Dao (2017) 16:145–163
DOI 10.1007/s11712-017-9544-z

Interpretational Paradox, Implicit Normativity, and Human Nature: Revisiting Weakness of Will from a Perspective of Comparative Philosophy

Yujian Zheng[1]

Published online: 5 April 2017

Abstract This essay critiques or engages a wide range of existing works on the ancient and well-contested issue of weakness of will, from a new perspective of comparative philosophy combined with a focus on a largely neglected Davidsonian paradox of irrationality. It aims at revealing the interplay between the descriptive and the normative in the very notion of critical interpretation, as well as a special relation between holding-true and making-true which helps to explain the non-accidentalness of the descriptive coat of the Plato Principle and some of the Mencian paradigmatic tenets. By the same token, it also sheds light on some holistic picture about a certain implicit type of dynamic normativity, which seems evidently applicable to, for example, the Mencius-Xunzi 荀子 dispute on human nature, but scarcely noticed or articulated in contemporary contexts of comparative philosophy.

Keywords Weakness of will · Paradox of irrationality · Critical interpretation · Implicit normativity

1 Practical versus Theoretical Problem Concerning Weakness of will

Weakness of will has been one of the most frequently discussed problems in the West since Socrates. Socrates' denial of its very possibility (in *Protagoras*) rests upon the idea that it is not "in human nature" to pursue freely what one takes to be the lesser good. To make the denial, a basic principle, called (by Davidson 1982: 294) the "Plato Principle," states that "no one willingly acts counter to what he knows to be best." Can this principle really hold in the face of so many apparently incontinent acts in our daily life? One might argue that it would be too strong or unrealistic to believe or sanction

✉ Yujian Zheng
zhengyj@ln.edu.hk

[1] Department of Philosophy, Lingnan University, Tuen Mun, Hong Kong

such a principle in practical domain, which characteristically involves various motivational factors likely involving causal forces of unruly nature.

John Searle, for one, wants to dismiss the classic model of (practical) rationality, contemporarily represented by Davidson's causal theory of action, by asserting the existence of an essential causal gap between the psychological antecedents of an action and its performance. Fraser and Wong applied this Searlean critique and his notion of the Background in the context of understanding the nature of the Chinese problem of "character weakness" and its distinctive solution (Fraser and Wong 2008). In his largely positive reply to Fraser and Wong, however, Searle expressed some disagreement with their characterization of the Western-Chinese distinction in dealing with weakness of will as a difference between the theoretical problem of explaining irrational action and the practical problem of overcoming it. "The reason I think this is misleading is that of course the problem of overcoming weakness of will is very much part of the traditional moral education in Western culture, as it is in other cultures" (Searle 2008: 335).

Without disputing the truth of this claim itself, I would like to point out, though, that looking at a larger comparative picture, the prevailing philosophical orientation in the West (especially its contemporary analytic tradition) is much more toward theoretical explanation (in modern time, typically against the paradigmatic scientific background), which not only tends to proliferate technical sophistication but also, characteristically or inevitably, leads to problematizing certain traditional notions or issues within a naturalistic framework informed by science. Take Searle's own notion of the "gap," for example, and ask a layman question: "Why is there such an *inescapable* gap between my deliberation and my resulting decision or intention?" One can hardly make sense of the idea here unless one becomes familiar with enough parts of the contemporary landscape of analytic philosophy, which, for example, involves a research agenda or program of relating (or even reducing) our familiar normative notions such as reasons, rules, or content to physical or naturalistic notions. In this respect, the above Searlean rejection of the so-called classic model of rationality is nothing but some in-house dispute over a certain specific thesis (here, the Davidsonian causal sufficiency thesis about one's best judgment for subsequent action; see Davidson 1980b) in an overall shared framework of explanation.

It is no wonder, therefore, that one may easily identify salient differences between traditional (East or West) approaches to weakness of will and the science-inspired naturalistic ones such as those relying on causal-theoretical notions of reason and action. Identifying such differences is no intent of this essay.[1] Neither will I attempt to assess the overall merits or demerits of each type of approach. Rather, what I want to explore here is a certain largely hidden dimension of normativity underlying both the theoretical and practical approaches to weakness of will—"hidden" in the sense of being not in any explicit form of an imperative or directive at the level of moral education or practical guidance. I will show that Confucianism, as a main representative of Chinese intellectual tradition, not only does not lack sensitivity to such a special normative dimension, but also distinctively made use of it long ago, despite the fact that the explicit, clear-eyed distinction between the descriptive and the normative is

[1] I do not suggest that identifying such salient (and perhaps deep-cutting) differences in some cogent way is not important for comparative philosophy. On the contrary, no fruitful in-depth comparative (and constructive) engagement will be possible unless such differences are properly understood.

relatively new in modern Western philosophy. My exploration will be mediated by a careful discussion of a paradox of irrationality broached by Davidson.

2 Implicit Forms of Normativity Manifested in some Representative Views of Weakness of will

Everybody agrees that there is something odd about any free action against one's own best judgment, should there be such akratic actions. Most people would regard such actions as irrational, at least in a narrow, subjective sense (i.e., regardless of whether one's best judgment is objectively best). Regarding something as irrational embodies an explicit, evaluative form of normativity, and perhaps also some further normative intention to correct it. So judgment about (ir)rationality is already beyond the domain of pure description and partially enters into the normative domain. The intriguing thing concerning alleged cases of akratic actions, however, is that the data to be judged, or the explananda, are by no means neutral or "theory-free." That is, for any given alleged case of akrasia, there seems always to be room for redescription or interpretation in light of certain background theory or principles of rationality.

Let me, following Huang (Huang 2008: 439–440), enumerate the following six types of behavior which are different from strict akrasia yet might resemble or get confused with it in one way or another: (1) recklessness or intemperance, (2) compulsion, (3) hypocrisy (in which case one does what she insincerely says she ought not to), (4) ignorance (e.g., a smoker does not know or really believe that smoking is bad), (5) negligence or forgetfulness (despite the fact that the person ought to know, or normally knows, what she is doing and whether she ought to do it), and (6) rational choice (e.g., one knows perfectly well the long-term bad effect or risk of smoking as opposed to its immediate pleasure or excitement, and clear-headedly chooses the latter). This list is not exhaustive.[2]

The point of enumerating them (and possible others) is to show that it is not easy to identify, and thus prove, the existence of a pure case of strictly akratic action as long as one has some doubt about, or a hard time in distinguishing, the alleged suspect case of akrasia from all these partially resembling cases of different phenomena. It is no wonder, therefore, that "many ancient philosophers argue that not only is there no weak-willed person in this sense, but weakness of the will is simply impossible. Any case considered as weakness of will is nothing but one of the above six other phenomena in disguise. Thus, whoever thinks that weakness of the will is possible (or even actual) has to bear the burden of proof to show how it is possible" (Huang 2008: 440).

