10] One question I have about Tenenbaum's theory of instrumental rationality concerns the relationship between the principles of instrumental *reasoning* in *ETR DERIVATION* and what rationality requires according to the theory. As John Broome points out in *Rationality Through Reasoning*, an agent could engage in good reasoning, but be under no requirement to do so. (For instance, the rational requirement to believe the logical consequences of what one believes—for instance, roughly, to believe q when one believes p and p→q—applies only when one cares about the relevant question (here, the question of *whether q*). But I could very well engage in good deductive reasoning about some matter I don't care about. That would be good reasoning that is not rationally required of me.) See Broome (2013: 157-159, 247). And it seems that Tenenbaum would agree with Broome's observa-

However, there might be cases where it's rationally permissible for one to knowingly fail to pursue some sufficient means to an end one is pursuing. Consider the following case:

> *Principled Patty*: Patty is the new chair of the Philosophy Department, and she is pursuing the end of *getting a hire*—in particular, she's aiming to get the Dean's permission to hire a logician. She knows that blackmailing the Dean would enable her to get a hire, but doing so runs afoul of her moral principles, and she refuses to do it. She instead pursues other means: lobbying members of the Dean's Hiring Advisory Committee, working on a detailed hiring request, trying to convince other departments of the value of having a first-rate logician at the university, and so forth. However, she is not sure these conventional means will be successful.

There's a difference between Principled Patty and my earlier case of the charitable donation. In the case of the donation, I know of two sufficient means to donating: writing a check and depositing the envelope. Patty, however, knows of only one sufficient means: blackmailing the Dean. The other, conventional means aren't thought by her to be sufficient, either individually or collectively, for getting a hire. It seems that Patty "knowingly fails to pursue some sufficient means to an end she is pursuing" yet she doesn't seem to be guilty of instrumental irrationality.

Now if we vary the case so that Patty thinks blackmailing the Dean is *both* sufficient and necessary, then there would be irrationality—at least if Means-Ends Coherence is correct. In that case, Patty would have the prohibited combination of intending to get a hire, believing that (intending to) blackmail the Dean is necessary, and not intending to blackmail him. But we're setting up the example such that she *doesn't* believe it's necessary, but *does* believe it's sufficient.

In Chapter 9, Tenenbaum mentions the possibility of a case structurally parallel to Principled Patty:

> However, in some cases, there are no sufficient means that I know will achieve my end, but I do not abandon the end. I try means that will likely, or at least possibly, achieve my end. So, for instance, I might realize that I know of no sufficient means to achieve my

tion: after all, while there is a rational requirement corresponding to the Principle of Instrumental Reasoning (Sufficient)—the requirement to take some sufficient means—he doesn't specify any requirement corresponding to the Principle of Instrumental Reasoning (Contributory). So, he seems to acknowledge the possibility that one could engage in good instrumental reasoning according to that principle without being under any rational requirement to do so. But that raises the question of what explains why there is an associated rational requirement when there is one. Why, for instance, does the Principle of Instrumental Reasoning (Sufficient) generate a requirement to take some sufficient means, but the Principle of Instrumental Reasoning (Contributory) generate no similar requirement? It's not clear to me what the answer would be. I'll set this question aside and focus instead on Tenenbaum's view about what rationality requires when it comes to means believed to be sufficient.

> end of earning a million dollars (or that the only means that I know will achieve this end, defrauding my great-uncle, is not a means I am willing to take), but that there are some actions I could perform that would have a good chance of achieving the end (becoming a lawyer) or that could at least make it possible (buying a lottery ticket). (209)

In this passage, he's primarily concerned with cases in which the agent knows of no sufficient means to achieve his end, but the suggestion in the parenthetical remark is that we could treat cases like Principled Patty (and more generally, cases in which the only sufficient means is "not a means I am willing to take") in the same way.

Tenenbaum's ingenious suggestion at this point is that in such cases, the agent's action is more accurately described as *trying to φ* rather than *φ-ing*, where *trying to φ* is an "essentially different action" from *φ-ing.* (210)[11] He makes the point with a different example, in which the bullies are trying to prevent the nerds from crossing the street. Tenenbaum, as one of the nerds (in the example), writes: "In such a case, it would seem that I would more naturally describe my action by saying, 'I am trying to cross the street,' rather than 'I am crossing the street.' (209-210)." And then the suggestion is that the same could be said in cases in which one in unwilling to take some sufficient means. Using our example, we could say that Principled Patty's end isn't *getting a hire*, but *trying to get a hire*—or, at least, she should revise her ends so that *trying to get a hire* is her end. As Tenenbaum puts it:

> We can now say that the agent who realizes that she cannot, or is not willing, to pursue means she knows to be sufficient for her end of φ-ing must revise her ends, and among the possible acts still available to her will be the act of trying to φ. (210)

But now note that if Patty's end is *trying to get a hire,* she does indeed take some sufficient means to her end. The conventional means (lobbying the Hiring Advisory Committee, etc.) do indeed suffice for *trying* to get a hire. (They aren't sufficient for *getting* the hire, but are sufficient for trying to do so.) And thus Tenenbaum could deliver the verdict that Principled Patty is indeed instrumentally rational—she's pursuing some sufficient means to her end of trying to get a hire—thereby avoiding the objection entirely.

Tenenbaum's suggestion here is that we need to change what gets put into the "A" in the schema of Principle of Instrumental Reasoning (Sufficient), where the starting premise is "Pursuing A" and "A" is a variable for agential ends. We should have Patty's end be "trying to get a hire" and then it's easy enough to

[11] As he notes in a footnote on p. 209, there a sense in which the first sentence of the previously quoted passage isn't entirely accurate: "I would now be pursuing a different action, so in some sense I did abandon the end."

maintain that Patty is indeed pursuing some sufficient means to her end, and is thus rational. I want to raise four concerns about this strategy in the remainder of this section.

My first concern is that this seems to distort Patty's practical reasoning. The "trying" is now presented as the *object* of Patty's pursuit, since we now have "Pursuing *trying to get a hire*" as the first premise in Patty's instrumental reasoning. But Patty herself would likely reject that characterization of her practical reasoning. She would likely say that *what* she is pursuing is the end of *getting a hire*, not a trying. Her trying is something that occurs *while* she is intentionally pursuing the end of getting a hire; it's not the *object* of that pursuit. The object, as she sees it, is getting a hire. Patty also knows, like the rest of us, that we aren't always successful in our pursuits.

Here's another way to think about this concern. In aiming to articulate her practical reasoning, Patty certainly wouldn't have the first premise of her reasoning be "I am pursuing the pursuit of a hire" or "I am trying to try for a hire." Such premises involve confusing redundancies, and it's not at all clear what these sentences mean. It would be much more natural for her to simply say "I'm pursuing getting a hire" or "I'm trying to get a hire." But I'm not sure that "I am pursuing trying to get a hire" is all that much better. (Just as it seems odd to say that *what* is being pursued is a pursuit, and *what* is being tried is a trying, it seems odd, though perhaps not to the same degree, to say that *what* is being pursued is a trying.) It would be much more straightforward to have "I am pursuing getting a hire" as the first premise in her reasoning, while acknowledging that this pursuit *also* involves Patty's trying to get a hire and that she knows she may or may not succeed in doing what she is trying to do.

My second concern is about how redescribing Patty's end as a trying would interact with *ETR COHERENCE*. According to that principle, "when an instrumentally rational agent realizes that her ends are incompatible (cannot be jointly realized), she abandons at least one of the ends from the smallest subset of her ends that cannot be jointly realized" (45). For instance, when I realize that I cannot both finish this paper tonight and prepare adequately for tomorrow's class, I will, if I'm instrumentally rational, give up at least one of the two ends. But I might realize these two *ends* cannot be jointly realized without thinking that the associated *tryings* cannot be jointly realized. After all, in this example, I know full well that I could give both ends my best shot and fail spectacularly at one or perhaps even both. In light of this point, the general concern is that when we redescribe φ-ings as tryings, we'll render *ETR COHERENCE* inapplicable to cases in which it should be applicable.

