Nature, Reasons, and Moral Meaningfulness

Pierre Charette
Department of Philosophy
McGill University, Montreal
August 2008

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of PhD

© Pierre Charette 2007
RÉSUMÉ .............................................................................................................................. iv

ABSTRACT .......................................................................................................................... vi

ACKNOWLEDGEMENTS ........................................................................................................ viii

INTRODUCTION .................................................................................................................. 1

CHAPTER 1. PETER STRAWSON: SOME ESSENTIAL FEATURES OF MORAL LIFE .................................................................................................................. 30

1.1. THE STRAWSONIAN FRAMEWORK .............................................................................. 32

1.1.1. The connective model of meaning and our fundamental cognitive structure .......... 35

1.1.2. A rationally acceptable language and ontology: Strawson vs Quine ...................... 39

1.1.3. Strawson’s metaphysics of experience ..................................................................... 46

1.1.4. The rational acceptability of secondary-quality terms ............................................ 49

1.2. FREEDOM, RESENTMENT, AND NECESSITY ......................................................... 52

1.2.1. Optimists, pessimists, and the contrast between “participant” and “objective” attitudes ................................................................................................................................................ 56

1.2.2. Moral sentiments, the “manifestation requirement”, and its “responsibility condition” .. 60

1.2.3. Strawson’s Humean argument: irremovability by nature ......................................... 67

1.2.4. The argument from the distinctness of languages (ca 1962) ...................................... 69

1.2.5. “Freedom and Necessity”: a compatibilist analysis of the free, responsible self ......... 73

1.3. STRAWSON IN PERSPECTIVE ......................................................................................... 81

CHAPTER 2. DAVID WIGGINS: HUMAN NATURE AND MORAL LIFE ...................... 87

2.1. THE WIGGINSIAN FRAMEWORK ............................................................................ 94

2.1.1. Underdetermination, anthropocentricity, and the practical nature of morality .......... 99

2.1.2. Wiggins’ metaphysics of valuational experience ...................................................... 111

2.1.3. A metalanguage for the interpretation and the critical assessment of discourse ...... 113

2.2. WIGGINS’ ANTHROPOLOGY OF MORAL LIFE ......................................................... 136

2.2.1. A Humean theory of human nature ........................................................................ 138
2.2.2. The unforsakeable and the first-order ethic ................................................................. 154
2.2.3. A libertarian conception of free will ............................................................................. 169

2.3. METAETHICS AND CRITICISM ........................................................................... 173
2.3.1. Weak moral cognitivism and the genealogical account of human experience .......... 175
2.3.2. Criticism of rival views ............................................................................................. 187

2.4. WIGGINS IN PERSPECTIVE .............................................................................. 210

CHAPTER 3. DENNETT'S FRAMEWORK AND ACCOUNT OF HUMAN NATURE

3.1. THE DENNETTIAN FRAMEWORK ...................................................................... 223
3.1.1. Essentialism, the cosmic pyramid, and "skyhooks" .................................................. 224
3.1.2. Gradualist accounts and the rationally acceptable language .................................... 232
3.1.3. Dennett’s ontology ................................................................................................. 240
3.1.4. Dennett's epistemology and the anthropocentricity thesis ...................................... 247
3.1.5. The targets of Dennett’s criticism ........................................................................... 256
3.1.6. The intentional level of discourse ............................................................................ 260
3.1.7. The functional level of discourse ............................................................................ 270

3.2. EVOLUTION: FROM SELF-REPLICATORS TO PRIMATES .................................. 274
3.2.1. The evolution of some phenomena ........................................................................... 275
3.2.2. The evolution of intelligent agency .......................................................................... 281
3.2.3. Cognitive abilities in animals .................................................................................. 291

3.3. PROTO-MORALITY IN PRELINGUISTIC HOMO SAPIENS ............................ 300
3.3.1. The emergence of proto-moral cognitive dispositions and norms: game-theoretical accounts ...................................................................................................... 303
3.3.2. Capacities and faculties of prelinguistic Homo sapiens ........................................... 313
3.3.3. Proto-moral sentiments ........................................................................................... 319

3.4. MIND, LANGUAGE, AND LANGUAGE-PRODUCED PHENOMENA ........ 324
3.4.1. Primitive cognitive areas and language .................................................................... 327
3.4.2. Dennett’s theory of the mind................................................................. 331
3.4.3. Language-produced phenomena......................................................... 346
3.4.4. The self ............................................................................................... 349
3.4.5. Rational criticism: the practice of giving and asking for reasons........... 354

CHAPTER 4. THE PERSONAL STANCE: DENNETT’S PHILOSOPHY OF
MORALITY ..................................................................................................... 360

4.1. THE GENEALOGY OF MORAL LIFE: THE “BIRTH OF VALUES” INSIGHT
...................................................................................................................... 360

4.1.1. Hobbes: the covenant and rational criticism........................................... 363
4.1.2. Hume: natural inclinations and rational criticism .................................... 365
4.1.3. Nietzsche: the autonomy of values.......................................................... 368

4.2. DENNETT’S ANTHROPOLOGY OF MORAL LIFE .................................. 371

4.2.1. Responsibility, free will and decision-making: a compatibilist account........ 373
4.2.2. Underdetermination, conversation-stoppers and metaethics..................... 408
4.2.3. Radical re-evaluations and attributions of moral meaningfulness............... 418

CONCLUSION.................................................................................................. 424

BIBLIOGRAPHY ............................................................................................. 431
**RÉSUMÉ**

L’“anthropologie de la vie morale”, ou “anthropologie morale”, consiste en une approche de la philosophie morale initiée, au sein de la tradition analytique, par Peter Strawson, et développée, de façons différentes et indépendantes, par David Wiggins ainsi que par Daniel Dennett. Je tiens les anthropologies morales respectives de Wiggins et de Dennett pour complémentaires, et je propose leur synthèse au sein d’un cadre doctrinal dennettien.

Le cadre doctrinal en question inclut la définition d’un langage “rationellement acceptable”. Les descriptions et comptes rendus énoncés dans ce langage sont interprétés ontologiquement à la lumière de l’ontologie de Dennett, et les énoncés candidats au statut de connaissance sont évalués selon son épistémologie, dont j’affirme qu’elle inclut la thèse de l’“anthropocentricité”. Cette thèse, également défendue par Wiggins, confère aux comptes rendus philosophiques auxquelles elle est directement liée, un caractère de validation. Aussi les anthropologies morales respectives de Wiggins et de Dennett valident-elles toutes deux, en grande partie, la vie morale ordinaire.

L’anthropologie morale montre comment la constitution dispositionnelle de l’espèce humaine sous-détermine (c’est-à-dire conditionne et contraint, sans pour autant déterminer) les standards de correction par référence auxquels nous
évaluons moralement la conduite, les sentiments et les jugements, y compris les jugements portant sur la "signification morale".

L’anthropologie morale de Wiggins propose une théorie largement humienne de la nature humaine, ainsi qu’une description pénétrante de la moralité, et des préoccupations, motifs, buts, besoins, aspirations et expectatives “inaliénables” qui y sont attachés, et qui en tant que tels la valident.

L’anthropologie morale de Dennett propose une théorie évolutionniste de la nature humaine, et la relie à un compte rendu compatibiliste de la responsabilité morale, du libre arbitre, et de la prise de décision morale. À propos de cette dernière, Dennett fait valoir que, compte tenu de notre condition d’êtres limités mais engagés, certaines maximes imposent un terme à la délibération. J’avance que ces maximes jouent dans l’anthropologie morale de Dennett un rôle similaire à celui joué par la “ sphère de l’inaliénable” dans celle de Wiggins.