Regardless of whether the burden of proof should be (entirely) borne by whoever affirms rather than denies weakness of will at the phenomenal level, one thing seems certain: directly claiming or dismissing its possibility by appeal to our common sense (as if everyone should share a belief in such a subtle or technical matter of conceptual distinction) is not satisfactory. For that matter, Searle could hardly be entitled to claim that Davidson as well as many others who follow him, let alone those who followed Aristotle or Socrates over a long time, have obviously betrayed our common sense.

[2] For example, I discussed one special type of quasi-rational (or quasi-irrational) behavior which shares a certain underlying pattern of causal mechanisms with weakness of will (see Zheng 2006).

Moreover, to be fair to works of Davidson on weakness of will, one should acknowledge that they exhibit nuanced sensitivity, as well as proper level of effort, to the need of reconciling both description of phenomenal data and explanatory coherence within a more unified theory of action and rationality.

For my present purpose of revealing fashions of normativity at some deep or hidden level, it seems crucial to note and emphasize this single most important fact: ruling in or out weakness of will, against the backdrop of an almost continuous spectrum of familiar and unfamiliar behaviors with overlapping or resembling features, is in itself a normative move or call, with or without explicitly theoretical or practical motives. I will demonstrate this point with a careful discussion of a Davidsonian paradox of irrationality, and its possible connection to a hidden dimension of the Confucian-Mencian way of thinking in the next two sections. Before I do that, I would like to give some brief comments on Jiang and Huang, with an eye to showing that their theses also manifest, though not explicitly make, the same normative point in their own distinctive ways (Jiang 2000; Huang 2008).

First, Jiang's paper is admirably clear in locating the theoretical problem about the possibility of weakness of will in an assumption about a necessary knowing-desiring-doing connection involved in any intentional action. Only for those who take the connection for granted, the actual or possible occurrence of intentional action against one's own best knowledge at the moment, which violates the connection, becomes puzzling and therefore needs some special explanation. Jiang then plausibly argues that both Aristotle and the Cheng-Zhu 程朱 School have similar answers to this challenge, which both take as legitimate since both subscribe to this assumption. Their strategy of responding to the challenge is "to show that the weak-willed agent does not really know what is best at the moment of incontinent action" (Jiang 2000: 242).

Of course, along this line, a modern philosopher can further distinguish taking the nature of the necessary connection as broadly logical from taking it as causal, the latter being something Searle would vehemently deny given his notion of the causal "gap." It would be inappropriate, or even anachronic, to demand either Aristotle or the Cheng-Zhu scholars to make such a distinction, though in a way it actually stimulates the contemporary fuss or resurgence of interests for weakness of will. Without this distinction in the hands of ancient scholars, however, we still have good reasons to view their commitment to the necessary knowing-desiring-doing connection as normative[3]—probably as an implicit form of normativity whose theoretical status is not explicitly brought to conscious reflection.

Second, Huang's subtle and insightful comparison of the neo-Confucian CHENG Yi 程頤 (1033–1107) and Socrates and Aristotle regarding the relationship between knowledge and action invokes an important distinction between two types of knowledge, namely, knowledge of/as virtue versus knowledge from seeing and hearing. The former type of knowledge is "internal" in a dual sense: unlike the latter type which results from external contact, it comes from inner experience such as sudden enlightenment, thus called "self-getting" (Huang 2008: 446); it also has the logically internal (i.e., entailing) relation to the action that follows it (Huang 2008: 444). My brief comment here is this. *If* this knowledge distinction exists objectively, that is, independently of a normatively assumed principle stipulating the internal relationship between knowledge and action—

[3] Here my use of "normative connection" is not equivalent to logical connection in some narrow sense. One difference is that the former, not the latter, is compatible with the so-called causal gap.

a principle that would conceptually exclude the possibility of akrasia without needing empirical evidence, then such a distinction can surely play the explanatory role in sorting out those superficially akratic behaviors from strictly akratic ones, and in justifying the impossibility of the latter by appeal to the existence of the knowledge of/as virtue. This is a big "if" and I do not want to rush a judgment at this point.

Even if the knowledge of/as virtue as an ideal type (of knowing-how, say) is well-established, or the normative principle of such knowledge and its implied action is well-accepted, it is still not equal to saying that anything short of such character-based knowledge of/as virtue at any moment of real life is no practical judgment at all (against which judgment an alleged akratic action can be measured); neither is it equal to demonstrating that such knowledge of/as virtue can be normal, or normally expected of ones who are, to varying degrees, on their way toward this virtuous state of knowledge. Furthermore, in matters of nonmoral or non-character-related nature, allegedly free action against one's best judgment (were it to exist at all) would seem only to take some minimal epistemic or rational virtue, which has little to do with any particular action or practice, for arriving at the subjective yet clear-eyed best judgment in question. In other words, Cheng's neo-Confucian standard may be too highbrow or demanding to be necessary for a diagnosis of a whole range of candidate akratic behaviors in our mundane life.

In a nutshell, even though it may turn out to be convincing that real relationship between (practical) knowledge and action is by no means a crude definitional matter or an arbitrary stipulation, the general approach to ruling out akrasia via a talk of knowledge types manifests some (implicit or underlying) commitment to a sophisticated normative stance. When I use "sophisticated," I mean to suggest that, in a sense, whatever descriptive part or basis associated with the stance can hardly be falsified—whenever someone produces presumable evidence for a candidate akratic behavior, the theorist of this normative stance can always resort to a strategy of (re)describing or interpreting it as belonging to one of those phenomena in disguise listed above. The practical function as well as theoretical status of this well-hedged normative stance seem to be worth examining and exploring in a more careful way.

3 A Davidsonian Paradox of Irrationality

It is actually in consideration of the same intriguing, sophisticated normative stance embedded in the overall explanatory task concerning irrational behavior like weakness of will that Davidson made the following remark:

> The underlying paradox of irrationality, from which no theory can entirely escape, is this: if we explain it too well, we turn it into a concealed form of rationality; while if we assign incoherence too glibly, we merely compromise our ability to diagnose irrationality by withdrawing the background of rationality needed to justify any diagnosis at all. (Davidson 1982: 303)

Many theorists who try to provide an adequate explanation of weakness of will and its bearing on the issue of rationality fail fully to appreciate the implication of this remark, something which I believe is important for understanding the source of many apparent puzzles around strict akrasia or subjectively irrational behavior. This failure may be

partly due to the fact that Davidson himself does not elaborate in what sense the remark captures a genuine paradox, and how it relates to some of his other ideas that figure centrally in his theory of action and theory of interpretation.

Here are some preparatory remarks for us to get an initial grip of this putative paradox. Apparently it draws on some general relation between explanatory (in)coherence, (un)intelligibility, and diagnosis of irrationality. First, incoherence breeds unintelligibility at the same level of description of the phenomenon (e.g., social or psychological level, as opposed to physical level). In other words, when we diagnose a certain action as irrational, it does not necessarily mean that this action is by its physical nature unintelligible. For instance, we can understand why someone smokes despite his earlier resolution to quit smoking, and think that he is irrational for continuing to smoke. We need not know the exact details of the underlying causal mechanism to make the diagnosis or criticism. Our judgment of irrationality could be a simple matter of whatever normative position we happen to adopt. Irrationality does not seem to have a necessary connection with being paradoxical in the sense of unintelligible at every level of possible description or explanation.