Let's apply this point to Patty's case in particular. In Patty's Department, the chair is automatically on the hiring committee, as of the very moment the hire is approved. While Patty knows full well that she can't both *get a hire* and *not be on a hiring committee*—and so *ETR COHERENCE* would prohibit her from pursuing both ends—she doesn't believe (because it's not true) that she can't both *try to get a hire* and *not be on a hiring committee*. (These ends *are* jointly realizable, and she knows it.) And so we would need some other explanation of why she's rationally prohibited from also intending to avoid being on a hiring committee. *ETR COHERENCE* would no longer be able to deliver this result.

My third concern is more of a dialectal one. In order for this strategy to get around the original objection, it has to be the case that Patty is pursuing the end of trying to get a hire and *not also* pursuing the end of getting a hire. It's not enough to note that there's *some* description of Patty's end (the one involving trying) that has it come out that she's taking sufficient means to her end. The original problem was that there's another description of Patty's end (the one involving intentional action) that has it come out that she's failing to take some sufficient means, and the *ETR* would then declare Patty to be instrumentally irrational. To avoid that, we have to *disallow* "getting a hire" as a correct description of what Patty is doing. But this seems to be a tall order. Let's suppose that Patty succeeds in getting a hire. A third-person observer (perhaps Patty herself at a later time) might reasonably describe the instrumental means Patty undertook (lobbying the Hiring Advisory Committee, writing the detailed hiring requests, etc.) as components of the extended action of getting a hire, much like one might reasonably describe, in Tenenbaum's example, the instrumental means he took (turning the oven on, mixing the eggs, etc.) as components of extended action of baking a cake. Of course, such an observer might very well also mention a trying, but they likely wouldn't do so *at the expense of* describing the extended action; they would likely say that Patty was *both* trying to get a hire and succeeding—that is, *getting a hire*. But, as we noted above, we have to disallow "getting a hire" as a correct description. That seems to be a significant cost.

My fourth concern is a normative one. Tenenbaum thinks that the agent who is "*not willing* to pursue means she knows to be sufficient for her end of φ-ing must revise her ends, and among the possible acts still available to her will be the act of trying to φ" (210, emphasis added). This helps with Principled Patty, since we can then say that in taking the conventional means (lobbying the Hiring Advisory Committee, etc.) she is indeed taking sufficient means to her end of *trying to get a hire*, and so is rational. It gets Patty off the hook as far as the charge of irrationality goes. But do we want to allow that a *mere unwillingness* to pursue means known to be sufficient can let one off the hook in this way? Consider:

*Phobic Patty*: Patty is the new chair of the Philosophy Department, and she is pursuing the end of *getting a hire*—in particular, she's aiming to get the Dean's permission to hire a logician. Matty is the new chair of the Mathematics Department, whose first (and last) proposal as chair is to give up one of his department's faculty lines to Philosophy, so that they can hire a logician. All Patty needs to do is walk from Philosophy Hall to Mathematics Hall and pick up the paperwork. But Patty has an intense phobia of Mathematics Hall, and refuses to walk over there and get the paperwork, even though she knows this will suffice for getting a hire. She instead pursues other means: lobbying members of the Dean's Hiring Advisory Committee, working on a detailed hiring request, trying to convince other departments of the value of having a first-rate logician at the university, and so forth. However, she is not sure these conventional means will be successful.

Whereas Principled Patty's unwillingness is based on good moral reasons, as is Sergio's unwillingness to defraud his great-uncle, Phobic Patty's unwillingness is based on an irrational fear of Mathematics Hall. But since both are equally *unwilling* to take some means they know to be sufficient, and are pursuing other conventional means to getting a hire, it seems that Tenenbaum's theory would treat the cases alike: if Principled Patty gets off the hook, Phobic Patty does as well. But that seems to be a bad result. We want it to come out that Phobic Patty is instrumentally irrational.[12]

What the pair of examples suggests is that it can't be that an agent's *mere unwillingness* to take some sufficient means to getting a hire lets us instead construe the relevant end as *trying to get a hire* and then see the conventional means as sufficient for the trying (thereby removing the instrumental irrationality). Rather, she must have *good reasons* for being unwilling. Principled Patty has good reasons while Phobic Patty does not. This raises a further question of what it takes to have good reasons for refusing to take some means known to be sufficient. That might be a difficult question to answer. But there's no principled reason for thinking that a theory of instrumental rationality couldn't provide an answer to that question. But note that in providing such an answer, the theory would not be simply applying *ETR DERIVATION* or *ETR COHERENCE*, but would be engaging in a substantive normative inquiry about reasons.[13] In any case, my main point here is that we need to find some grounds

[12] If the phobia is not Patty's fault, we may not want to *blame* her for her irrationality. But it's clear that her phobia is interfering with her rationality, and, specifically, making her instrumentally irrational with respect to her end of getting a hire.

[13] Moving in such a direction way may require that we revise *ETR COMPLETE*, which takes these two principles to be the only basic principles in our theory of instrumental rationality. Or, alternatively, it could be seen as a supplement to the two principles that helps us understand how they are applied.

for letting Principled Patty off the hook that don't extend so far as to let Phobic Patty off the hook as well.

Let's sum up the argument of this section of the paper. I've focused on Tenenbaum's claim about rationality and *sufficient means*:

> But at the very minimum we can say the following: *an agent is instrumentally irrational if she knowingly fails to pursue some sufficient means to an end she is pursuing.* (47)

I've argued that Principled Patty is a counterexample, since she is not instrumentally irrational in knowingly failing to pursue the known sufficient means of blackmailing the Dean. I've then considered a reply suggested by Tenenbaum's remarks in Chapter 9—namely, that Patty (if she's rational) only has the end of *trying* to get a hire and she does take some sufficient means to that end. And I've raised four concerns about this reply: (1) it distorts the first premise of Patty's instrumental reasoning in having *trying* as the *object of her pursuit*; (2) it makes it unclear how we can apply *ETR COHERENCE* with respect to the new end (the trying, as opposed to the φ-ing); (3) it requires that we reject as false any third-personal report which has *getting the hire* as the relevant extended action (perhaps alongside *trying to get the hire*); and, (4) it proves too much in also letting Phobic Patty, who is also unwilling to take some sufficient means, off the hook as well.

3. In this paper, I've focused on two components of Tenenbaum's *ETR* that will be exciting and interesting to those steeped in the structural rationality literature, where Means-Ends Coherence has been a standard requirement of instrumental rationality. First, whereas Means-Ends Coherence is a requirement governing *intentions*—specifically, a requirement forbidding one from intending to *X*, believing intending to *Y* is necessary for *X*-ing, and not intending to *Y*—Tenenbaum says that "instrumental rationality is, roughly, a relation between *intentional actions*" (2, emphasis added), and the principles of reasoning in *ETR DERIVATION* are formulated to reflect that ("Pursuing A," Pursuing $B_1$," etc.). I've here avoided discussion of the contentious question of the conclusion of practical reasoning—specifically, of whether the Aristotelian Thesis is correct—and focused instead on the "premises" or inputs—specifically, on the idea that intentional actions, not intentions, put in place the requirements of instrumental rationality. I've argued that there's a cost to accepting the *ETR*, since many standard cases of instrumental irrationality, covered by Means-Ends Coherence, wouldn't be covered by the *ETR*. And I've argued that Tenenbaum's attempt, in Chapter 5, to remedy this difficulty by appealing to "gappy actions" with a gap at the start won't do enough to resolve the worry.