À la fin, j’affirme avoir montré comment intégrer les anthropologies morales respectives, et complémentaires, de Wiggins et de Dennett au sein d’un cadre doctrinal dennettien.
ABSTRACT

The “anthropology of moral life”, or “moral anthropology”, is an approach to moral philosophy which I take to have been initiated by Peter Strawson, and developed, independently and in different ways, by David Wiggins and Daniel Dennett. I take the respective moral anthropologies of Wiggins and Dennett to be complementary, and I propose to synthesize them within a Dennettian framework.

The framework involves the definition of a “rationally acceptable language”. Descriptions and accounts stated in that language are ontologically interpreted in the light of Dennett’s ontology, and the knowledge claims made in the language are assessed in the light of his epistemology, which I take to include a “thesis of anthropocentricity”. That thesis, also propounded by Wiggins, confers a vindicatory character to those philosophical accounts to which it is directly related. Thus, both Wiggins’ and Dennett’s respective moral anthropologies have a strong vindicatory character in regard to common moral life.

Moral anthropology shows how the dispositional constitution of the human species “underdetermines” (i.e. conditions and constrains, but does not determine) the standards of correctness by reference to which we morally assess conduct, sentiments and judgments, including judgments about what is “morally meaningful”.

vi
Wiggins’ moral anthropology proposes a largely Humean theory of human nature, as well as an insightful description of morality, and of the “unforsakeable” concerns, motives, purposes, needs, aspirations and expectations that are attached to it, and which as such vindicate it.

Dennett’s moral anthropology proposes an evolutionary theory of human nature, and relates it to a compatibilist account of moral responsibility, free will, and moral decision-making. Regarding the latter, Dennett emphasizes, given our predicament as limited but committed beings, the importance of deliberation-stopping maxims, which I take to play in his moral anthropology a role similar to that played by “the sphere of the unforsakeable” in Wiggins’.

In the end, I claim to have shown how to synthesize Wiggins’ and Dennett’s respective, and complementary, moral anthropologies within a Dennettian framework.
ACKNOWLEDGEMENTS

My deepest thanks go to my supervisor, David Davies, and my co-supervisor, Jeff Speaks, for the competence and the patience with which they guided this project and supported my endeavors.

I wish to thank, for all their help and kindness, the members of the administrative personnel who have worked at the philosophy department during those long years: Barbara Winston, Angela Fotopoulos, Mylissa Falkner, Saleema Nawaz, and in particular Claudine Lefort.

Merci aux amis qui m’ont apporté leur appui tout au long de ce parcours, et avec qui j’ai pu discuter du contenu de cette thèse, notamment: Gerry Beaulieu, Daniel Derome, Frédéric Fournier, et Daniel Hébert.

Enfin, un grand merci à ma famille.
This dissertation presents and defends a distinct approach to moral theory, which can be characterized as a naturalist, genealogical phenomenology of moral life, and which I call, following David Wiggins, the “anthropology of moral life”, or “moral anthropology”.

Moral anthropology describes the essential features of ordinary moral life as involving a potentially species-wide “ordinary morality” with both a deontological and a valuational dimension. It accounts for human moral life (and “ordinary morality”) by tracing its relation to human nature, including notably certain moral dispositions, as well as free will. This account is not only explanatory, but also vindicatory. That is, certain essential features of moral life — of our ordinary morality — are explained by their origins in certain species-wide dispositions, and vindicated by the idea that these dispositions are in some sense (following Wiggins again) unforsakeable.

This approach to moral theory is set within a non-reductionist framework, according to which adequate standards of correctness (notably moral standards) need not be grounded in principles independent from the dispositional constitution of the human species. This framework also involves a metaphysics of experience according to which no logical or metaphysical gap intervenes.
between the perceiving and judging Subject, and the Object which she perceives and of which she judges.

Moral anthropology is not a common approach within analytic philosophy. However, a partial version of what I just described is to be found in the respective works of Peter Strawson and of Daniel Dennett, and a full-fledged version is to be found in the work of David Wiggins\(^1\).

The respective moral anthropologies of those three philosophers share importantly similar traits. Their respective doctrinal frameworks also have this in common that they are opposed to reductionism. Still, there are important differences between the respective doctrinal frameworks of those three philosophers — notably between those of David Wiggins and of Daniel Dennett, on whom the present dissertation focuses. Those divergences can be summarized as follows.

Dennett implicitly defines, for purposes of “serious, literal” descriptions and accounts, what I call a “rationally acceptable” language, distinct from ordinary language. As I read Dennett, his rationally acceptable language is characterized

\(^1\) Wiggins is the only one who uses the expression “moral anthropology”. He does not use it systematically, however, even though he sometimes presents his philosophy of value in a programmatic way.
by four qualities — three constraints and one “opening” (which will be listed shortly). Dennett’s framework also includes an ontological view (which states how the language is to be ontologically interpreted) and an epistemological view (which states how the knowledge claims formulated in the language are to be assessed.) His moral anthropology, then, is couched in his rationally acceptable language, as interpreted by the attendant ontology and epistemology.

I contend that Dennett’s rationally acceptable language is adequate, and that adequate descriptions and accounts must be couched in it. Specifically, I impose as an adequacy condition on moral anthropology that it be couched in Dennett’s rationally acceptable language. The latter is significantly different from the rationally acceptable language traditionally defined by reductionist doctrines. It will be helpful to list already what I take to be the four defining qualities of Dennett’s rationally acceptable language, even though the definition of the key terms featured in my formulations of these qualities (e.g. the term “skyhook”) will have to wait until Chapter Three. The first three qualities are constraints, the fourth is an “opening”.

First, it excludes a kind of concept which Dennett calls “skyhooks”.

Second, it excludes skyhook-inducing “essentialist thinking”, and substitutes for it “gradualist thinking”.

Third, it is regimented into four irreducible levels of discourse: the physical; the functional; the intentional; and the personal/moral. The interrelations between the four levels of discourse, and the relations between their respective subject matters, are characterized sufficiently to account for their irreducibility, and to
provide credible answers to the “placement question” as it concerns those subject matters, i.e. how they relate to the physical world, as the subject matter of the physical sciences.

Fourth, it is open to “non-referential terms”. That is, it admits of terms which do not purport to refer to material objects. Moreover, it admits of non-referential vocabularies, and considers acceptable that a whole level of discourse might be non-referential in this sense.

A moral anthropology couched in Dennett’s rationally acceptable language is thus skyhook-free and gradualist (and so anti-essentialist), and “places” morality in the physical world by the characterization of the relations between the four levels of discourse.

Wiggins’ framework, for its part, is not based on the definition of a rationally acceptable language distinct from ordinary language, and so does not impose at the outset constraints on the rational acceptability of descriptions and accounts. It is also explicitly opposed to reductionism. Wiggins’ framework is centered on a Davidsonian theory of interpretation, which states two things: norms for the intersubjective practice of interpretation and the critical assessment of judgments; and five marks of truth. The theory, then, tells us how to interpret and assess ordinary discourse, and how to recognize true judgments as such. At the outset, no “rational acceptability” constraints are defined (besides the principles of interpretation itself), no kind of concept is excluded at the outset (however abstractly defined), and there are no constraints regarding the
regimentation of areas of discourse into ordered levels of discourse, such that the characterization of their relations, and of the relations between their respective subject matters, would automatically provide satisfactory answers to the placement question.

I find Dennett’s framework more satisfactory than Wiggins’, principally because I share his view of a rationally acceptable language. Overall, his doctrinal framework is clearer than Wiggins’, and it provides more straightforward answers to important questions such as the placement question.