Second, irrationality is a failure within the house of reason (Davidson 1982: 289). That is, if a person never engages in relevant reasoning, he cannot be judged irrational; we may call him nonrational at most. Though the irrational must not lie totally outside the ambit of the rational, it must not lie totally inside either. "Totally inside" means that everything involved in the putatively irrational behavior has an explicit or implicit reasonable explanation which seems so good as to "turn it into a concealed form of rationality"; or, in other words, the agent himself may appeal to such reasonable explanations to justify his apparently irrational behavior.

Third, "[w]hat is special in incontinence is that the actor cannot understand himself: he recognizes, in his own intentional behavior, something essentially surd" (Davidson 1980b: 113). This challenges us to explain a certain kind of putatively irrational behavior without thereby rendering that very behavior *subjectively* rational—that is, from the agent's own point of view, his action is irrational. So something must be partially outside the ambit of reason that has done the job of making irrationality. In other words, the irrational actions must involve both a rational element (viz., appealing to reasons) and a nonrational element (viz., being susceptible to non-reasons) at the core; short of either element we will not have irrationality.

Davidson follows this line of thought when he appeals to a partitioning model of mind and claims that reason has no jurisdiction across the boundary (Davidson 1986). However, many theorists have tried to point out that Davidson's treatment of incontinent actions in reconciliation with his principles concerning the relationship between reasoning and acting is unsuccessful, even on his own terms (e.g., see Margolis 1981), and too intellectualistic (e.g., see Robinson 1991; Audi 1990).[4] Such criticisms, however, have not touched the notion of the paradox of irrationality.[5] The question is: where does the paradox lie when one tries to diagnose irrationality?

Let us look closely at the two horns that purportedly make the Davidsonian dilemma: if we explain too well, for example, how weakness of will arises—that is,

[4] Robert Audi proposes, instead, a wider, holistic, and nonintellectualistic conception of rational action.

[5] I partially addressed the paradox by using a dynamic picoeconomic model to explain both the causal mechanisms giving rise to akrasia and the formation/retention of self-critical judgment of one's own irrationality (see Zheng 2001). I am not aware of any other direct and serious treatment of this Davidsonian paradox.

if we find a causally and mentally too coherent picture for it, we turn it into a concealed form of rationality; while if we assign incoherence too glibly, that is, if we exclude weakness of will from the class of intelligible phenomena, then we merely mystify the otherwise understandable (thus rationally diagnosable) human behavior by giving up the always possible effort of successful interpretation of the other.

The second horn seems relatively easy to understand. If someone fails to understand a complicated phenomenon because of his laziness or incompetence and yet excuses himself by blaming the object for incoherence, that would not make the object puzzling and apt for a diagnosis of irrationality. As for the first horn, however, we may ask: what is the demarcation between explaining appropriately and explaining "too well?" If coherence is the objective of our explanation, how can it be criticized as "too coherent?"

In order for any complaint against "a concealed form of rationality" to get off the ground at all, it seems that one needs first to offer a plausible account of how causal explanation can relate to rationality ascription under normal or standard situations; and this, hopefully, will yield a needed background principle that a proper causal explanation of an intentional behavior can legitimately put it in a form of rationality, at the personal (as opposed to sub-personal) level. Only after this picture is well-placed could one start talking about possible scenarios in which (agent-)causal explanations could go astray, including "too well," so as illegitimately to turn the behavior into a certain form of rationality.

Here is a candidate version of the background principle, call it **P1**: "any causally explanatory reason (as a causal antecedent) for an action (as the causal consequent) can rationalize the action."[6] Now we may recast the akratic action in terms of the divergence of the explanatory reason from the *ex ante* justificatory one. For one is guaranteed in principle to find a certain explanatory reason for any action already occurred while its divergence from the *ex ante* justificatory reason, stated in the agent's best judgment, is implied by the very definition of akratic action. In typical akratic situations, such a divergent, explanatory reason is not hard to find: it usually is or corresponds to a desire, a feeling, or something of that ilk, toward certain attractive feature(s) of the inferior option against which the overall best judgment is formed. Even when it is hard distinctly to identify such a causally efficacious item ready in one's consciousness, it is still easily imaginable that a causal item (which may not take any explicit form of a reason for the agent himself) lurks somewhere in his brain (or body).

Given **P1**, one can quite easily conceive a possibility that a serious attempt to make an odd action (viz. the *explanandum*) look more coherent against its backdrop may run some risk of illegitimate rationalization: the explainer may put some imaginary item(s) (viz., part[s] of *explanans*) into the targeted explanation. If the imaginary item is deposited as a reason—not necessarily a good enough reason—of the agent which fills the gap in a causally sufficient explanation of the action, then following **P1**, this

[6] What I mean by "reason" here conforms to the broadly accepted use of the term, that is, consisting mainly of desire or belief or both in the agent's psychological profile preceding the action concerned. Such reason does not always entail rationalization of a relevant kind; otherwise **P1**, or its skeleton "reason rationalizes," would be nothing but a tautology. One may specify, for instance, that in a situation where there is a minor as well as a major reason for a particular action and only the major one is causally responsible for the action, the minor, causally idle reason does not rationalize it. It is surely beyond the scope of this paper to discuss other (perhaps improved) candidate versions of such a principle, and their possible connections to another, more well-entrenched principle (call it **P2**) in Davidson's action theory. **P2**, being the converse of **P1**, claims that the (primary) justificatory reason for an action is its cause. See Davidson 1980a for the initiation and defense of **P2**.

imaginary explanatory reason will rationalize the action in some relevant sense. Thus, there is a real danger of falsely or arbitrarily ascribing such a degree of rationalizability to the irrational action that turns it into "a concealed form of rationality."