Second, whereas Means-Ends Coherence is concerned exclusively with means believed to be necessary, Tenenbaum's *ETR* is concerned with means believed to be sufficient. My suspicion is that Means-Ends Coherence has enjoyed a certain popularity in the rationality literature in part because it seems easier to say what rationality requires when it comes to means believed to be necessary, and matters become trickier when it comes to non-necessary means. And if the argument in the previous section is correct, that suspicion is confirmed to some extent. I've focused in particular on Tenenbaum's claim that instrumental rationality requires that one not knowingly fail to pursue some sufficient means to an end she is pursuing. I've presented a counterexample to that claim (Principled Patty) and argued that Tenenbaum's strategy for dealing with such cases, suggested by his remarks in Chapter 9, will generate further difficulties for his theory.[14]

John Brunero
Department of Philosophy, University of Nebraska-Lincoln
jbrunero2@unl.edu

## *References*

Bratman, M., 1981, "Intention and Means-End Reasoning," in *Philosophical Review* 90(2): 252-265.

—, 2009, "Intention, Practical Rationality, and Self-Governance," in *Ethics* 119(3): 411-443.

Broome, J., 2013, *Rationality Through Reasoning*, Wiley Blackwell, Oxford.

Brunero, J., 2020, *Instrumental Rationality: The Normativity of Means-Ends Coherence,* Oxford University Press, Oxford.

—, 2021, "The Conclusion of Practical Reasoning," in *The Journal of Ethics*, 25: 13–37.

Clark, P., 2001, "The Act as Conclusion," in *Canadian Journal of Philosophy* 31(4): 481–505.

Dancy, J., 2014, "From Thought to Action," in R. Shafer-Landau, ed., *Oxford Studies in Metaethics*, Vol. 9, Oxford University Press, Oxford.

Dancy, J., 2018, *Practical Shape: A Theory of Practical Reasoning*, Oxford University Press, Oxford.

Davidson, D., 1978, "Intending," in *Philosophy of History and Action* 11: 46-60.

Fernandez, P., 2016, "Practical Reasoning: Where The Action Is," in *Ethics* 126 (4): 869-900.

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

[14] Many thanks to the graduate students in my Fall 2021 Ethical Theory seminar for helpful discussion of Tenenbaum's book.

Lord, E., 2018, *The Importance of Being Rational*, Oxford University Press, Oxford.

Paul, S., 2013, "The Conclusion of Practical Reasoning: The Shadow Between Idea and Act," in *Canadian Journal of Philosophy* 43(3): 287-302.

Podgorski, A., 2017, "Rational Delay," in *Philosophers' Imprint* 17(5): 1-19.

Setiya, K., 2007, "Cognitivism about Instrumental Reason," in *Ethics*, 117(4): 649-673.

Singh, K., in press, "New Work for a Theory of Instrumental Rationality," in *Analysis.* https://doi.org/10.1093/analys/anac020

Tenenbaum, S., 2007, "The Conclusion of Practical Reason," in S. Tenenbaum, ed., *New Trends in Philosophy: Moral Psychology*, Rodopi, Amsterdam, 323-343.

Tenenbaum, S., 2020, *Rational Powers in Action: Instrumental Rationality and Extended Agency*, Oxford University Press, Oxford.

Worsnip, A., 2021, *Fitting Things Together: Coherence and the Demands of Structural Rationality*, Oxford University Press, Oxford.

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

Erasmus Mayr[1]

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(2) Whenever I ask myself, during the course of the summer, whether I should start or continue reading, my current preferences at that moment and my other ends may speak sufficiently strongly against reading some pages 'just now' that it is rational for me not to start (or continue) reading then. E.g.,

[5] See also (Mayr 2022).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

First, he could accept that the principle of instrumental reasoning cannot be action-guiding after all for the pursuits of long-term, indeterminate ends, at least not with regard to 'Positive Contribution,' if these pursuits share the features (1) to (3) presented at the beginning of sub-section 1.1. The principle might still be action-guiding in other contexts and for the pursuits of long-term, indeterminate ends in other respects (e.g., when it comes to ruling out courses of action which would make reaching the end impossible). But with regard to 'Positive Contribution,' it would merely be an evaluative standard.

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

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

## 2. *Principles for imperfect agents*

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

### 2.1. Dealing with the risk of our own future irrationality

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

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

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

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

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

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

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

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

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

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

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

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

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

### 2.2. Uncertainty of success

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

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

But it seems hard to accept the consequence Tenenbaum draws from this, namely that we can draw no inference as to the instrumental rationality of an agent who, on learning that the envisaged means may fail to lead to the aim, (and who can therefore no longer decide to reach this aim, but only decide to try to do so) does not (even) try to reach this aim. "Suppose I was on my way to meet Mary at her office, and I now realize that Mary might not be in her office. (…) Nothing about my basic given attitudes here determines whether I will, insofar as I am rational, engage in the action of trying to meet Mary at her office" (RPA: 210).

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

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

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

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

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

## *References*

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

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

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

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

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

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

# Tenenbaum on instrumental reason and the end of procrastination

Matthias Haase

*Abstract*: In *Rational Powers in Action*, Sergio Tenenbaum argues that instrumental rationality is constitutively rationality in action. According to his theory, we not only reason *to* action, we also reason *from* action: both the major premise and the conclusion of instrumental reasoning are intentional actions in progress. In the paper, I raise four challenges: (a) The view rests on the assumption of a symmetry between the starting point and the conclusion of instrumental reasoning. But in the cases of telic actions like building a house, proper reasoning concludes with the completion of the action. (b) Tenenbaum conceives of the nexus between ends and means in terms of the relation between a temporally extended whole and its parts. This fails to do justice to the distinction between movement and conduct. (c) The theory suggests that it is instrumentally irrational to abandon all particular ends. But it is hard to see why this should be so. (d) Tenenbaum holds that his theory of instrumental rationality can explain why procrastination is a vice. Yet the argument seems to rest on a simplification of the phenomenon.

*Keywords:* action, activity, conduct, instrumental reason, pleasure, prudence

## 1. *The rational and the real*

In *Rational Powers in Action*, Sergio Tenenbaum proposes to turn the received theory of instrumental reason from the head to the feet. The prevailing conceptual framework puts the spotlight on ordering preferences, forming intentions, and modeling plans. Their realization in action appears to be another matter. Strictly speaking, the work of reason seems confined to the inner recess of the mind while leaving all the rest to the forces of nature. According to Tenenbaum, the prevailing view not only fails to explain the rationality of action; it also rests on distorted picture of our ends and purposes. Properly conceived, the action in the external world is the first thing to consider rather than the last: "Instrumental rationality is rationality *in action*."[1] Its proper work is the realization of ends; and its home office, so to speak, isn't the inner realm furnished

[1] Tenenbaum 2020, viii. In what follows cited as RPA.

Submitted: 22 November 2022
ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i1.447
Accepted: 29 January 2023

with a set of conative states: the taking of means proceeds from the temporally extended action of pursuing the end. The book develops the position through devastating critique of various alternatives on offer in the literature. Here, I will focus on the positive proposal.

The final line of the treatise reads: "If all went well, this book has helped us to see that, at least when practical reason is flawlessly exercised, the real is the rational and the rational is the real." As Tenenbaum is well aware, the allusion to Hegel's infamous formula may come as a surprise at the end of a book devoted to instrumental reason. That is not quite what Hegel had in mind: he was talking about the actuality of the good in ethical life. My question in what follows will be whether the claimed unity of thought and action can be understood within the confines of a theory that leaves open whether any of it is actually good.

It all turns, of course, on what is meant by "action." Consider the opening paragraph of Christine Korsgaard's *Self-Constitution*:

> Human beings are *condemned* to choice and action. Maybe you think you can avoid it, by resolutely standing still, refusing to act, refusing to move. But it's no use, for that will be something you have chosen to do, and then you will have acted after all. Choosing not to act makes not acting a kind of action, makes it something that you do. (2009, 1)

Going by this line, it is the human plight to act. But where choosing to refrain from it is already a case of it, one might ask in light of what it all counts as action. In the course of Korsgaard's investigation, it turns out to be the great old question of how to live, in face of which we can't but act. Accordingly, the sense of agency is essentially ethical. Tenenbaum, by contrast, investigates the power to realize ends, whatever they might be. A theory of instrumental rationality puts no restriction on their content, apart from requirements for successful realization. That is what he calls the Toleration Constraint. (RPA, 20) This suggests that, *as* far as instrumental reason is concerned, Hume was right when he pronounced: "'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger." (Hume 1978, 416) So, how could it be instrumentally irrational to prefer always postponing to finishing this paper? As befits the topic, I shall leave the latter question for the end. I will begin with an outline of Tenenbaum's approach to instrumental reason and then turn to his account of action.