I find no aporia in Wiggins’ framework, and I think that, capitalizing as it does on interpretation and on critical assessment in the intersubjective practice of giving and asking for reasons, it can exclude, on a piecemeal basis rather than by specification at the outset, whatever concept proves to be unacceptable; and it can adequately answer the placement question regarding various subject matters. But it remains an extremely abstract, and rather vague framework, and the answers it provides to our queries are not as straightforward as those that are provided by Dennett’s.

I take Dennett’s moral anthropology to be largely adequate, but incomplete, as it focuses on just a few features of moral life, namely moral responsibility, personhood, free will, and moral deliberation and decision-making. Moreover, Dennett does not engage in in-depth reading of classical philosophers — at least not as regards their moral theories.
By contrast, Wiggins’ moral anthropology is quite elaborate. It is rich with deep insights, and features in-depth critical reading of many classical works of philosophy—first and foremost that of Hume, but also those associated to the consequentialist tradition, to “deontological theory”, and to “virtue theory”.

I contend that the respective moral anthropologies of both philosophers are complementary, in a way which calls for their mutual synthesis. I contend that Wiggins’ moral anthropology is compatible with Dennett’s framework; and I propose to “import” the former into the latter, and to effect in this way a “synthesis” of the respective views of these two philosophers.

Peter Strawson has proposed something of an “intermediary” framework. Strawson specifies certain constraints on the rational acceptability of discourse, which imply the rejection of both reductionist doctrines (which have an overly restricted conception of rational acceptability) and of certain specific theoretical traditions, such as intuitionism or libertarianism (which have too lax a conception of rational acceptability). Mostly, Strawson has been a pioneer of analytic moral anthropology. In the threefold alignment of philosophers that I propose, Strawson is the initiator. We will return to Strawson shortly.

In this Introduction, I will first present a generic sketch of central “doctrinal components” of moral anthropology, after which I will reiterate my “Dennettian” position. Then I will present the content of the chapters to come. Finally, following
Wiggins, I will briefly discuss one feature of moral life which moral anthropology is the only approach to adequately capture: moral meaningfulness.

Moral anthropology involves four “doctrinal components”, which I will now define. Three of them, in particular, are central to the respective moral anthropologies of all three aforementioned philosophers. The sketch that follows corresponds to an extremely rapid tour around a hermeneutic circle that the dissertation will cover throughout its four chapters. Hopefully, the numerous promissory notes featured in this generic portrait will be fulfilled in the body of the dissertation.

What I call the thesis of the underdetermination of common life and of the ordinary outlook by human nature (or simply “the underdetermination thesis”) has to do with the relation of human nature, dispositionally construed, to common life and the “ordinary outlook” (which is constituted, basically, by ordinary language, common sense, common stocks of knowledge, as well as by what the phenomenologists call the “natural attitude”). The ordinary outlook includes standards of correctness. That is, humans critically assess their judgments, actions, and emotions in the intersubjective practice of giving and asking for reasons (or “rational criticism”), by reference to certain standards of correctness (which can be epistemic, moral, practical…) The contention of the thesis of underdetermination is as follows.
Although human beings enjoy considerable behavioral and cognitive autonomy, and behave and make judgments in pro-active ways, they are endowed with *natural dispositions* (behavioral, cognitive and emotional), which *condition and constrain, but do not determine*, action, emotion and judgment. The same dispositions *condition and constrain, but do not determine*, the *standards of correctness* by reference to which humans assess actions, judgments and emotions (i.e. the standards to which the latter are *answerable*). It is in this sense that dispositions *underdetermine* common life and the ordinary outlook.

The underdetermination, by human nature, of the *standards of correctness* by reference to which action, judgments and emotions are assessed, is a phenomenon that is probably easier to circumscribe than the underdetermination of action, judgments and emotions themselves. In the present dissertation, the focus will be on the underdetermination, by dispositions partaking in human nature, of *moral standards*. In this dissertation, I mostly speak of how certain dispositions which partake in human nature are *manifested* in action, emotion and judgment; and how they *underdetermine the standards* to which the latter are answerable.

The underdetermination thesis can be characterized as “naturalist”, but such a characterization is not very informative, given the vagueness of the “naturalist” doctrinal label. Although not all “naturalist philosophers” would endorse the underdetermination thesis as formulated here, my contention is that Strawson, Wiggins and Dennett would. As we will see in Chapter Two, Wiggins argues that Hume defended such a thesis.
The thesis of the underdetermination of moral standards by human nature is linked to the insight that morality does not come down to moral discourse (and so the investigation of moral life does not come down to the analysis of moral discourse), even though language is a necessary component of morality.

The second component is an antireductionist doctrine. Moral anthropology is antireductionist, in the sense that it considers that adequate descriptions and accounts of moral life cannot be articulated in non-moral terms. It must make use of a moral vocabulary, which itself presupposes an intentional vocabulary.

As it will appear, Strawson, Wiggins and Dennett defend different forms of antireductionism. They are all opposed to the reductionist view according to which "serious, literal" discourse about the world (and our relationship to it) is to be entirely articulated in a regimented language limited to two levels of "description and explanation", namely the physical and the functional.

First, I will briefly sketch the reductionist doctrine. Second, I will present what I call the "distinct languages" antireductionist argument, to which I take Strawson and Wiggins to adhere. Finally, I will explain what Dennett retains from that antireductionist argument, what he sets aside, and I will highlight some distinctive traits of his antireductionism.

---

2 I sometimes use the expression "non-reductionism" to characterize Dennett's doctrine, but I do not make a systematic use of that expression.
During the 1960’s, although the quest for a criterion of cognitive meaningfulness is abandoned, the heirs of logical empiricism retain the idea that “serious, literal” descriptions and accounts require a regimented language, distinct from ordinary language. Thus, the physical sciences use a regimented physicalist language to articulate descriptions and explanations of the constitution and behavior of material objects, and those accounts are superior to those of “folk physics”. The logical properties of the physicalist language (together with its regimented character) allow us to keep track of what is being talked about, and they make it possible, by the same token, to eliminate “empty terms”, i.e. terms which fail to successfully refer to anything that satisfies credible reality criteria.

The rise of functionalism in the philosophy of mind, during the 60’s, establishes a second level of description and explanation, over and above the physicalist level, and (by and large) irreducible to it, viz. the functionalist level\(^3\), which bears the hope of accounting for what the physicalist vocabulary, by itself, does not seem able to account for. That notably includes biological phenomena, and mental phenomena. The functionalist level may not be (entirely) reducible to the physicalist level, but it is related to it. Although accounting for that relation proves to be difficult, most advocates of functionalism consider that functional items and phenomena are involved in causal chains, and readily fit in the world described and explained by the physical sciences.

\(^3\) “Reducibility” and “irreducibility” are multi-faceted properties, and the extent to which the functionalist level of description and explanation is “irreducible” to the physicalist is a matter of debate. The point here is that the idea that the functionalist level could be entirely reduced to the physicalist level has rapidly become a minority option.
The result, then, is a two-level regimented language, which aspires to the elimination of empty terms, and which is closely associated to science. Roughly speaking: the physicalist level is that of the physical sciences; the functionalist level is that of the life sciences, and is also supposed to be that of “scientific psychology”, and of a scientifically credible philosophy of mind.

The scope of the regimented language has no clear limits, and universalist aspirations. If “scientific psychology”, together with scientifically credible philosophy of mind, can allow us to account for linguistic meaning and understanding; if the same disciplines can allow us to account for human conduct as something intelligible to humans; and if they can allow us to account for moral phenomena; then, the deployment of the regimented language will demonstrate the universality of science, while qualifying the thesis of the unity of science: it might be possible to identify science to one regimented language, but that language has more than one level of description and explanation, and methodological rules may vary between levels.