One may wonder, along this line, how anyone can ever be sure of his judgment that a certain deposited explanatory item is imaginary rather than really operative (say, in some not yet recognized way). Although such a wonder can arise in explanation of any natural phenomenon, it becomes more important when one faces a human phenomenon typically involving psychological properties of various kinds. The distinctiveness of a psychological event is this: it can be both causal and mental at the same time, meaning that it can be described in both neurophysiological (and ultimately physical) terms and in terms of its propositional content. What is special about mind is not just that it creates and understands meanings, that is, it has certain mental states or events with propositional contents, but also that the mental states describable by folk psychological terms do not have clear conceptual boundaries among themselves. Typical folk psychological terms include belief, desire, reason, judgment, want, intention, will, volition, emotion, and many others; and they normally overlap, interweave, or change into one another, and may do so variably in various contexts. For those who do not readily see this fact, let me quote one passage, out of many similar ones, to illustrate it a little bit:

> [O]ur ordinary mental terms display riotous overlap, alluring vagueness, categorial ambiguity, and rich shades of nuance … for example, instead of ascribing a belief to an agent we could often ascribe instead a memory, an opinion, an assumption, a prejudice; or something held, taken for granted, known or recalled … is it a *belief* that I have two hands, or should it rather be described as something I take for granted? When a dog is barking beneath a tree, can we say that he *believes* that a cat is up it? … Our mental life seems to be an excellent example of a Heraclitean flux (cf. William James: "Consciousness is nothing jointed; it flows") which we can divide up as we will and which rarely suggests or prescribes any single method of categorial division. (Wilkes 1981: 152)

What this fact implies for the issue of diagnosing irrationality is that we often have enough maneuvering room for either explaining (thus rationalizing in some sense, given **P1**) or dismissing (ruling out) whatever puzzling phenomena as we find exigent under particular practical situations.

Precisely because of this unavoidable conceptual looseness, we almost always have a choice between assigning this or that level of coherence to this or that part of mind with corresponding propositional contents or logical relations. This implies that there is in theory more than one possible way to obtain the targeted holistic explanatory coherence. It is thus perfectly conceivable that a certain way of obtaining holistic coherence could better (than other ways) preserve the integrity of the normative rules we accept as our basic principles of reasoning and/or action.[7] When we rule out, for instance, an assumed mental item of an agent as nonexistent and base our critical

[7] Apart from pure logic rules, there are varieties of rules which reasonable people either invariably or most likely accept. For instance, Carnap and Hemple have argued for a principle which all rational people will accept though it is no part of the inductive logic. It is the *requirement of total evidence for inductive reasoning*: give your credence to the hypothesis most strongly supported by all available relevant evidence. See Davidson 1980b: 112. The other example is the Plato Principle indicated in the beginning, to which I will turn shortly.

diagnosis on the breach thus "created" on the part of the agent, we are often not doing it arbitrarily but rather following relevant rules as far as possible.

Even when we talk about easy and natural behavior under normal circumstances, we are also following rules of various kinds; and the rules we follow, or the ways of applying them, could be contextually wrong without our recognition. Hence there is *always* a question in mapping out the mind under explanation in a certain way: that is, whether we have struck the best balance between fitting individual mental items with the overall observable behavioral patterns on the one hand,[8] and maintaining our critical ability in following rules, including those basic norms of rationality that constitute the background for our making sense of things, on the other.

However, again, where exactly is the paradox, except for the likely tension embedded in the above question of striking the best balance? In almost every sphere of human practice, there are tensions related to attaining optimal balance of certain kinds. We do not usually call such tensions "paradoxes." Perhaps the paradoxical element lies not so much in difficulties of ruling out fictitious items according to certain rules in action explanations, as in a normative *necessity* of rational criticism of some incoherence putatively identifiable in the irrational actions under explanation.

The degree of rationality with which a person behaves seems to be some inherent property of his, which can thus never be dependent upon the third-person view of his behavior. At least at any particular moment of his life, it seems to be a fixed fact about his mental capacities or tendencies that he exhibits certain general patterns of thinking and motivation. Without disputing this, how can we then make sense of the claim that the interpreter is entitled to assign a certain degree of coherence/rationality to the person being interpreted?[9] In other words, isn't it paradoxical to say that some independent, fixed fact about one mind is dependent upon another mind's decision to, how to, or even not to, understand it?

One reply to this question is perhaps that this apparent paradox can be easily eliminated by a more careful expression drawing on the distinction between ontological independence and epistemological independence: that is, the ontological independence of a mind from the interpreter does not contradict the epistemological dependence of whatever constitutes a perceived "fact" about the mind upon the conceptual apparatus and interpretive strategies of the interpreter. So one can consistently hold that any "factual" statement about an object is nothing more than a descriptive form of understanding from a certain interpreter, while the interpreter need not deny the ontological independence or coherence of the object. The ontological assumption concerned here is that any natural thing or event is coherent and thus mind must be coherent in virtue of its being a natural thing (or mental events in virtue of their being physical events).[10] This assumption is independent from, or prior to, any particular human (and likely hermeneutic) understandings of mind.

When we judge some action internally or subjectively irrational, that is, the agent intentionally violates the criteria he has freely adopted and is presently holding, are we

[8] Theoretically speaking, such best balance struck in our interpretation may not necessarily be the balanced point or state in the agent's own deliberation when he acts.

[9] This expression certainly allows for the possibility that the interpreter and the interpretee are one and the same person, that is, one wants to understand himself.

[10] It is certainly beyond the ambition of this essay to demonstrate that this is not a mere assumption or we have good reason to believe it.

ascribing some natural incoherence to the agent as if it were a fundamentally mysterious defect in his natural makeup? If the answer is positive, a problem of misconception seems to have occurred: we transgress the ontological territory with reasons that merely have epistemological relevance. If, on the other hand, we contend that what we are ascribing is not ontological incoherence but rather an explanatory one, namely, the incoherence in *our* explanatory scheme between a certain supposed mental item of the agent and his other more or less well-established mental parts as well as certain well-accepted rules of action/thought, then the question is why we should ascribe to the agent the source of such an interpretive problem, which seems only to reflect some imperfect or incompetent state of ourselves as interpreters. A further, related question is: what purpose does it achieve to judge certain actions *internally irrational* rather than hermeneutically and temporarily unintelligible?

Here is a possible approach to the plausible answer. No matter what other possible sources of the explanatory incoherence there could be on the side of an interpreter, there is always one possibility that it is caused by, or corresponds to some real mismatch or loopholes on the side of the agent between his own conscious reasons for action and other mental items associated with certain underlying (perhaps unknown) mechanisms of action. Therefore it is always, or *in principle*, open, as a default or alternative strategy, for interpreters to ascribe the origin of whatever explanatory residue to the agent's side. Where or when they are fully justified to do so is a partially normative question, that is, depending on many empirical, contextual factors as well as certain general normative concerns, such as those involved in "the principle of charity" in radical translations.[11]

Interpreters must be so careful as to avoid premature accusation of irrationality, that is, to distinguish such shallow sources of incoherence as some of the enlisted phenomena in (1)–(6) of Section 2 from other deep sources of incoherence on the side of the agent, for only the latter could provide the possible ground for the criticism of irrationality. When I say "shallow," I mean that it is relatively easy to eliminate or circumvent the sources once they are identified; correspondingly, the so-called "deep sources" refer to those that are harder either to identify or to eliminate.

Without getting into any detailed discussion of these possible deep sources of incoherence that could give rise to the so-called "internal irrational" actions, I am more interested in exploring a more subtle layer of normative interpretation. Let us pick up the Plato Principle, mentioned at the beginning of the paper, as one important principle in our charitable interpretation. Recall, the Plato Principle states that no one willingly acts counter to what he knows to be best. It is interesting to notice that this expression of the Plato Principle sounds like a pure *description* of a natural fact, that is, everybody by nature acts (or reasons) intentionally according to what he knows to be best (or correct). However, this could hardly be true unless you would define "what he knows to be best" here as nothing but whatever momentary mental state is revealed by his subsequent, ostensible act, and thus make this statement trivially true, that is, true by definition.