## 2. *The extended theory of instrumental rationality*

As Tenenbaum conceives it, a theory of instrumental rationality must contain the following elements: an account of the "input", an account the "output", and an account the principles connecting the two. The input is what the subject reasons *from*: her "basic given attitudes." The output is what the subject reasons *to*: her "practical exercises." The task of a theory of instrumental rationality is accordingly to articulate the "principles governing the exercises of a (finite) rational agent's active powers in light of her given attitudes." (RPA, 17) The account Tenenbaum proposes is radically parsimonious. His theory only needs a sole principle for the articulation of the rational connection and one category for the representation of the conative elements so connected.

The sole principle of instrumental reason is the *principle of derivation* according to which an instrumentally rational agent takes sufficient means to her ends. The articulation in terms of sufficient means allows deriving the *principle of coherence*, which excludes the pursuit of incompatible ends. (RPA, 45) After all, pursuing ends that can't be jointly realized makes it impossible to take sufficient means to one's ends. Accordingly, an instrumentally rational subject is efficacious and coherent. What the power thus specified governs is the relation between intentional actions: doing something for the sake of something else one is doing. It is a familiar Aristotelian doctrine that the conclusion of practical reasoning is action. According Tenenbaum, the same holds for its starting point: we not only reason *to* action, we also reason *from* action. The "basic given attitudes" figuring as conative input are intentional actions of pursuing ends; the corresponding "practical exercises" figuring as output are intentional actions of taking means. The relation is of course mediated by the agent's cognitive conception of the means-end connection: the minor premise of the instrumental syllogism.[2] But "both the conclusion and major premise are intentional actions." (RPA, 44)

The argument for the thesis proceeds through a critique of the received views. One of Tenenbaum's central objections is that these positions fail to account for the rationality of action, since they present the reasoning as stopping short of the doing and issuing instead in choices, decisions, or intentions that stand in causal relations to movement or change in the external world. Properly conceived, the reasoning "reaches all the way down to, for instance, the movements of one's limbs." (RPA, 16) That the reasoning must also be taken as descending *from* intentional movement is said to follow from the impossibility of assessing rationality in light of what would be available in a momentary snapshot of subject's conative attitudes. (RPA, 50) According to classical decision theory, the rationality of the output is supposed to be determined by reference to the *fully determinate* desires or preferences that the subject has *at a moment*. But in the pursuit of projects that take time to complete, our ends are usually indeterminate because (a) their content is vague, (b) our initial conception doesn't

[2] For the most part of book, Tenenbaum treats the minor premise as expressing knowledge; the task of the final chapter is to show how the theory can accommodate conditions of uncertainty and risk.

rule out all inacceptable realizations, and (c) the relevant degree of perfection isn't settled in advance. The *realized* end is determinate; but the determination takes place in the course of the realization: in the process of reasoning out the means in reaction to the challenges arising along the way and in coordination with one's other ends. For this reason, the "given attitudes" figuring as conative "input" must be conceived as the temporally extended pursuit of ends. Hence the name of the position: the Extended Theory of Instrumental Rationality.

## 3. *The symmetry thesis*

It is central to Tenenbaum's teaching that instrumental reasoning not only concludes in action but also begins with action: anything apt to provide a starting point for instrumental reason must belong to the same category as the conclusion. This thesis of a formal symmetry between input and output is something that the position shares with the standard view where both appear as conative mental states. That is not how Aristotle seems to present the practical syllogism. He says that the conclusion is an action; but he doesn't make an analogous remark about the mayor premise. In fact, he seems quite concerned to stress that the rational source of movement isn't always another movement.[3] To the untutored mind, it would at any rate appear that we aren't always in the midst of motion. So, a central task for Tenenbaum's approach is to explain how we are to understand the concept of intentional action such that everything fits into this mold.

Going by Tenenbaum's introduction of the term, an intentional action is "an event or process in the external world." (RPA, 12) For the purposes of the treatise, mental actions are set aside. The official paradigm is "bodily action." (RPA, 15) But Tenenbaum works with a specific conception of what that amounts to. For beings like us, realizing an end usually takes time and involves taking several steps. Consider building a house, writing a book, or training for a marathon. Such things aren't done in a day. The action is temporally extended and divides into phases. Where this is so, it is usually also possible to truly predicate the respective action concept "φ" in a judgment that exhibits what is sometimes called the *broad progressive* where the truth of "*S* is φ-ing" is compatible with *S* currently not making any progress in her φ-ing. (RPA, 71) A person can be truly described as being in the process of building a house, even though she is currently sitting of a sofa taking a nap or having a sandwich. Tenenbaum describes those phases as "gaps" in the overarching action – as opposed to its "fully active

[3] The premises appear under the heading of the good and the possible. (See *De Motu Animalium*, 701a23-24; *Nicomachean Ethics*, 1147a29-32.) Of course, the good figures as the object of pursuit. But not all pursuit falls into the category of movement. (See *Nicomachean Ethics*, 1139b1-4.) I will come back to the latter point in the next section.

parts" where the agent *is currently* making progress by taking concrete means to her end. Accordingly, he calls temporally extended actions, which include such inactive phases, "gappy" actions.

The notion of "gaps" in the fully active engagement makes space for the idea of the coordinated pursuit of multiple ends: say, scheduling the training sessions for the marathon in such a way that there is still enough time in the day to also make some progress on the book and the house one is working on. However, the concept of "gappy" action is also supposed to provide the conceptual resources to conceive of *any* end or purpose – any "basic given attitude" providing the conative "input" for instrumental reason – as physical action. What about future-directed intentions where there isn't anything yet that the agent is or was doing *actively*? Tenenbaum holds that forming an intention can be treated as a limiting case of the engagement in a temporally extended action; it is just that the "gap" is at the beginning – "prior to the fully active parts of the action." (RPA, 124) However, the proposal seems incompatible with the initial introduction of the term where an intentional action appears as a countable particular: an event or process in the external world. The latter idea also seems contained in the official definition of "gappy" action:

> [W]e can call a (token) action *A* 'gappy', if it extends through an interval of time $t_0$-$t_n$ such that at some intervals contained in the $t_0$-$t_n$ interval, the agent is not doing anything that is a (constitutive or instrumental) means to *A*.[4]

The talk of a token-action suggests a countable individual in the external world. But how is that particular to be individuated where there is only a "gap" without any active parts around it? In a footnote Tenenbaum suggests that Helen Steward's account of intentional actions as processes would be congenial to his approach. (RPA, 12, Fn 29) According to her view, however, processes are modally robust individuals that are individuated by reference to the spatio-temporal location of their "initial segments."[5] When one spells out the proposed treatment of future-directed intentions through Steward's definition, the so-called "(token) action" will end up as an item that has its original home in the mind or at least somewhere within the inner limits of the agent's body.[6]

[4] RPA, 71. The passage is meant as a preliminary definition or "first approximation." But the further complications introduced by the final definition make no difference for the present considerations.

[5] See Steward 2013, 807. Tenenbaum refers to an earlier paper where Steward doesn't articulate the criteria of individuation. But the early paper already contains the claim that processes are spatio-temporally located individuals.