The regimented language, then, together with the theories — the descriptions and explanations — whose articulation it makes possible, would enjoy primacy over ordinary language, and over the multiple “folk theories” which partake in the ordinary outlook that we entertain in common life. If the scope of the regimented language — the scope of the sum of all the theories it can articulate on its two conceptual levels — can be shown to be universal, then the regimented language can be said to enjoy universal primacy over ordinary language (i.e. scientifically credible accounts enjoy primacy over “folk” accounts,
couched in ordinary language). In regard to the ordinary outlook, then, such a doctrine is essentially *revisionary*.

This doctrine, then, entertains the following reductionist projects: that of accounting for intentional phenomena, and for human conduct as a whole (ordinarily accounted for by belief-desire “folk” psychology) in non-intentional (specifically, functional) terms; and that of accounting for morality in non-moral terms (e.g. functional-psychological and sociological terms). However, it is important to acknowledge that most advocates of such “reductionist” projects consider that the functionalist level of description and explanation, although related to the physical, is not reducible to it, or at least not entirely.

After WWII, an antireductionist doctrine has stemmed from Oxford, appearing as a Wittgensteinian take on traditional hermeneutical arguments, which radically distinguish the causal explanation of physical phenomena on the one hand, and the interpretative understanding of human, rational agency on the other. I propose to reconstruct and to schematize this antireductionist picture as what I call the “distinct languages” argument.

With this reconstruction I mean first and foremost to capture the generic form of an argument that both Peter Strawson and David Wiggins make use of (as we will see in Chapters One and Two) in their respective defenses of moral anthropology. At the same time, I mean to show that in this connection, their

---

4 For example, in the view of Wilfrid Sellars, science has a privileged status, following which the “manifest image” (the ordinary outlook) is to be revised in the light of the “scientific image”.

12
respective argumentations are variations on a generic antireductionist picture that
came out of Oxford after WWII. A historically rigorous reconstruction would
require of course more references. As it stands, my reconstruction is partly based
on Walter Cerf’s general review of the six books that composed, as of 1962, the

The “distinct languages” argument is directed against the idea that a
regimented, scientifically acceptable language could have a universal scope, and
therefore enjoy primacy (in the sense specified above) over ordinary language, in
particular over the areas of ordinary language that constitute what could be called
a “moral-psychological language”, suitable for the description, explanation, and
assessment of human conduct (more on this shortly).

According to the antireductionist argument, the moral-psychological
language is irreducible to the “scientific language”, because it expresses
concerns and carries out tasks that cannot be expressed and carried out by the
scientific language (which expresses different concerns, and fulfills different
functions). One way to put it is to say that these are two distinct languages bear
on distinct subject matters. Let us see more precisely how this goes.

We can individuate, within ordinary language in general, two areas which
seem to work together: the intentional-psychological area, and the moral area.
First, let us acknowledge that each vocabulary can be analytically individuated.
Thus, the intentional-psychological area of ordinary language bears on human conduct, responsive as it is to rational principles, which make it intelligible. In other words, human beings act for reasons, and human conduct is ordinarily, and philosophically, made intelligible and accounted for, by identifying the reasons that rationalize it. The use of the intentional vocabulary is, to a certain extent, organized into an “ordinary psychology” — a folk theory that trades in rational-purposive descriptions and explanations of human agency, and in which (so the argument goes) causality plays no significant role.

The moral vocabulary may also be analyzed on its own, as it bears on the character and conduct of human beings as something morally assessable.

Now, our ordinary psychology is, to a certain extent, a moral psychology. That is, not only we tend to make simultaneous use of the intentional and of the moral vocabularies, but they come to merge at certain junctures. For that reason, they may be treated (at least in some contexts) as constituting a “moral-psychological language”, bearing on the rational, intelligible and morally assessable conduct of the free, morally responsible agents that we are.

On the other hand, there is a “scientific language”. It is, at basis, a physicalist language, used to describe and explain phenomena, much of which come down to causal processes: scientists establish correlations between distinct

---

5 Moral sentiments, and the concepts we use to express them, are typical moral-psychological items. This appears clearly in Strawson's analysis of them in (Strawson 1968 [1962]), as we will see in 1.2.
event-types, and propose to subsume them under laws. To the extent that the correlations are thought to correspond to causal connections, the laws under which event-types are subsumed are causal laws, and the behavior of material bodies is causally explained.

The event-types and material bodies described by the scientific language may belong to the living world as well as to the inanimate world. Thus, the scientific language includes within its scope event-types which occur in the central nervous system of human beings, including those which cause bodily movement.

The “scientific language”, as depicted in the argument, is basically a physicalist language. Eventually, the critics of reductionism have acknowledged the establishment of a second level of description and explanation, but that does not radically change the argument.

Each of the two languages just described (the moral-psychological and the scientific) has a vocabulary of its own, and is answerable to standards and principles of its own. Each of the two languages can be said to have a subject matter of its own, to the extent that each language expresses distinct concerns, and fulfills distinct functions. This view is supported by the “logical connection argument”, which can be summarized as follows. (1) There is a “logical connection” between an action and the reason that rationalizes it and makes it intelligible; (2) by contrast, the notions of “cause” and “effect” are logically distinct, and talk about causes and effects has to keep them logically distinct; therefore
(3), human action does not lend itself to causal explanations (where some physical state or event would be the cause, and action the effect).

In sum, each language has its conceptual and logical limits, and a limited scope. The undeniable growth of scientific knowledge does not justify the universalist aspirations of reductionist philosophers. When it comes to accounts of human agency, “human action” might appear as a unitary subject matter describable in two distinct languages, one scientific and rigorous, the other ordinary and deficient. But that is to neglect the sense in which the two languages can be said to bear on distinct subject matters.

The relation between the two languages and their respective subject matters comes down to this: what the scientific language describes and explains are necessary conditions for the existence of rational, free beings with a mental life, conducting themselves in their rational, intelligible, morally assessable ways.

In regard to the universalist aspirations of reductionist philosophers, the scientific language is overly restricted. Attempts at describing and explaining human mental life and human agency with the scientific language can only be inadequate, and deliver in the end a worldview impoverished to the point of being literally absurd.

As we will see, this argument from the distinctness of languages is used by Strawson and by Wiggins in their respective defenses of moral anthropology.
Considering my brief description of Dennett’s framework at the outset of the present Introduction, it is already visible that his antireductionism is not to be assimilated to the argument just described. Dennett recognizes the conceptual and logical limitations of the physicalist and the functionalist levels of discourse, as well as the limits of their respective scopes. Talk about the behavior of “intentional systems” (we will see in Chapter Three what that means) requires an intentional vocabulary; and talk about morality requires a moral vocabulary. Now, Dennett advocates the use of a rationally acceptable language, distinct from ordinary language, regimented into levels of discourse, whose interrelations are minimally characterized. What he proposes is to add, over and above the functionalist level, an intentional level of discourse; and over and above the latter, a properly moral level of discourse. The intentional and the moral levels must not be conflated. Moreover, the moral level of discourse is not “scientifically acceptable” so much as — more simply — “rationally acceptable”.

My position, as I will make clear, is that the polarized form of the antireductionist doctrine, as propounded by the “distinct languages” argument, does not adequately answer the placement question, while Dennett’s non-reductionist framework does answer it adequately.
The third component of moral anthropology is the doctrine of *anthropocentric standards* (or “anthropocentricity thesis”), which contrasts with the doctrine of cosmic standards⁶.

First, I will briefly sketch the doctrine of cosmic standards. Then I will sketch the doctrine of anthropocentric standards.