[11] "Principle of charity" was first put forward by Quine in his account of radical translation; Davidson nevertheless is its foremost proponent in interpreting both languages and actions. The basic tenet of this principle is to discourage judgments of irrationality. See Quine 1960: 26–79 (Chapter 2); Davidson 1984. For a comparative discussion of principles of charity of different degrees of severity, see Thagard and Nisbett 1983.

A plausible view about the nature of the Plato Principle seems to be that, despite its descriptive look, it is a *normative* principle, whose proper function is not to describe or predict, but to prescribe, guide or cultivate. In this light, we can make good sense of the point of judging an incontinent action as irrational: because it does not live up to the normative standard which the agent's behavior would naturally or normally match and could not rationally reject if he were to become aware of it, and because we believe that the action can be rectified according to that standard. On the other hand, however, we never need to suppose that the irrational action is ontologically incoherent. Thus we always keep the possibility open that one day we may be able to fully explain the natural genesis of the action in its utmost details. Even then we are not to "turn it into a concealed form of rationality," for we will still hold it as normatively incoherent—unless one day we modify or abandon our normative principles in light of whatever newly-discovered psychological and/or other truths.

In a nutshell, as long as we are fully aware of the normative nature of these basic principles or requirements by which we judge and criticize irrational behavior in our justifiable human practice, the original Davidsonian paradox of irrationality seems to dissipate.[12]

4 Relation between Holding True and Making True: Revealing a Confucian-Mencian Paradigm

Even after one is convinced of the normative nature of the Plato Principle, one may still wonder why we should adopt this rather than that particular principle as a normative principle of *interpretation* (as opposed to, say, *education*). Are there any natural, ultimate constraints on our freedom to enact and sanction such a normative principle? Should the content of the principle be constrained by those actual existing capacities of the minds, whose actions are supposedly to be governed by the principle? In other words, is there any issue of descriptive accuracy or factual approximation with regard to a normative principle?

We can easily think of some normative rules whose sole purpose is but to demand conformity, for instance, a work dress code. The rule-makers need not pay much attention to whether, or to what extent, the people involved already have those behaviors required of them by the rules; the only concern here is that the required behaviors obtain, no matter whether, or how much, they are out of the people's efforts to comply with the rules. Obviously, there is no question of descriptive accuracy in these rules about the *status quo* or average condition of the people to whom they apply. A notable thing here, however, is that a presupposition of any rule-maker must be that the people to whom the rule applies are *capable* of living up to it,[13] or changing themselves in that direction, if they are not already so; otherwise there would be no

[12] My solution of the Davidsonian paradox, however, distinguishes itself from a *prima facie* similar solution suggested by David Henderson, whose approach is to accept all attributions of explicable irrationality even without a complete causal explanation of it: the latter solution does not stress the normative nature of irrationality attributions, nor does it draw on the distinction of incoherence at different levels (Henderson 1987).

[13] A well-accepted principle corresponding to this presupposition is "ought implies can."

point in making a rule. In short, *status quo* is no concern whereas potentiality or "perfectibility" (in a narrow sense) is a must for this category of rules.

When it comes to the domain of interpretation, however, things become quite different. As interpretation presumably targets at some existing and fixed phenomenon, it is truth-oriented and descriptive at its core. Then, where is any room left, in a substantive interpretation, for normative content of a compliance-oriented and prescriptive principle? Hence there seems to be a sense of paradox in the very notion of "a normative principle of critical interpretation." That is, on the one hand, interpretation demands truth about the *status quo* of the person being interpreted whereas on the other, the special (i.e., critical) normativity of the principle demands adaptation to a standard from the very person, which presupposes his potentiality for change. How can one put these two opposite requirements in one place without being self-contradictory?

It is plain that attempts of rational criticism, based on normative standards, of the alleged irrational object, are conceptually distinctive and theoretically separable from attempts to interpret the same object. Were these two kinds of attempts also separable in reality of human practice, there would be no credible ground at all to talk about normative principles of critical interpretation, and whatever paradox thereof. Unfortunately, however, the hermeneutic situation in reality is as remote from this picture as one could possibly imagine. In Davidson's expression of the paradox, and some of his dialectical remarks elsewhere, there seems to be an intuition underlying his emphasis on the necessity of the background of rationality for diagnosing irrationality: that is, it is impossible to separate interpretation from normative critique of irrationality against the largely rational backdrop (wherever it comes). The idea is that without sufficient number of fundamental attributes of rationality granted to the agent being interpreted, no meaningful interpretation can ever get off the ground.

At certain levels, there could only be one possible mode of understanding—the critical understanding *constituted* as well as *regulated* by basic rational norms which we have no choice but to take as the necessary preconditions of having thoughts at all. As Davidson himself puts it, "to identify at least some irrationalities with inner inconsistencies … is not to explain, or even to go very far in describing, such psychological states; indeed, it makes the problems of description and explanation seem impossible" (Davidson 1985: 346). Here the "description and explanation" refer to those efforts of understanding that are free of rational norms like the Plato Principle.

If certain rational norms and principles are necessary grounds or indispensable components for any meaningful understanding of human behaviors, an inevitable implication is that they approximate, or cannot deviate too much from, existing conditions of the normal people or their potentialities.[14] On the other hand, however, it does not seem to make practical sense to ask what level of the present descriptive accuracy of these norms and principles is perfect to our normative human practice as a whole. For, as one may plausibly argue, most of these norms are not products of deliberate human designs but rather "filtered deposits," as it were, out of some long evolutionary process of human adaptations (to nature as well as to each other).

[14] It is worth emphasizing, if it is not already crystal-clear, that the "human behaviors," as well as the "norms and principles" for understanding them, are not domain-specific behaviors such as moral or legal ones; rather, we are talking about the most general, common, or minimal aspect or layer of any understandable behavior. That said, it is no contradiction to make the present claim while simultaneously yielding the possibility that a morally abnormal person could fail the normal ethical conditions most of the time.

There is another sense in which talking about choice or redesign of our basic rational norms seems incredible. Since these norms have long been an indispensable *constitutive* part of our social practical life of communicating, interpreting, and criticizing each other, a large proportion, if not all, of our behavior and thought, with their patterns, habits, dispositions, and so on, must always already be shaped by these norms to such an extent that any wholesale change or radical departure is unthinkable. In short, rational criticisms normally have their force of validity because the norms on which they are based tend to be believed to have such a degree of natural veracity that their gross violation would be deemed as perverse, absurd, or unnatural. The more "natural" or descriptive the coat of these norms looks, the stronger the normative-constitutive power or effect they possess or produce. This is an insight which, I believe, is derivable from reflection on the Davidsonian paradox of irrationality—an insight about some interesting relation between *holding true* (corresponding to the so-called "descriptive coat" above) and *making true* (corresponding to the "normative-constitutive power" above).