[6] When Michael Thompson argues that intention is, metaphysically speaking, on a par with action in progress, he insists that the progressive is "general" and reserves the introduction of "a genuine particular" for the perfective. (See Thompson 2008, 137.) This might provide an alternative way of ensuring the symmetry between mayor premise and conclusion. By the same token, however, the

On the face of it, the trouble with pure intending as potential input is related to a corresponding difficulty on the side of the output. The aim is to provide an account of the rational realization of ends in the material world. But in the case of finite ends like building a house, this would appear to introduce a categorial asymmetry between starting point and conclusion. At least that is what the philosopher of common sense suggests when presenting the pure form of the technical syllogism in his *Logic*.[7] As Hegel has it, the *intended end* is general and subjective, while the *realized end* is particular and objective. The transition is the *taking of the means*: the action in progress or the reasoning as rational realization unfolding in time. The reasoning concludes in the completed action: the doing folded into a fully determinate particular. In the case of the example at hand, quite literally a thing: the house that was built. Or so Hegel suggests. Tenenbaum, by contrast, avoids the puzzle about the transition from mind to world by situating already the *intended end* in objectivity. But why should the reasoning be described as reaching down to movement, if movement already figures as its given starting point?

The idea would be absurd, if we were meant to take the talk of a token-action as signifying the *fully determined* particular that stands at the end: the done deed where everything has been settled. The action figuring as input is meant to be a *determinable* that gets determined through the execution of the project. However, the same should hold for the action figuring as output. After all, the theory represents both by action sentences in the progressive. Tenenbaum suggests that the "active parts" of the overarching action can't themselves all be "gappy": there must be some basic actions that *only* have active parts. (RPA, 72) Still, qua being in progress they must be conceived as determinable rather than fully determined. As long as the finish line lies still ahead, it isn't all settled yet, and something might interfere. So conceived, the reasoning seems to stops short of the realized end. Completion or success appears to fall outside its scope. Whatever explains the transition from the *determinable* to the *fully determined*, it doesn't appear to be the rational realization. And what holds for the overarching process, should equally apply to any step along the way: its completion will also lie beyond the scope of the rational realization. But unless the completion of some of the phases can be understood through the reasoning, it is unclear how the respective process can count as a progressing physical ac-

question would arise whether the reasoning so conceived reaches all the way to the respective particular that is under the relevant descriptions the completed action or realized end. For the discussion of the analogous point about practical knowledge see Haase 2018.

[7] In his *Science of Logic*, Hegel treats the instrumental syllogism in the chapter "Teleology" abstracting from the idea of the good, which is introduced a later in the book. For the asymmetry thesis see esp. Hegel 2010, 12.169.

tion in the perspective of the reasoning. So, how can we say that the rational is the real and the real is the rational?

## 4. *The monolithic conception of action*

Another question to ask of an account of rational realization is whether it can do justice to the variety of the things we pursue. G.E.M. Anscombe once complained that "modern philosophy of the Anglo-American tradition" is guilty of "a great fault": she called it "the monolithic conception of desire, or wanting, or will." (2005, 154) One might also describe it as the view that one pro-attitude operator will do for all intents and purposes. A possible motive for seeking such uniform representation is the commitment to the program of decision theory where anything that plays a role in rational choice must fit into the slot of preferences to be compared and weighted. Tenenbaum adamantly refuses this program, inter alia on the ground that the ends we pursue are non-comparative. At the same time, he also seems concerned to ensure that anything figuring as input for instrumental reason fits one category: "[A]ny kind of policy, project, long-term action, and so forth can be understood [...] as a continuous (though "gappy") action." (RPA, 126) Accordingly, it seems sensible to wonder whether Tenenbaum endorses what one might call a monolithic conception of action; and if so, whether that is a mistake.

The difficulty is that the notion of action was originally introduced through reflection on the pursuit of finite ends like building a house or writing a book. While all individual doing arguably stops at some point, such telic action verbs specify what it is for the action so described to end on its own terms: by reaching completion instead of being interrupted. This isn't always so. Some ends are infinite in that they don't define a terminus or stopping point to be reached through the act of realizing. Take the end Tenenbaum discusses under the heading of the policy of faithfulness. (RPA, 133) Traditional marriage vows tend to mention a natural stopping point, but death doesn't enter the formula as the target state to be brought about by the having and the holding. It wouldn't be in the spirit of the vows either to take them by analogy to holding a weight until the gym trainer calls time. The difference hasn't escaped Tenenbaum's notice. But he treats it in a certain way.

The contrast between finite and infinite ends is initially introduced by way of the following example: "Unlike the end of running a marathon, singing has an internal structure that never fails to give purpose to one's life." (RPA, 56) Of course, singing a song usually goes by more quickly than a marathon. What Tenenbaum has in mind is a distinction between two kinds of terms: telic action verbs like "to run a marathon" and activity verbs like "to run" or "to go for a run." In the present tense deployment, the former describe action on the way to completion. The latter, by contrast, don't specify a terminus internal to the act and accordingly represent action going on indefinitely. So, when the bare terms figure in the place of the object of current pursuit, the corresponding act of realizing will be directed in the one case at the *completion* and in the other at the *maintenance* of the respective action. As Tenenbaum has it, both kinds of verbs can figure in the description of "gappy" action. After all, it can be true that you are running a marathon or going for a run, even though right at this moment you are standing still to have a drink or to take a phone call. In each case, your instrumental rationality will be assessed with the view to how you manage the "gaps" and the "active parts" in relation to the respective temporal profile of the overarching action of which they are phases. (RPA, 127) Stopping for drinks and phone calls all the time tends to undermine the aim of reaching the finish line or, for that matter, the goal of maintaining a run.

Formally speaking, this is supposed to be all that is needed. Any kind of end, Tenenbaum suggests, can be accommodated in this framework when one notes "a continuum of indefinite length of gaps between fully active parts." (RPA, 126) Personal policies are said to have "the same structure as activities" insofar as "the constitutive and instrumental means for the end of the activity are for its continuation or perseverance, not for its completion." (RPA, 127) Once one conceives of "policies [as] instances of long-term gappy actions", the account developed in reflection on the management of the "gaps" and "active parts" in "'mundane' long term actions like baking a cake" can be "extend[ed] to a policy such as 'exercising regularly'." (RPA, 130) In the latter case, the instrumental rationality of the agent is a matter of whether the relevant interval contains sufficient "active parts" for maintaining the policy. (RPA, 131) In the case of a relaxed exercising regimen, it is only required to exercise often enough. Other personal policies, like the traditional take on faithfulness, are *strict* rather than *loose* in that don't allow the occasional night outside. (RPA, 133)

With these details in view, one might say that the proposed conception of action isn't uniform but rather binary: "An instrumentally rational agent engages in the fully active parts […] for the sake of […] the acceptable *completion or maintenance* of the larger action." (RPA, 74, my italics) Note, however, that the disjunction appears within one structure. What does the work for the account of the instrumental reason is the idea of "engag[ing] in the fully active parts […] *for the sake of* […] the larger action." (RPA, 74, original italics) In the resulting picture, the coordinated pursuit of multiple ends involves distinguishing various levels and keeping track of intricately nested action descriptions. But the instrumental nexus is always couched in terms of the relation between an overarching action and its active phases. In this respect, the conception seems monolithic. The question is whether that is problematic.

Among the purposes to be accommodated in the framework is "the end of engaging in the enjoyment of pleasant activities." (RPA, 70) This can mean that I'm seeking something pleasant to do or that I'm aiming at making time for activities I know to be enjoyable. Yet the basic case is surely the one where I'm taking pleasure in what I'm currently doing. Say, I'm eating gummy bears. Having another one is a means to maintaining this pleasant activity. But in what sense does what is so maintained appear in the perspective of the maintaining as a "larger action" for the sake of which I'm engaged in a present "active part"? Conceiving of writing this sentence as a "part" of writing a paper goes together with understanding my present act in relation to the steps I have taken up to this point. In the gummy bear scenario, I may be aware that I have been doing it for a while. Yet those gummy bears I ate during that time are nothing to my current pursuit. Sitting in my tummy they will eventually become an impediment to keeping going. But until they do that possibility might not enter my mind. My only concern is to keep the supply constant. Experience might teach me to take a more structured approach. In turn, you may find me in a restaurant having a five-course meal. Now, there is a "larger action" with "fully active parts", but here it also holds that I'm engaging in the former for the sake of engaging in the latter. And if it all works out, the times between the courses aren't "gaps" in my enjoyment of this pleasant activity. Provided abundance of resources, addiction extends this structure to infinity. While the occasional smoker chooses to have a cigarette thinking that it will be enjoyable, the true smoker relishes all day in gapless pleasure spending the times between the smoking sessions in joyful anticipation. And yet smoking doesn't thereby figure as a purpose of life.