According to the doctrine of cosmic standards, all adequate *standards of correctness* are ultimately grounded in “extra-human principles”, by which I mean principles independent of the dispositional constitution of the human species, as well as of the human predicament in general, which notably includes our limited life span.

A non-exhaustive list of extra-human principles would include apodictic truths, necessary principles of the transcendental sort, logical principles, analytic truths, and principles of objectivity linked to the physical sciences. Following Wiggins, I will call standards which purport to be grounded in extra-human principles “cosmic” standards, and I will call the “cosmic standards view” that according to which adequate standards are ultimately grounded in such principles. According to this view, scientifically acceptable descriptions and explanations are in principle answerable to such standards. Like the reductionist doctrine, this view has revisionary implications, with a specific focus on standards. Thus, “ordinary standards” (standards of reasonableness, of

---

⁶ Here, I elaborate upon a contrast to be found in the works of Wiggins. See notably (Wiggins 2002 [1976], p.119 ff.)
knowledge, of morality, of beauty, of affective experience...) must be replaced by “cosmic” standards. In the case where we cannot find cosmic standards that human thought and action can actually live up to, we face scepticism.

When the doctrine of “cosmic standards” is combined with to the reductionist view, it delivers a kind of reductionist doctrine according to which the descriptions and explanations articulated at the functional and physical levels are answerable to “cosmic standards”. Note that not all forms of reductionism are linked to the “cosmic standard” view.

Against that, the doctrine of anthropocentric standards states that “anthropocentric” standards of correctness, i.e. standards grounded in the general predicament of the human species, may be adequate. The anthropocentricity thesis thus dismisses the sceptical problem conjured up by the quest for “cosmic standards”. We may refer to adequate, anthropocentric standards to critically assess various ordinary judgments, notably moral judgments, affective judgments and judgments of secondary qualities. Specifically, moral standards are anthropocentric standards.

The anthropocentricity thesis is not as such antireductionist, but as it goes against the cosmic standard view, it goes against certain forms of reductionism.

The anthropocentricity thesis also excludes relativism, i.e. doctrines to the effect that standards are necessarily relative to cultures, or even to individuals.
Now consider the framework composed by the underdetermination thesis and the anthropocentricity thesis. As applied to moral life, it delivers the view that human nature underdetermines the standards of ordinary morality; and that these standards are anthropocentric, and adequate as such.

The anthropology of moral life, based as it is on those two theses, has a vindicatory character. That is, it not only accounts for ordinary morality, it vindicates some of it. Ordinary, anthropocentric standards may be adequate not in spite of being grounded in nothing deeper than human nature, but in virtue of being grounded in something as deep as human nature.

The fourth doctrinal component of moral anthropology is a metaphysics of experience according to which our experience of the world is “direct”, in the sense that no logical or metaphysical gap intervenes between the perceiving and judging Subject and the Object which she perceives and of which she judges.

In regard to moral life, the metaphysics of experience states that we directly experience genuine moral properties of items (characters, human conduct, states of affairs …). The ontological status of moral properties can be construed as being at once objective and subjective, because their existence depends on species-wide human dispositions attuned to their perception. On this count, moral properties are like affective properties (disgustingness, amusingness, etc.) and secondary qualities.

I count Strawson (who was inspired by Kant) as one of the pioneers of this sort of “metaphysics of experience”. Wiggins also defends such a metaphysics,
and acknowledges its historical dimension. For Wiggins, as we will see in 2.3.1.,
the metaphysics of experience is linked to a metaethical theory, as it supports his
view that a certain class of moral judgments are good candidates for truth.

Now, I contend that Dennett offers the most rigorous treatment of the
underdetermination thesis, of the antireductionist view and of the
anthropocentricity thesis attached to it, because he shapes those doctrinal
components with the constraints of his own rationally acceptable language. He
also proposes a metaphysics of experience which I take to be adequate, but
which will not be discussed in this dissertation, because it is not crucial to his
moral view.

The four doctrinal components are also treated in interesting ways by
Strawson and by Wiggins, as will appear in the chapters that are dedicated to
them. In particular, Wiggins introduces, in relation to his treatment of the
underdetermination thesis, the very important notion of the *unforsakeable*. But let
us survey in order the content of the four chapters that make up this dissertation.

Chapter One exposes Strawson’s anthropology of moral life. This is
preceded by a presentation of some central features of Strawson’s overall
philosophical framework, which constitute the doctrinal “skeleton” of his moral
anthropology.
In Chapter Two, I show (2.1.) that all four doctrinal components described above can be found in the philosophy of David Wiggins. Regarding the doctrine of anthropocentric standards, Wiggins polarizes the opposition between on the one hand those reductionist doctrines which adopt moreover the “cosmic view”, and on the other hand antireductionist doctrines, to which he ties the anthropocentric thesis. Wiggins describes their opposition with a spatial metaphor: whereas reductionism (cum the cosmic view) constitutes an “external” standpoint on the world and on the ordinary outlook, antireductionist doctrines like his own constitute an “inner view” on the same things. He stresses that “external standpoint” doctrines usually take the scientific discourse as enjoying primacy over the ordinary outlook, thus implying that the latter must be revised in light of the former; while antireductionist cum anthropocentrist doctrines (like his own) have a strong vindicatory character, and dispute the idea that scientific discourse and its restricted language (that defined by reductionist doctrines) have a universal scope and a preeminent status.

I show (2.2.) that Wiggins’ doctrinal framework is fully applied to moral life, thus issuing in a self-conscious moral anthropology. Wiggins adopts Hume’s theory of human nature (2.2.1.), and adapts it to draw it closer to the moral anthropology he envisages. He also acknowledges that the Humean theory is to be completed by evolutionary theory. Wiggins then focuses (2.2.2.) on certain dispositions (or, as he prefers to say, “concerns” and “motives”), which he characterizes as *unforsakeable*, and from which the essential features of moral life originate. Wiggins’ defense of a libertarian conception of free will (2.2.3.)
completes the picture. Wiggins argues that our ordinary moral standards are by and large adequate, to the extent that they originate from unforsakeable concerns and motives.

I present Wiggins’ metaethical theory in 2.3., and I show (2.3.1.) how Wiggins’ metaphysics of experience supports the “weak cognitivist” position delivered by his metaethical theory. I also present his criticism of rival views (2.3.2.). These are essentially reductionist doctrines, which Wiggins indicts on the count that they deliver worldviews impoverished to the point of absurdity — notably moral absurdity.

In 2.4., I present my own take on Wiggins’ work. I contend that his description of ordinary morality, and his vindicatory scrutiny of it are largely adequate, and that the notion of “unforsakeable” concerns and motives acts like the pivot of the description and of the vindication. I make three critical points about his doctrinal framework, regarding which I contend that Dennett’s framework is better.

Chapter Three presents Dennett’s doctrinal framework (3.1.) and his evolutionary theory of human nature. The latter is articulated in the rationally acceptable language defined in the former (and interpreted in the light of the attendant ontology and epistemology, which includes the anthropocentricity thesis). I present Dennett’s evolutionary theory of human nature with a special concern for the evolutionary history of the conditions of morality. It begins with the evolution of agency and of related phenomena, from self-replicators to primates
Chapter Four presents Dennett’s philosophy of morality. First, it presents (4.1.) Dennett’s take on three philosophers (Hobbes, Hume, Nietzsche) who propounded, in each his own way, the insight according to which values in general (and morality in particular) were brought about by language and rational criticism. Dennett’s moral anthropology proper is presented in 4.2. It includes a compatibilist account of moral responsibility, free will, and moral decision-making. Regarding the latter Dennett emphasizes the role of deliberation-stopping maxims, which I take to play in his moral anthropology a role similar to that played by “the sphere of the unforsakeable” in Wiggins’. As we will see, what Dennett has to say about the issue of moral meaningfulness is related to his view of decision-making.
My conclusion is that the respective moral anthropologies of David Wiggins and Daniel Dennett are compatible and complementary, in a way that calls for their synthesis. I show that the most adequate anthropology of moral life is obtained when we insert Wiggins’ moral anthropology in Dennett’s framework, within which they complete one another.