Now, with this insight, let us take a new look at some of the familiar passages in Confucian classics, hoping to gain some deeper understanding of the relation and relevant issues. Mencius' most distinctive and well-known idea about the goodness of human nature is not only a foundational part of the Confucian picture of Heaven and man but also a most profound inspiration for subsequent Confucian moral practice or orientation. In a justly famous passage about one's seeing a child on the verge of falling into a well, he remarks that "whoever is devoid of the heart of compassion is not human, whoever is devoid of the heart of shame is not human, whoever is devoid of the heart of courtesy and modesty is not human, and whoever is devoid of the heart of right and wrong is not human…. Man has these four germs just as he has four limbs" (*Mencius* 2A6).[15]

Just by the explicit linguistic forms of these expressions, it is obvious that they are descriptive instead of normative (i.e., involving words like "ought" or "should"). Even though the underlying or ultimate point of such expressions is (at least partially) normative, their descriptive outlook may have some unique, irreplaceable role in attaining or inducing the moral objective corresponding to the point. The obvious formal similarity of this paradigmatic remark of Mencius' to the Plato Principle illuminated above is not accidental. Or so I shall contend.

Mencius seems to place the greatest emphasis on the heart of shame: "great is the use of shame to man…. If a man is not ashamed of being inferior to other men, how will being their equal have anything to do with him?" (*Mencius* 7A7) If the heart of shame is an incipient tendency deep-seated in human nature, Mencius is certainly consistent in holding both that man, by his very nature, tends to feel shame about his actual falling short of others' dutiful or continent behavior exhibiting characteristic feature of the (normative) humanity itself, and that "the (descriptive) difference between man and the brutes is slight" (*Mencius* 4B19). On the one hand, given the factual slight difference between man and brutes, nobody (short of sages) could automatically be warranted in always, or even normally, exhibiting moral/ rational behavior; on the other hand, nevertheless, everybody has aspiration as well as potentiality to become, or live up to the standard of, a morally better person whose ideal type, exemplified by sages, is *definitive* of being a true human being. The latter aspect means that it is by no means

[15] The English translations of the *Mencius* here and after are from Lau 2003.

arbitrary, fantastic, or obscure to *describe,* rather than *merely wish*, the mature, fully developed state of a gentleman as the *natural* (as opposed to *artificial* or *queer*) status of man in the universe.[16]

No matter whether, or to what extent, the descriptive form of relevant expressions is intended with a normative-constitutive meaning or component, an undeniable moral-psychological fact seems to be that nobody could really bear the awareness that he is intrinsically inferior to others, something less worthy of the name "human being," or something merely at the rank of brutes. One might be, for various reasons, ignorant or self-deceived about possible evidence pointing to such inferiority. As Mencius observes, "When one's finger is inferior to other people's, one has sense enough to resent it, but not when what is inferior is the heart. This is what is called ignorance of priorities" (*Mencius* 6A12). That precisely implies, however, that had one got clear evidence, or faced universal inescapable criticism, of one's own morally inferior behavior, the motivation to change the resentful situation by rectifying the behavior would be natural. Hence the source of the possibility for moral progress.[17]

The same thing can be said about the rational inferiority or subjective irrationality in the Davidsonian sense, which also falls under the general, Mencius' category "inferiority of the heart" here. More pertinently, we had better use one specific type of Mencius' incipient tendencies in the heart, namely, "the heart of right and wrong," to make salient a unique feature of rational appraisals, that is, their built-in *critical intent*. We shall, following D. C. Lau, note a crucial connection between cognitive judgment and conative sanction, which seems to be suggested by Mencius' theory of human nature:

> "[T]he heart of right and wrong" has a twofold significance. First, it refers to the ability of the heart to distinguish between right and wrong. Second, it can also refer to the approval of the right and the disapproval of the wrong by the heart. Now this ability of the heart is relevant to the understanding of the reasons for Mencius' holding the view that human nature is good. For even when we fail to do what is right we cannot help seeing that what we have failed to do is right and feeling disapproval toward the course of action we have chosen, with its accompanying sense of shame. In this way the statement that human nature is good is given a sense which is completely independent of the way in which human beings in fact behave. (Lau 2003: xix–xx)

The remarkable thing here is the emphasis on the logical independence of Mencius' (implicitly) normative notion of human nature from the actual level of people's behavioral performance or moral/rational development, though what is implied by such a normative nature is in principle not beyond the reach of humanity, that is, the potential human capacities naturally endowed by Heaven,[18] or from evolution (if put in terms of modern science).

[16] One may want to say that it is comparable to the Aristotelian sense of "second nature" in which Mencius understands man's natural status here.

[17] It may take some adequate amount of empirical evidence in experimental psychology to confirm my hypothesis here, that is, that descriptive, or implicitly normative, forms of expressing principles of human action tend to be motivationally more effective than direct prescription or explicit commandments.

[18] Lau's interpretation here is by no means idiosyncratic; for example, ZHANG Dainian 張岱年 seems also to share such an interpretation of Mencius (see Zhang 2005: 185–188).

The special normative status of human nature for Mencius can be made clearer against a contrastive backdrop well-afforded by Xunzi's 荀子 distinction between *xing* 性 ([human] nature, or whatever one is readily born with) and *wei* 偽 (deliberate effort, or what is accomplishable only through learning, nurturing, or deliberate practicing).[19] Despite the analytic clarity and descriptive utility of the *xingwei* 性偽 distinction, which obviously empowers Xunzi's criticism of Mencius on the issue of appraising human nature, it by no means closes the alternative possibility of defining human nature in a way that is not constrained by this distinction.

Mencius' approach to human nature is to pick up the slight yet distinctive features of certain mature human representatives, for example, sages, as evidence for his conviction that states embodying these features are in principle within the reach of everyone; for every member of the human species must have the same germs for the accomplishment as sages do, otherwise it would be impossible even for the sages to accomplish them. Potentiality, instead of actuality or probability of realization, must be intrinsic to, or representative of, a particular species. In contrast, Xunzi's logic behind his approach lies in treating what is complete and ready-made (by heaven, so to speak) at the beginning of each individual life, thus unalterable by any later deliberate efforts, as exclusively belonging to *xing*. Potentiality, from this perspective, does not qualify as *xing* because, by definition, potential is neither complete nor indicative of any particular moral outcome without "cumulated *wei*" (*jiwei* 積偽). "How could *li* 禮 and *yi* 義, being nothing but *cumulated wei*, be the *xing* of human being?" "Therefore, what is common between sages and the mass belongs to *xing*, whereas what is different between the two is due to *wei*" (*Xunzi*, "Xing E 性惡" ["Human Nature Is Bad"] chapter; see Knoblock 1988–1994: [III] 153). So Xunzi, following this logic, would be able to deny that any particular accomplishment of a sage can entail universal existence of moral germs as part of the fixed *xing*.