With the view to the artful management of addiction or lured by Kant's alleged remarks about the positive effects of tobacco consumption on contemplation, a person might also adopt a smoking-policy. But that seems like a different kettle of fish. As Tenenbaum uses the term, it covers a wide range: from plans or "intermediate policies" adopted in the pursuit of finite ends like training for a marathon all the way up to such things as the "policy of loyalty." (RPA, 189, 133) The proposed account is meant to cover all of them: "[A]ny kind of policy [...] can be understood [...] as a continuous (though "gappy") action." (RPA, 126) Initially, the line isn't put forward as a thesis about the "metaphysics of actions or policies"; it is said to only express the claim that "from the point of view of the theory of rationality, there are no differences between actions and policies."[8]

[8] In the dialectic of the book, intentions for the future, plans, and policies come up, because philosopher like Michael Bratman suggest that they are related to additional principles of "diachronic rationality." (Bratman 2018) Tenenbaum argues that there is no need for such further principles, once one realizes that ordinary action already involves managing one's agency extended in time. But that still leaves the question how to fit those items into his theory.

Yet it is not clear how there can be a division of labor here. 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. A few lines further down Tenenbaum indeed presents it as a consequence of his theory that "policies [...] must be regarded as ordinary actions." This suggests that, metaphysically speaking, they are to be counted as such. But how do we count them?

With respect to ordinary actions like crossing the street or going for a run, one can ask: "Are you are *still* doing it, or are you are doing it *again*?" Suppose an hour ago I saw you moving across. When I look up now, there you are a little further up the street doing the same as before. So, I wonder. Your answer will have consequences for the list I'm keeping on how many runs you go on per week. On the assumption that personal policies are to be situated on a continuum with activities, it would seem to follow that I could make an analogous list counting your policies of the year. Take a loose exercising regimen or a strict drinking policy. In both cases, the question may arise whether you are *still* on track or *again*, after having fallen off the wagon. And yet when you do the same as before, it makes no sense to ask whether it is the same one. During a year one can lose a habit and acquire it afresh; it doesn't follow that there are two within that interval. It is the same here. In the little book I'm keeping on you, a personal policy is something that you do, but it doesn't fit the category of token-action extended through an interval.[9] The nexus of realization is not a relation of "larger action" to its "active parts"; it is rather akin to the relation between your general conduct in a certain area and its manifestation on a particular occasion.

On reflection, it sounds strange to describe the execution of a policy in terms of the management of "gaps" and "active parts." Say, you have a policy to break up fights. It would seem that if your policy is *loose* rather than *strict*, then it can also be on active service when you choose to let those two go on with their brawl. And if your policy is *strict*, then it should be at work in any social situation to assess whether some fight is going on. Going by Tenenbaum's account, a "fully active part" would be the breaking up of a given fight. But what if no one around you is fighting? Does maintaining your policy require you to get yourself into situations where people are fighting or, if you can't find any, arrange for people to have a fight? That can't be right. The description in terms of engaging in a "fully active part" for the sake of maintaining the existence of a "larger action" seems to introduce the wrong kind of connection.

[9] On the face of it, the alternative presented in the book doesn't exhaust the philosophical options. Bratman originally introduced his notion of "personal policies" as an enrichment of the furniture of the agent's mind: it's not just beliefs and desires, as the standard story would have it. (See Bratman 1989) Tenenbaum insists that policies aren't mental states but rather token-actions. However, one can deny that policies are items in the mind without thereby affirming that they are on a continuum with going for a run.

There is a yet another purpose that figures in the theory: "the pursuit of happiness, or the pursuit of a good life." Tenenbaum stresses that this is "an end that the agent pursues." (RPA, 47) According to the theory, the pursuit of an end is an intentional action. It follows that the pursuit of happiness must be an action. So, one would want to know of what kind. Lenny Kravitz sings about it in terms of motion: "My mama said that love's all that matters. But I'm always on the run." (Kravitz, 1991) Still, he isn't literally talking about going for a run. So perhaps it is an instance of very long-term gappy action. Yet where are the gaps? Sleeping better not be a hiatus in the practice of living well. The relevant sense of agency seems to fit neither of Tenenbaum's two categories: activity verbs and telic action verbs. Living well is arguably not like running as if it could go on forever; nor does leading a good life appear to be analogous to running a marathon. Aristotle does say that in choice one's whole life is at stake. But he doesn't mean that one should make all choices with the view to the bucket list. He excludes the children from choice and *praxis*. And yet he wouldn't deny that a little one might resolve to always run away when father comes home.

## 5. *Preferring not to*

As a power of reflection, practical reason puts us in the position to step back from any particular purpose or, for that matter, from all of them. A few pages after coining his famous formula about the rational and the real, Hegel presents this possibility as a distinguishing mark of human agency: by contrast to a brute animal, a human being can "abandon all things" and "renounce any activity of life, any end."[10] Take the writer from Melville's story of Wall Street: Bartleby, the scrivener. Towards any determinate course of action that comes up as option, Bartleby eventually adopts the stance expressed by his infamous formula: "I would prefer not to." Ultimately, he abandons all ends and renounces of any activity of life. Korsgaard would of course insist that even Bartleby can't escape the human plight to act: from the perspective of her theory, the scene appears as self-constitution done badly. Yet she also holds that the notion of agency can't be understood within the confines of a theory of instrumental reason. So, what is to be said about the scenario from the standpoint of Tenenbaum's teaching about the flawless exercise of the latter power?

It is part of the theory that instrumental rationality can require giving up some ends. According to the *principle of coherence*, the following holds: "When an instrumentally rational agent realizes that her ends are incompatible (can-

[10] The famous formula appears close to the end of the Preface of his *Philosophy of Right*; the above remark is from §5 of the Introduction. See Hegel 1991, 20 and 38.

not be jointly realized), she abandons at least one of the ends from the smallest subset of her ends that cannot be jointly realized." (RPA, 45) But the principle doesn't tell us how to choose between incompatible ends. An arbitrary choice between *A* and *B* can be instrumentally required, if pursuing either the one or the other serves a further purpose. Yet it can't be presupposed that the subject always has a further end for which this is true. In his introduction of the notion of "basic given attitudes", Tenenbaum says that they provide the "standard of success" and are not themselves "subject to direct evaluation in the theory of instrumental rationality." (RPA, 11) But the *toleration constraint* doesn't quite hold for the subject the theory about – at least not when one takes "given attitudes" to the particular purposes the subject might find herself pursuing. By Tenenbaum's own lights, the status of given attitudes changes when the idea of their totality enters the scene. In forming such conception, the pursuing subject distinguishes herself from each of them: from the particular objects of her pursuits or the contents of her will. From the standpoint of such reflective stance, any one of them appears as something that is potentially to be renounced or abandoned when it turns out that they hinder each other. The question is what a theory of instrumental rationality can say about how to proceed from here.