Now, at the beginning of “Truth, Invention, and the Meaning of Life” (2002 [1976]), Wiggins suggests that many ordinary, lay people are concerned with the issue of *moral meaningfulness*, in the sense that they wonder “What is the meaning of life?”, and readily assume that «the whole issue of life’s meaning presupposes some positive answer to the question whether it can plainly and straightforwardly be *true* that this or that thing or activity or pursuit is good, has value, or is worth something.» (2002 [1976], p.87) Wiggins suggests moreover that those ordinary people «suppose that questions like that of life’s meaning must be among the central questions of moral philosophy.» (ibid)7

Wiggins concurs: the question of life’s having a meaning *should* be central to moral philosophy, although it is not, for the reason that this question «is under the suspicion of belonging in the same class as “What is the greatest good of the greatest number?” or “What is the will?” or “What holds the world up?”. This is the

7 «Il n’y a qu’un problème philosophique vraiment sérieux: c’est le suicide. Juger que la vie vaut ou ne vaut pas la peine d’être vécue, c’est répondre à la question fondamentale de la philosophie. Le reste, si le monde a trois dimensions, si l’esprit a neuf ou douze catégories, vient ensuite. Ce sont des jeux. Il faut d’abord répondre.» (Camus 1942, p.17)
class of questions not in good order, or best not answered just as they stand.»
(2002 [1976], p.87-88) There is a tradition, Wiggins suggests, according to which the central questions of moral philosophy concern rather happiness, as well as the “good/bad” “right/wrong” contrasts, as geared to questions of action guidance (“What shall I do?”, “How should we act?”), understood in ways that are more casuistic than existential. But he claims that this tradition is mistaken.

For Wiggins, the question whether judgments of moral meaningfulness are candidates for truth is central to an inquiry into moral meaningfulness as such:

In what follows, I try to explore the possibility that the question of truth and the question of life’s meaning are among the most fundamental questions of moral philosophy. The outcome of the attempt may perhaps indicate that [...] these questions and the other questions that they bring to our attention are a better focus for ethics and meta-ethics than the textbook problem ‘What [under this or that or the other circumstance] shall I do?’ My finding will be that the question of life’s meaning does, as the untheoretical suppose, lead into the question of truth — and conversely. [...] I shall also claim to uncover the possibility that philosophy has put happiness in the place that should have been occupied in moral philosophy by meaning. (2002 [1976], p.88)

However, Wiggins’ inquiry into the question of life’s meaning remains, for the time being, fragmentary. The reason for that, I take it, is that the issue of moral meaningfulness must be tackled from within an anthropology of moral life (to which a metaethical theory, assessing whether moral judgments are good candidates for truth, is according to Wiggins an indispensable complement, but I
will dispute this statement). A moral anthropology, then, must be sketched before
the issue of moral meaningfulness can be tackled. But I concur with Wiggins that
the issue of moral meaningfulness is an important issue that has been neglected
by traditional approaches to moral philosophy, and that it should occupy as
central a place in moral philosophy as it occupies in common moral life. I contend
that the anthropology of moral life is the one approach that can do justice to that
question, to which I will return when I assess Wiggins’ moral anthropology (2.4.),
in Chapter Four (on Dennett’s moral philosophy), and in the conclusion.

Here, I want to provide a few notes on the issue of moral meaningfulness,
which help understand the originality of moral anthropology as an approach to
moral philosophy.

The question of life’s having a “meaning” involves what I describe as an
ordinary, existential, teleological construal of “the good”. Something that is good
in this sense is “morally meaningful”: it has intrinsic (non-derived) worth,
constitutes a non-instrumental finality of human action and existence, and invests
human experience with an ultimate sense of point.

Such a construal of the good, directly related as it is to the goal-
directedness of human agency and existence, could be characterized as
“Aristotelian”, but also as phenomenological. The idea is that human agency and
existence imply the self-conscious setting of goals, hierarchically related from the
most immediate and less comprehensive, up to some ultimate goal(s), which,
however vaguely defined, one sets to one’s existence as a whole, and to which
one’s inferior goals are instrumentally related. Moral anthropology, as such, is not specifically Aristotelian (although Wiggins' philosophy as a whole is deeply influenced by Aristotle), but it could be described as combining a *phenomenology* of moral life (how do we experience morality in common life?) to a *genealogy* of moral life (what are the origins of the essential features of moral life, and how do they contribute to account for them?)⁸.

There certainly is something phenomenological in the way moral anthropology brings out the dependence, not only of moral meaningfulness, but of moral life as a whole on *our ordinary sense of responsible selfhood and of freedom*. Moral anthropology brings out how moral life as a whole depends on our capacity to establish, one way or another (libertarian or compatibilist) that we are indeed morally responsible and we do have free will.

Nihilism appears as the negative side of the good as moral meaningfulness. In the case where we would conclude that our sense of free, responsible agency is illusory, it would be difficult not to interpret such a conclusion as nihilistic. That would be an “elementary” form of nihilism, by contrast with a “higher” form of nihilism, which would follow from the impossibility to find anything intrinsically worthy, and able to constitute a non-derived finality to

---

⁸ One way of putting the existential “version” of the question “what is the good?” is the question whether, for humans, there are any goals more worthwhile than self-preservation and reproduction (since all animal species have self-preservation and reproduction “programmed” as the ultimate goals of behavior, to which hierarchically inferior goals are instrumentally related).
human action and existence — what Camus called “the absurd”, and of which he found an illustration in the myth of Sysiphus\(^9\).

As the negative side of moral meaningfulness, the two varieties of nihilism that I just distinguished (the “elementary” form, and the “higher” form) are also at the center of moral life, and should be at the center of moral philosophy. Wiggins suggests as much in (Wiggins [1976]).

The “higher challenge” that moral anthropology takes up, then, is that of moral meaningfulness itself: can we find anything that is intrinsically worthwhile, and so can constitute a non-derived finality to human action and existence? If not, we face the higher form of nihilism, the absurd. A convincing answer to the higher challenge presupposes that we have a convincing answer to the elementary challenge, i.e. that we can establish that our ordinary sense of having a free, morally responsible self, is genuine.

Moral anthropology brings out the phenomenological, incarnated, and connected character of those issues. I contend that it has to be articulated in Dennett’s rationally acceptable language, so that moral meaningfulness may be discussed without involving “skyhooks” (these will be defined in 3.1.1.), and may be “placed” in the physical world. In this way, we may have a grasp of our nature, of our reasons, and of what moral meaningfulness we may find in life.

\(^9\) Richard Taylor has also discussed the myth of Sysiphus in (Taylor 1970). In (Wiggins [1976]), Wiggins briefly evokes Camus, but mostly discusses Taylor’s take on the myth.
CHAPTER 1. PETER STRAWSON: SOME ESSENTIAL FEATURES OF MORAL LIFE

Central features of Strawson’s overall framework are presented in section 1.1., namely: the underdetermination thesis, the antireductionist view, and the metaphysics of experience.

In Strawson’s framework, the underdetermination thesis appears as the view according to which human beings are endowed with a “fundamental cognitive structure” which conditions and constrains human thought, but does not determine it. This thesis is connected to an antireductionist view, to which, I contend, the anthropocentricity thesis is implicitly tied.