Here is no place to ponder and adjudicate on the dispute between Xunzi and Mencius. What is particularly interesting and relevant to our present purpose is a connection of this dispute to the idea of implicit (as opposed to explicit) normativity which seems only associable with Mencius' approach. Let me elaborate a little bit.

It goes without saying that Mencius and Xunzi share the ultimate Confucian ideal of realizing *ren* 仁, *yi* 義, *li* 禮, and *zhi* 智 for everyone. That is, both acknowledge the realizability of the ideal. Their difference, one may claim, lies in the road or method to attain this ideal. Given Xunzi's conception of *xing*, moral or rational behavior in the direction of this ideal must be crafted through remolding our given *xing* according to certain explicit normative procedures. Mencius, on the contrary, starts with an implicit normative conception of *xing*, and takes certain actual end-results conforming to this ideal as evidence for *xing*'s fitness or homogeneity to the ideal, hence binding the elements of the ideal, or their germs, with *xing*. Then the moral task for everyone is just to enrich and enlarge *xing*'s potentials, perhaps against all extrinsic odds, to let it fulfill its destiny (*ming* 命), that is, the actualization of the ideal. Such a diachronically holistic move of bundling together two terminals of a process, that is, certain teleological results at one end and a certain status of initial conditions at another end, reveals a distinctive form of dynamic (or evolutionary) normativity, regardless of the degree or specific form

[19] I largely follow ZHANG Dainian here in his concise exposition of Xunzi's dispute with Mencius (see Zhang 2005: 189–191).

in which Mencius, or anyone, gains explicit self-consciousness of it.[20] To comprehend the same bundling move in different guises, let us take a close look at another eye-catching passage in Mencius:

> The way the mouth is disposed towards tastes, the eye towards colours, the ear towards sounds, the nose towards smells, and the four limbs towards ease is human nature, yet therein also lies the Decree. That is why the gentleman does not ascribe it to nature. The way benevolence pertains to the relation between father and son, duty to the relation between prince and subject, the rites to the relation between guest and host, wisdom to the good and wise man, the sage to the way of Heaven, is the Decree, but therein also lies human nature. That is why the gentleman does not ascribe it to Decree. (*Mencius* 7B24)

Why doesn't the gentleman ascribe the dispositions or functions of our sensual organs to nature, but instead to some normative Decree, while explicitly acknowledging that they belong to human nature? Correspondingly, why doesn't the gentleman ascribe the moral proprieties to Decree, but instead to some descriptive nature, while explicitly acknowledging that they belong to the Decree? Obviously, Mencius does not deliberately blur the conceptual distinction between nature and decree, or celebrates the gentleman's arbitrary ascription based on such a blur.

The plausible answer, I propose, has something to do with the bundling idea concerning the deeper relation between natural endowment and moral destiny prescribed by the Decree. More concretely, fitting exercise of endowed nature points to a certain normative destiny while feasible realization of the normative ideal requires certain potentials rooted in such nature. Were there no normative constraint on our given nature, the difference between gentlemen and "little men" (*xiaoren* 小人, or morally undeveloped men) would not emerge; were there only and explicitly normative high talk (without heeding our natural constraints), the similarity among all human beings or even their similar origin to other animals would be neglected—in other words, the naturalistically grounded universal possibility for men to become moral would not be emphatically revealed. The bundling idea is a reflection, as well as a promotion, of the hidden, dialectic relation between holding true and making true.

One special reason I call Mencius' deeply normative approach "implicit" is that it does not regard the ultimate moral ideal as something imposed from outside, like an explicit order from some arbitrary external authority which may be indifferent or irrelevant to the development of the natural endowments of the subjects who receive the order. Rather, the ultimate moral ideal is the central, constitutive part of the type of beings the subjects are supposed to be or become. More specifically, this normative supposition is "implicit" in a dual sense: first, the germ metaphor suggests that the subjects already have the essential ingredients only to be properly developed; second, whatever normative force for the subjects to achieve the ideal had better come from within, or from stimulation via their internal dynamic motivational structure, rather than from compulsory remolding associated with certain psychic engineering programs which Xunzi seems to be fond of.

[20] It might be interesting to note, in passing, that the Chinese character *duan* 端, which is a keyword in the key passage quoted above (i.e., *Mencius* 2 A6), can be standardly translated as "terminal" as well as "germ."

Both implicit aspects of this Mencius-style normativity evidently fit and support the insight, indicated earlier, about the practically significant relation between holding true and making true. The first implicit aspect, related to the descriptive coat of Mencius' germ metaphor, has a role in warranting one's basic confidence in one's own potential for the expectable result; while the second implicit aspect, bearing on endogenous source of change or correction, has a role in mobilizing hidden energy associated with, for example, "the heart of shame."

Since the main objective of this essay is not about moral education/cultivation, I will not pursue the practical side of the relation between holding true and making true. Instead, let us take a new look at the theoretical side of the relation, that is, the question about possible merit of the Mencius-style normative approach in explaining weakness of will from a vantage point which clearly emerges with our reflection on the relation.

Given what has just been presented, we can see better what lacks in Searle's assertion that the almost ubiquitous mundane existence of weakness of will falsifies the Davidsonian perplexity over its possibility. A contemporary Mencius-minded reply to the Searlean accusation would be this. It is perfectly legitimate to wonder how weakness of will is possible, given our fundamental belief in the goodness/rationality of human nature qua some kind of descriptive truth (which implies that weakness of will is ontologically problematic). At the same time, however, we, as theorists, should not forget the deep implicit normativity of such a belief, with its natural inevitability in the holistic cosmological process as well as its actual fragility in real life which is always full of various contingencies. In other words, we should never lose sight of the persisting background fact that the difference between man and brute is slight, while we recognize the rational need and the realistic chance or prospect for fulfilling Heaven's decree that man be moral. It is the tension between these two sides, or the ineradicable disparity between the temporal factuality and the timeless hidden normativity that ultimately accounts for how weakness of will is possible.

In short, Mencius' fundamental insight is this. Not only does man have natural roots similar or homologous to animals, but, no less importantly, man also has a natural destiny (*ming*) to become fully rational and moral; in other words, normativity, explicit or implicit alike, is also natural in a distinctively Confucian-Mencian sense.[21] No matter whether, or to what extent, that may ring a paradoxical overtone for thorough-going reductivist or physicalist ears, this is a destiny we cannot help finding for ourselves in the natural world.

5 Some Concluding Remarks

This essay has an ambitious ring in the following two senses. First, it tries to critique or engage a wide range of existing works on a hotly-contested, age-old topic, weakness of will, from a new perspective of comparative philosophy combined with a largely neglected Davidsonian paradox of irrationality. Second, it ultimately aims at some unifying thesis about a certain implicit type of dynamic normativity which, as far as I know, has never been well-articulated.[22]

[21] Arguably it is also in a reinterpretable Aristotelian sense of taking man as "rational animal."