According to Hegel, instrumental reason reaches at this point an impasse that it cannot move beyond by its own resources. Going by the notes of his students, he pronounced in his lectures that arbitrarily "putting oneself in only one of them setting all the others aside" would mean to give up the standpoint of reflection and thus to "relinquish [one's] universality, which is the system of all drives." Yet "the idea of forming a hierarchy to which the understanding (*Verstand*) usually resorts," the possibly apocryphal quotation continues, "is equally unhelpful since no criterion for ordering is available here so that the demand tends to run out in tedious general platitudes."[11] Taken by itself, instrumental rationality can't provide much guidance once we leave the idealized scenario where the philosopher assumes for the purposes of presentation that the only concern on the agent's mind is how to get a cover or, for that matter, how to maximize gains in the stock exchange. When the "sum total of satisfaction" is at stake, there is nothing for the "calculating understanding" (*berechnende Verstand*) to compute. On an admittedly flatfooted reading of Melville's story on Wall Street, Bartleby might be described as the unsettling embodiment of that impasse, situated fittingly right in the heart of what is arguably the original home of decision theory. In the cool hour of reflection, one must admit that the material reflected upon doesn't contain a standard for comparing. Accordingly, there is no rational ground for affirmation and denial –

[11] The line is from the Addition to §17 of the *Philosophy of Right.* I amended the translation. See Hegel 1991, 50.

pursuit and avoidance. From the logical point of view, the only way to maintain the stance of rational reflection instead giving oneself over to arbitrary particularity is to politely decline each invitation or demand: "I would prefer not to."[12]

It is an intricate question how Tenenbaum's theory stands to the Hegelian verdict on the limitations of instrumental reason, taken by itself. On the one hand, the arguments against the familiar story about maximization seem analogous: our ordinary ends are non-comparative and the appeal to strength of desire ultimately depends on normative hedonism so that it runs afoul of the *toleration constraint.* (RPA, 62) On the other hand, Tenenbaum accepts the challenge to show that his own theory can provide an account of rational ordering in the pursuit of multiple ends. Roughly speaking, the proposal is this. Tenenbaum introduces the following auxiliary hypothesis: our ordinary ends have an "internal structure" that allows the distinction between better or worse actualizations. (RPA, 47) Given the hypothesis, he argues, the theory can "generate preference orderings out of its basic non-comparative, non-graded attitudes." (RPA, 54) In this way, the theory is meant to incorporate the insights of decision theory and in effect supply an account of the rational standards guiding the revision of incompatible project with the view to the coordinate pursuit of the totality of one's ends. For the present purposes, the crucial question is whether the conceptual framework can provide a cure for Bartleby's ailment and show that "when practical reason is flawlessly exercised, the real is the rational and the rational is the real." (RPA, 229)

Note that the *toleration constraint* would seem to exclude not only normative hedonism but just as much its denial. By the same token, it cannot be ruled out either that from the point of view of the subject all that modifying and revising comes at a cost. After all, the ensuing work of coordinating and scheduling may seem like a nuisance. Considering this, the subject might arrive at the reflective preference not to engage in any of that. As Tenenbaum has it, forming a conception of the totality of one's *particular ends* goes together with the introduction of what he presents as a *general end*: "the pursuit of happiness." (RPA, 47) As I argued above, such pursuit doesn't seem to fit Tenenbaum's category of temporally extended (though "gappy") action. In fact, it has been disputed that the definition of the term introduces a link to the concept of physical action. A person might take it as a substantive question whether happiness is to be sought in living an active life (by taking means to particular ends) or rather in reaching a state of blissful inactivity (by freeing oneself from such worldly ambitions).

[12] In his reading, Gilles Deleuze brings out this character of Bartleby's formula: it expresses neither acceptance nor refusal – not even a preference, just a "non-preferred." (Deleuze 1998, 71). Of course, Deleuze would resolutely refuse a Hegelian framing. Going by his terms, "Bartleby is not the patient, but the doctor of a sick America." (90)

The space for that question would appear to be opened by Tenenbaum's own observation that the pursuit of happiness can also give rise to further ends. He calls them *general means* such as "wealth, health, and the cultivation of our skills and talents." (RPA, 47) But skills and talents are not only things that an instrumentally rational subject might come to regard as in need of cultivation. Observing that their ends or purposes tend to hinder each other, the instrumental reasoner might devise a more radical solution than the mere adjustment and coordination of their given drives. Sometimes the rational thing to do is to look for other things to pursue. And if you can't get no satisfaction, why keep trying in that way? Going by the *toleration constraint*, it looks as if the theory will also have to allow for the subject to adopt "general means" of the following kind: aims like avoiding the frustration of one's will by interfering forces, the disappointment of facing the meager fruits of one's labors, or the dread of noticing that the only point of completing the task at hand appears to consist in providing the resources for engaging in the next project of the same kind. In light of such reflective attitudes, abandoning or renouncing all particular ends would appear as an instrumentally rational conclusion. The purest version of this posture of mind would arguably consist in maintaining the general stance of reflection by insisting like Bartleby: "I am not particular." (Melville 2002, 30)

One might try saying that this is one of the shapes that the unity of the rational and the real might take. However, this would be tantamount to giving up on the thesis that instrumental rationality is rationality *in action*, at least in the sense suggested by the line that the reasoning reaches all the way down to the movements of one's limbs. Resolutely standing still or refusing to move are of course intentional actions in the relevant sense. But such endeavors will be among the projects that those reflective attitudes would recommend to resolutely renounce. If instrumental reason is exercised here, its work will be entirely within the inner limits of the agent's body. To hold on to the official line about the rational and the real, it would have to be denied that instrumental rationality is flawlessly exercised in that scenario. But this seems to infringe on the *toleration constraint.*

## 6. *Instrumental virtue and the end of procrastination*

The debate about what is to be expected from a theory of instrumental reason is at the same time a dispute about which topics properly belong to ethics. Tenenbaum contrasts "instrumental practical rationality" and "substantive practical rationality." (RPA, 23) The former is concerned with the rational realization of ends, whatever they happen to be. The office of the latter is the determination of what is good to pursue. As Tenenbaum has it, these are "two separate powers" whose perfections are "prudence" and "[practical] wisdom" respectivey. One of the marks of their separation is that "a purely instrumentally rational agent" is conceptually conceivable. Even in our case, they can come apart in two ways. It is not just that the evil and the shameless may be clever; the good or practically wise might fail to be prudent: "Lack of prudence is one of these obstacles that stepmotherly nature can put between the good-willed agent and the object of her will." (RPA, 23) Cleverness or prudence is the same excellence of mind whether it operates in evil or in good people. The task of a theory of instrumental reason is to provide a general account of "what the prudent agent knows." (RPA, 24)

Presented in this way, the definition of the proper scope of instrumental rationality puts at the same time a limitation on the reach of substantive practical rationality. Aristotle would beg to differ. On his view, practical wisdom is a kind of knowledge that one only has insofar as one does act well. Another aspect of this disagreement comes out in a later chapter where Tenenbaum argues that courage and resoluteness are to be treated as "instrumental virtues" that the shameless might manifest as well. (RPA, 169) So conceived, defining courage doesn't require venturing into ethics; it belongs to the office of the theory of instrumental rationality. By way of illustration, Tenenbaum discusses a character called Shifter: someone who abandons their end whenever danger arises. Doing so is in line with the principles of derivation and coherence. Nevertheless, Shifter is said to exhibit instrumental irrationality insofar as they lack the proper disposition of the will:

> An ideally rational agent not only takes means that are available to her will in pursuing her ends, but her power to pursue ends is also not restricted by the internal shortcomings of her own will. In other words, cowardice undermines the agent's powers to bring about ends not necessarily by leading the agent into incoherence in the pursuit of certain ends, but by simply restricting the ends that are available to the agent. (RPA, 180)

The same verdict should apply to the reflective attitudes of avoidance considered in the last section. After all, they certainly present a restriction to the ends available to the agent. So, either the doctrine solves the Bartleby conundrum, or it runs into the same problem. In an earlier passage, Tenenbaum seems to admit that the agent's concern with "the ends she *might* have" presupposes that "her continued rational agency is among her ends." (RPA, 41) But if the theory was to assume that the latter purpose mustn't be abandoned, suicidal tendencies would also have to be ruled as instrumental vices, not to mention preferring the destruction of the world to scratching one's finger.[13]

[13] This is not what Tenenbaum seems to have mind, for he allows that courage may be exhibited

In the respective chapter, Tenenbaum presents an argument that appeals to Kant. Here, the "constitutive" character of instrumental virtues gets derived from the thesis that instrumental reason is "inextricably connected in the successful and paradigmatic case with the power to pursue good ends." In effect, the verdict of irrationality is grounded in the diagnoses of a "restriction to the general power to pursue the good." (RPA, 181) So conceived, one couldn't talk about instrumental virtue in connection with the idea of a merely instrumental creature. The Kantian derivation appears to presuppose the metaphysical impossibility of such a kind of being. Moreover, the relevant conception of the good couldn't be left in the abstract, for that wouldn't provide an inextricable connection between those two powers in the successful and paradigmatic case. This looks like an ambitious program that would require venturing into ethics. In the book, Tenenbaum appeals to it only for the purposes of elucidation; the notion of instrumental virtue is not meant to depend on it. But it is hard to see how the teaching could be developed from the reflection on prudence or cleverness, considered on its own.