Although Strawson’s metaphysics of experience makes no direct contribution to his anthropology of moral life, it is presented here in schematic form, for the reason that Wiggins’ own metaphysics of experience is apparently related to it. Giving an idea of Strawson’s metaphysics, then, should help understand Wiggins’.
My presentation of Strawson’s moral anthropology (1.2.) is extracted from two complementary texts of his: “Freedom and Resentment” (Strawson 1968 [1962]) and “Freedom and Necessity” (Strawson 1992 [1983])

In Strawson’s moral anthropology, the underdetermination thesis appears as the view that human concerns widely share certain “expectations and demands” which (as I interpret Strawson) condition and constrain, but do not determine, ordinary moral standards by reference to which we assess human conduct, as well as moral sentiments and the judgments that accompany them. This appears in the context of an analytic and phenomenological inquiry into our moral sentiments and the judgments that accompany them\(^\text{11}\. More precisely, Strawson examines how certain sorts of moral sentiments (and the judgments that accompany them) are answerable to a certain standard of correctness which in turns presupposes the genuineness of our ordinary sense of a free, morally responsible self. I contend that we can say, speaking in Strawson’s name, that those features of moral life manifest dispositions which partake in human nature,

\(^{10}\) The year between brackets is the year of original publication. In the case of references without page numbers (i.e. references to titles by year of publication), I will use only the year of original publication, between brackets.

\(^{11}\) By “the judgments which accompany moral sentiments”, I refer to those judgments which bear on the same object as the sentiments. For example, witnessing a case of abuse, one might feel indignation and think “This is unacceptable!”
and that these dispositions underdetermine the standards to which the features are answerable.

Strawson examines those features of the moral life by way of a question on the potential consequences for them of the truth of determinism. Many philosophers suggest that in the case where determinism is true, science has a universal scope, and a privileged status; and that consequently, the manifest image must be revised in the light of the scientific image.

Against that sort of suggestion, Strawson opposes antireductionist arguments, but he mostly suggests that the irremovability of the ordinary standards to which our moral sentiments and our attributions of responsibility are answerable (underdetermined as they are by dispositions deeply rooted in human nature), makes idle the issue of their vindication.

In 1.3., I take a position in regard to Strawson’s work. I contend that, in spite of certain shortcomings, Strawson sets moral anthropology on the right path.

1.1. THE STRAWSONIAN FRAMEWORK
This section presents four features of Strawson’s framework, as they can be found in his *Analysis and Metaphysics: an Introduction to Philosophy* (Strawson 1992)\(^\text{12}\).

The first feature (1.1.1.) consists in Strawson’s elucidative conception of analysis, together with his “connective model” of the meaning of concepts, and his thesis that our vocabulary is (as I put it) underdetermined by a species-wide “fundamental cognitive structure”.

The second feature (1.1.2.) consists in Strawson’s view of what I call a “rationally acceptable” language and ontology. Strawson suggests that the fundamental cognitive structure fixes an “upper limit” and a “lower limit” to what is rationally acceptable in the way of language and ontology; and that the specifics are to be worked out within the intersubjective practice of rational criticism.

Strawson sets his view of rational acceptability against two rival conceptions of the same topic: those which he sees as overly liberal, and those,

\(^{12}\) This book consists of a collection of ten texts, seven of which present the content of an introductory course which Strawson gave at Oxford from 1968 to 1987. In this introductory course, Strawson in fact deploys his own doctrinal framework. Thus, under its generic “introductory” self-presentation, the book presents a synthetic view of Strawson’s own, historically influential framework: «The first seven chapters of this book preserve, virtually unchanged, the content of the first ten or eleven of the 1968 lectures, to which I made few alterations in succeeding years.» (Strawson 1992, p.vii)
associated to reductionism, which he sees as overly restricted. Strawson condemns the former more severely than the latter, to which he nevertheless dedicates much critical attention. He notably discusses Quine’s doctrine of physicalistic reductionism. Strawson indicts Quine for assimilating the idea of a rationally acceptable language to a “scientifically acceptable language”, answerable to scientific standards, to which he confers a universal scope, and a universal primacy over ordinary language. The assimilation of rational acceptability to scientific acceptability is, Strawson suggest, the fundamental mistake of scientistic doctrines.

The third feature (1.1.3.) is what I call Strawson’s metaphysics of experience. Strawson argues that the relation between Subject and Object does not involve any gap, logical or metaphysical. According to Strawson, in the basic case (where we do not have any good reason to suspect any kind of cognitive failure), the world is as it appears in experience; and sense experience delivers judgments that are true, in a world-involving way.

This goes against those doctrines (notably empiricism) according to which there is a gap between the perceiving and judging Subject, and the Objective Reality about which she judges. That sort of “gap-involving” metaphysics is associated to various forms of reductionism.

The fourth feature (1.1.4.) stems from the previous ones, and implicitly calls upon what I have called (in the Introduction) the anthropocentricity thesis:
Strawson argues that a rationally acceptable language and ontology admits of secondary qualities and secondary-quality terms, even though they are banned (according to Strawson) from a strictly physicalist language and ontology. He suggests moreover that various areas of ordinary language are answerable to standards of correctness which, although they do not have the objectivity of scientific standards, are nevertheless adequate.

1.1.1. The connective model of meaning and our fundamental cognitive structure

Strawson’s theory of meaning appears as a non-formal, antireductionist form of conceptual role semantics. According to what he calls the “connective model”, the meaning of a particular concept (or “term”) is a function of its relations to other concepts. Consequently, the meaning of concepts is 

*elucidated* by tracing their interrelations.

The elucidative approach tolerates a certain circularity in the definition of terms: as a term is defined by other terms, which are in turn defined by other terms, etc, the term being defined in the first place may very well occur somewhere down the line of the definitional process. The informativeness of the process is not ruined for that, provided that it traces an elaborate enough network of concepts:
Let us imagine [...] the model of an elaborate network, a system, of connected items, concepts, such that the function of each item, each concept, could, from the philosophical point of view, be properly understood only by grasping its connections with the others, its place in the system — perhaps better still, the picture of a set of interlocking systems of such a kind. If this becomes our model, then there will be no reason to be worried if, in the process of tracing connections from one point to another of the network, we find ourselves returning to, or passing through, our starting point. (Strawson 1992, p.19)\(^{13}\)

Elucidative analysis reveals the hierarchical structure of our overall ordinary vocabulary, as it traces the interrelations between concepts, including asymmetrical relations of dependence, which mark the hierarchy of concepts: «the ability to operate with one set of concepts may presuppose the ability to operate with another set, and not vice versa. In this case we may say that the presupposed concepts are conceptually prior to the presupposing concepts [...]» (1992, p.21)

As the rules of ordinary language include the conditions of the correct use of concepts, of their presuppositions and entailments (given the approximate place of each concept in the network of the overall vocabulary), the elucidation of concepts partakes in the explicitation of the rules of ordinary language. The explicitation of those rules constitutes a “commentary”, which stands to ordinary language as an informal metalanguage to its object-language (Strawson 1995 [1949], p.196).