[22] I call this thesis "diachronic holism" and have recently attempted some articulation of it via a new analysis of the well-known Davidsonian thought experiment Swampman. See Zheng 2016.

Given this ambitious background intention, it is probably no wonder that the essay is almost inevitably incomplete (partly due to the space limitation). For one thing, the incompleteness lies in its obviously insufficient treatment of Confucian (especially neo-Confucian) resources that apparently have direct or indirect bearings on some of the central issues involved here.[23] For another, the incompleteness lies in its being unable to develop a highly significant line of thought which is barely broached here, that is, a diachronic holism of which the Mencian insight presented above is but one representative.

Although it is impossible to amend such incompleteness in a few short paragraphs, I do want to summarize, in broadest strokes, the main ideas of the paper so that, hopefully, certain structural or logically progressive relations among them become more visible. The first idea is the notion of implicit normativity, which apparently can be manifested in different ways. One of the most interesting ways it can manifest itself is that of a dynamic process via which a certain descriptive potential is normatively realized, with a possible mixture of foresight and hindsight. Descriptive elements and normative elements, though conceptually distinct, often blend and penetrate into each other. The second idea is the interplay between the descriptive and the normative in the very notion of critical interpretation. This notion is centrally involved in understanding the Davidsonian paradox of irrationality. The third idea is the unique transition from (a deep-normative solution to) the paradox of irrationality to the relation between holding-true and making-true. Here seems subsumable the non-accidentalness of the descriptive coat of the Plato Principle as well as some of the Mencian paradigmatic tenets. The fourth idea is about how the diachronically holistic tendency (or the underlying paradigm) of the Mencian approach, under some charitable reading, can provide a proper resolution to the Searle-Davidson dispute over weakness of will, which is in fact a plausible and practically significant solution of the Davidsonian paradox of irrationality. The final idea is about the interdependent bi-directions in comparative philosophy, no matter how incomplete my intended exemplification of them turns out to be: in one direction, I interpret Mencius in light of a Davidsonian line while at the same time I interpret the Plato Principle (associated with the Davidsonian paradox) in light of the Chinese line of diachronic holism. At the end of the day, what matters is not the minute accuracy or loyalty to the "original" sources but rather the coherent and enlightened new synthesis.

Weakness of will is a practical as well as a theoretical challenge to our unique place in nature or our conception of it. Our successful responses to this challenge, though perhaps never perfect, reveal the inescapable, fundamental nature of ourselves as normative beings in a deepest sense. Let me end the paper with a well-known saying of Confucius, which hopefully may show some new layer of meaning in light of our discussions above. "Is benevolence really far away? No sooner do I desire it than it is here" (*Analects* 7.30).

[23] DAI Zhen 戴震 (a well-known scholar in the Qing 清 dynasty), for instance, has certain insightful critiques of neo-Confucian representatives in the Song 宋 dynasty for their problematic understanding of the Mencius-Xunzi dispute on human nature, which seem able to fit with my treatment of the issue (see Dai 1982: 25–38).

References

Analects, *The*. 1993. Trans. by Raymond Dawson. Oxford: Oxford University Press.
Audi, Robert. 1990. "Weakness of Will and Rational Action." *Australasian Journal of Philosophy* 68: 276–281.
Dai, Zhen 戴震. 1982. *Annotation and Interpretation of the Mencius* 孟子字義疏證, 2nd ed. Beijing 北京: Zhonghua Shuju 中華書局.
Davidson, Donald. 1980a. "Actions, Reasons, and Causes." In his *Essays on Actions and Events.* Oxford: Clarendon Press.
____. 1980b. "How Is Weakness of the Will Possible?" In his *Essays on Actions and Events.* Oxford: Clarendon Press.
____. 1982. "Paradoxes of Irrationality." In *Philosophical Essays on Freud*, edited by Richard Wollheim and James Hopkins. Cambridge: Cambridge University Press.
____. 1984. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.
____. 1985. "Incoherence and Irrationality." *Dialectica* 39: 345–354.
____. 1986. "Deception and Division." In *The Multiple Self*, edited by Jon Elster. Cambridge: Cambridge University Press.
Fraser, Chris, and Wong Kai-yee. 2008. "Weakness of Will, the Background, and Chinese Thought ." In *Searle's Philosophy and Chinese Philosophy*, edited by Mou Bo. Leiden: Brill.
Henderson, David. 1987. "A Solution to Davidson's Paradox of Irrationality." *Erkenntnis* 27: 359–369.
Huang, Yong. 2008. "How Is Weakness of the Will Not Possible?—Cheng Yi's Neo-Confucian Conception of Moral Knowledge." In *Education and Their Purposes*, edited by Roger T. Ames and Peter D. Hershock. Honolulu: University of Hawai'i Press.
Jiang, Xinyan. 2000. "What Kind of Knowledge Does a Weak-willed Person Have?—A Comparative Study of Aristotle and the Ch'eng-Chu School." *Philosophy East and West* 50.2: 242–253.
Knoblock, John. 1988–1994. *Xunzi: A Translation and Study of the Complete Work*, 3 vols. Stanford: Stanford University Press.
Lau, D. C., trans. 2003. *Mencius*. Hong Kong: The Chinese University Press.
Margolis, Joseph. 1981. "Rationality and Weakness of Will." *Journal of Chinese Philosophy* 8: 9–27.
Quine, W. V. 1960. *Word and Object*. Cambridge, MA: MIT Press.
Robinson, Kirk. 1991. "Reason, Desire, and Weakness of Will." *American Philosophical Quarterly* 28: 295–296.
Searle, John R. 2008. "Reply to Chris Fraser and Kai-yee Wong." In *Searle's Philosophy and Chinese Philosophy*, edited by Mou Bo. Leiden: Brill.
Thagard, Paul, and Richard E. Nisbett. 1983. "Rationality and Charity." *Philosophy of Science* 50: 250–267.
Wilkes, K. V. 1981. "Functionalism, Psychology, and the Philosophy of Mind." *Philosophical Topics* 12: 152–153.
Zhang, Dainian 張岱年. 2005. *An Outline of Chinese Philosophy* 中國哲學大綱. Nanjing 南京: Jiangsu Jiaoyu Chubanshe 江蘇教育出版社.
Zheng, Yujian. 2001. "Akrasia, Picoeconomics, and a Rational Reconstruction of Judgment Formation in Dynamic Choice." *Philosophical Studies* 104: 227–251.
____. 2006. "*Ex Ante* vs. *Ex Post* Rationalization of Action." In *Philosophical Anthropology / The Proceedings of The Twenty-First World Congress of Philosophy*, edited by Stephen Voss. Istanbul: Philosophical Society of Turkey.
____. 2016. "The Swampman Puzzle and Diachronic Holism." *Philosophical Forum* 47: 171–193.