Kant himself seems to express skepticism about the latter kind of project when he discusses the distinction between "imperatives of skill" and "imperatives of prudence" in the *Groundwork*. The former are said to be *problematic*, insofar as they concern *possible* purposes: ends that one might or might not pursue, like building a house. The latter, by contrast, are *assertoric*, since happiness is an end that all human beings *actually* pursue by natural necessity. One might think that imperatives of prudence therefore present action as necessary. Kant denies this on the ground that it is impossible for us to determine by principle what would make us truly happy. So, it all comes down to "empirical counsels." In this connection, Kant mentions frugality and reserve; but he doesn't appear to think of them as requirements of rationality, for he stresses that they "are to be taken as counsels (*consilia*) rather than as commands (*praecepta*) of reason." (Kant 1997, 4:418) The same should hold for courage and resoluteness insofar as they are considered from the standpoint of prudence, taken by itself.

It seems worth mentioning another remark Kant makes in this connection. He observes that "in early youth it is not known what ends might occur to us in the course life." For this reason, "parents seek above all to have their children learn *a great many things* and to provide for *skill* in the use of means to all sorts of *discretionary* ends." Kant connects the observation with a complaint about the common neglect of teaching the little ones "the worth of things that they might make their ends." (4:415) Arguably, this is ultimately for them to decide.

by jumping into a shark infested pool to retrieve a five-dollar bill – provided that the person has "a fetish for five-dollar bills or […] no reflective preferences between seriously risking their lives and marginally adding to their wealth." (RPA, 186, Fn 35)

But whatever they end up doing with their lives, it will appear as a restriction or limitation in light of the infinite possibilities of what they might have become or could have done. That is what it means to lead a life: with any choice one determines oneself and limits oneself such that one will eventually be judged not by one's potential like a child but rather by one's actuality. Considered in abstraction, the idea of the irrationality of restricting the ends available to oneself would be analogous to the wish to remain forever young. Leaving aside that growing up among human beings tends to create a great impediment for the possible end of running with the wolves, this looks like another guise of the impasse Hegel was talking about.

When one steps back from all particular purposes, one's will appear as general or universal: as infinite potentiality. In Hegel's dialectic, this appears as the merely "negative notion of freedom": the reflective retreat from any determination. (Hegel 1991, §5) Of course, Hegel deems this is hopeless confusion: "A will [. . .] that wills only the abstract universal, wills nothing and is therefore no will at all. In order to be a will, [it] must restrict itself in some way or other." (§6) Unless one pursues particular ends, one doesn't realize oneself as agent. This is the impasse, put in abstract terms. Moving beyond it requires, according to Hegel, thinking the unity of the general and the particular in the singular: "self-determination" or "concrete freedom." (§7) That is what he complained Kant failed to achieve. Another name for it is the formula about the identity the rational and the real in ethical life. By the same token, what is there for us to know in matters of prudence figures in Hegel's system as something can't be separated from the standards of ethical life: it is part of practical wisdom.

One of Tenenbaum's central cases for an independent account of prudence is the treatment of the vice of procrastination. According to a famous argument by Korsgaard, it would be impossible to violate the principle of instrumental reason, if it was the only principle of practical reason. For, any action that would be a candidate for a violation of the principle to take means to one's end introduces another end for which the agent *is* taking means. Accordingly, one could always say that they changed their mind. (Korsgaard 1997) Tenenbaum argues that procrastination provides a counterexample. Say, I am pursuing the end of writing a paper for a book symposium. Writing sentences is the characteristic way of taking means. Then my usual tendencies set in: I keep fiddling with the introduction while looking around for passages to quote. According to Tenenbaum, I would be instrumentally irrational, if I consequently failed to produce during the relevant interval sufficient "active parts" for my "gappy" action to be completed in time. One might try to defend my sanity by saying that I must have changed my mind before it was too late and abandoned the end of finishing the paper. But this, Tenenbaum argues, wouldn't save me from the charge of irrationality, since it implies that I was taking means without pursuing the end. (RPA, 202)

It seems to me that the argument rests on two assumptions that are disputable. The first is that procrastination doesn't introduce its own propose. Often procrastinating is a means to the end of writing: it provides the leisure to come up with ideas. In that case, missing the deadline may be due to the cognitive mistake of losing track of time. But procrastination can also be purposive in other ways. It might, for instance, be a manner of venting anger about there being a deadline or a way of manifesting one's freedom from the task, proving to oneself that one isn't a scheduling machine. It can also be a way of holding on to the infinite potential of one's work in progress instead of eventually facing the meager reality of one's final product. The second assumption underlying the argument is that, despite being indeterminate in many other respects, my pursuit of writing a paper was from the beginning fully determinate in the following respect: it is all about the product. But one might engage in working on a paper not just for the sake of its completion, but also with the view to maintaining an activity that seems worthwhile: because it provides an occasion for learning from a wonderful book, because it is enjoyable, or simple because it gives one a task. By the same token, it wouldn't be instrumentally irrational to keep going without aiming to finish in time.

Matthias Haase
University of Chicago
haase@uchicago.edu

## *References*

Anscombe, G.E.M., 2005, "Practical Truth," in M. Geach and L. Gormally, eds., *Human Life, Action and Ethics: Essay by G.E.M. Anscombe*, Imprint Academic, Exeter.

Aristotle, 2000, *Nicomachean Ethics*, tr. by R. Crisp, Cambridge University Press, Cambridge.

Aristotle, 1978, *De Motu Animalium*, trans. by M.C. Nussbaum, Princeton University Press, Princeton.

Bratman, M., 1989, "Intention and Personal Policies", in *Philosophical Perspectives*, 3: 443-469.

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

Deleuze, G., "Bartleby; or, The Formula," in *Essays Critical and Clinical*, tr. by D.W. Smith and M.A. Greco, Verso, London, 68-90.

Haase, M., 2018, "Knowing What I Have Done," in *Manuscrito* 41(4): 195-253.

Hegel, G.W.F., 2010, *Science of Logic*, trans. by George Di Giovanni, Cambridge University Press, Cambridge.

—, 1991, *Elements of the Philosophy of Right*, A.W. Wood, ed., trans. by H.B. Nisbet, Cambridge University Press, Cambridge.

Hume, D., 1978, *A Treatise of Human Nature*, ed. by P.H. Nidditsch, Oxford University Press, Oxford.

Kant, I., 1997, *Groundwork of the Metaphysics of Morals*, tr. by M. Gregor. Cambridge University Press, Cambridge.

Korsgaard, C., 1997, "The Normativity of Instrumental Reason," in G. Cullity, G. and Gaut, B., eds., *Ethics and Practical Reason*, Oxford University Press, Oxford, 215-254.

—, 2009, *Self-Constitution*, Harvard University Press, Cambridge MA.

Kravitz, L., 1991, "Always on the Run," in *Mama Said*, Virgin.

Melville, H., 2002, "Bartleby, the Scrivener: A Story on Wall Street" in Dan McCall, ed, *Melville's Short Novels*, Norton & Company, New York, 3-34.

Steward, H., 2013, "Processes, Continuants, and Individuals", in *Mind* 122: 781-812.

Tenenbaum, S., 2020, *Rational Powers in Action: Instrumental Rationality and Extended Agency*, Oxford University Press, Oxford.

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

# 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