\(^{13}\) For the rest of this chapter, Strawson’s name will be omitted from references to his work.
For Strawson, *those concepts which stand at the basis of our vocabulary* have great theoretical importance, as he takes their articulation to reflect the articulation of the *fundamental cognitive structure* that all humans are endowed with: «A concept or concept-type is basic in the relevant sense if it is one of a set of general, pervasive, and ultimately irreducible concepts or concept-types which together form a structure — a structure which constitutes the framework of our ordinary thought and talk [...]» (1992, p.24; emphasis added)\(^{14}\)

What are those basic concepts? Typically, they are much more general than the ordinary terms that we use in ordinary talk. That is, they are typically generic — but irreducible — terms, such as "body" or "material object", which entail the more specific terms that are ordinarily used:

We should not ordinarily call a chair or a mountain a "body". We do not ordinarily have occasion to use an expression of quite such a general application as the philosopher makes, or used to make, of this word. But this difficulty is easily resolved. If a philosopher claimed that the concept of "body" was basic in our conceptual structure, his claim could be understood as a kind of shorthand for the claim that it was a basic feature of our conceptual structure that it contained a range of concepts of a certain general type, namely, concepts of different kinds of body; and he could maintain this consistently with admitting that we

\(^{14}\) Strawson uses, in an apparently interchangeable way, many different expressions to designate the fundamental cognitive structure.
ordinarily had no occasion to make use of so comprehensive a classification. (1992, p.23)

Although Strawson does not put his hypothesis in terms of dispositions, one gathers that the fundamental cognitive structure that he speaks of can be construed as a species-wide set of cognitive dispositions (thus partaking in the general cognitive constitution of the human species). As such, it enables human thought and discourse, but also sets limits and constraints on it. In other words, the fundamental cognitive structure conditions and constrains human thought and discourse, although it does not determine them. In yet other words, the fundamental cognitive structure underdetermines human thought and discourse.

It is already apparent how the elucidative and the reductionist forms of analysis differ from one another. According to the reductionist form, all legitimate concepts are in principle definable in physicalist terms, or at least (in newer versions of the reductionist doctrine) in functionalist terms. This makes it in principle possible to reduce all significant sentences, and classes of sentences (and so “areas of language”) to the physicalist, or at least the functionalist, level(s) of description and explanation.

In Strawson’s connective model, by contrast, basic concepts do not constitute a topically unitary vocabulary, bearing on physical items and phenomena (his examples of “bodies” and “material objects” notwithstanding); and, as explained above, the existence of asymmetrical relations of dependence
between certain concepts does not exclude from the definitional process an acceptable form of circularity.

1.1.2. A rationally acceptable language and ontology: Strawson vs Quine

The elucidative approach to analysis involves the explicitation of the rules that underlie the various areas of language. Beyond that, Strawson proposes to investigate how language and thought, together with the fundamental cognitive structure which underdetermines them, relate to reality. In the process, he proposes a rationally acceptable language and ontology.

Strawson’s view is set against a critical discussion of both those doctrines whose conception of a rationally acceptable language and ontology he finds overly liberal, and those whose conception of the same thing he finds overly restrictive.

In an early article on intuitionism ([1949]), Strawson already criticizes both logical empiricism for its overly “tough-minded” conception of a rationally acceptable language and ontology, so restricted that it negates the significance of moral judgments; and intuitionism, for its overly “tender-minded” conception of language and ontology, so liberal that it comes down to supplementing facts with
“fairy tales”. Of those two doctrines, Strawson considers that the former is less inadequate than the latter:

This descriptive vagueness [of moral judgments] is no defect [...] But by being mistaken for something more than it is, it leads to one kind of metaphysics: the metaphysics of substance (the thing-in-itself), or of intuited unanalyzable ethical characteristics. And by being ignored altogether, it leads to another kind of metaphysics: the tough metaphysics of translation, the brutal suggestion that we could get along just as well without the ethical language. Neither metaphysics — neither the tender metaphysics of ultimacy, nor the tough metaphysics of reduction — does justice to the facts: but the latter does them less injustice; for it doesn’t seek to supplement them with a fairy-tale. (1995 [1949], p.196; emphasis added)

The problem, then, with overly liberal conceptions of rational acceptability is that they suggest that ordinary language, and ordinary ontological interpretations of it, are adequate as they stand, even for purposes of “serious, literal” accounts of reality.

As for the overly restrictive conceptions of rational acceptability, Strawson mostly discusses, in ([1992]), the doctrine of Willard Quine. As he remarks, his and Quine’s respective views share a common doctrinal basis. However, this common ground is minimal.

Thus, the respective doctrines of both philosophers are based on the idea that “reality” is the sum of our ontological commitments. According to Quine’s doctrine (as Strawson presents it), «we are committed to belief in the existence of
whatever kinds of things we seriously refer to, or attempt to refer to, whether we refer generally, by way of variables of quantification, or determinately, by way of names or other definite singular terms.” (1992, p.43) Moreover, both philosophers agree that those ontological commitments are objects of critical examination in the intersubjective practice of rational criticism.

However, they disagree on the principles that guide and constrain the practice, on the standards to which it is answerable, and on the views of rational acceptability and of reality that it delivers, as we will now see.

1.1.2.1. Quine’s physicalist reductionism

For Quine, a rationally acceptable language is a *scientifically* acceptable language, more precisely a physicalist language; its scope is universal; and its standards of realist interpretation (or “ontological commitment”) are highly restrictive.

According to Quine’s doctrine (as Strawson describes it), the practice of rational criticism must in the end come down to a reductionist procedure of “critical paraphrase”, defining both a rationally acceptable language, and certain standards for the realist interpretation of terms. The rationally acceptable language comes down to a scientifically acceptable language, and its realist interpretation is allowed on the basis of highly restrictive principles:
[The] procedure of critical paraphrase [...] is to be guided by two maxims. The first maxim requires that we employ only a vocabulary which is clear and scientifically acceptable; the second that we restrict our ontology to the minimum which would be theoretically sufficient for the expressions of our beliefs, even if the price in practical convenience of observing such a restriction would be unacceptably high. The second maxim might be called the maxim of ontological economy. (1992, p.45)

Strawson describes this view as a scientistic doctrine of ontological reductionism, according to which reality is to be exhaustively described and explained by a physicalist language, endowed with a supposedly universal scope, and characterized by highly restricted standards of realist interpretation (or “ontological commitment”).

Those terms (or “concepts”) which do not survive “critical paraphrase” into scientifically acceptable vocabulary are considered as “empty terms” (or empty concepts). They fail to successfully refer to anything that satisfies adequate reality criteria, and they have no scientifically acceptable use. The fact that many such terms are commonly used in ordinary talk, does not disqualify ordinary language for practical uses, but we should acknowledge that those terms are not ontologically committing:

The apparent ontological excess which goes with the proliferation of nouns and noun phrases in our ordinary talk we can then charge to the account of mere practical convenience and brevity in discourse. We need not suppose we are seriously committed to belief in the actual existence of all the items such phrases may seem to stand for. (1992, p.45)
And when it comes to serious, literal descriptions and explanations of reality, those terms will be eliminated by the critical paraphrase process.

1.1.2.2. Strawson’s view of rational acceptability

Strawson’s more liberal view of a rationally acceptable language and ontology involves the following ideas: for one thing, the “scientific language” has conceptual and logical limitations, and therefore a limited scope; for another thing, as our fundamental cognitive structure underdetermines the structure of our vocabulary, it underdetermines the hierarchy of our ontological commitments, and sets limits on our discourse and thought. Precisely, it sets an “upper limit”, and a “lower limit”. Let us take those points in turn.

The hierarchical structure of our vocabulary as a whole reflects our ontological commitments in regard to the items designated by concepts: the more basic a concept is, the more committed we are to the existence of what it designates. Ontological commitments are not an all-or-nothing affair, decided on the basis of a firm criterion (namely “surviving critical paraphrasing”). Rather, they are a matter of degree, following the hierarchical status of a term in our vocabulary, and in our cognitive structure. As a result, instead of a clear-cut distinction between what is real and what is not real, we have a hierarchy of
ontological commitments, corresponding to a hierarchy of concepts (i.e. their ordering in respect of priority):

Instead of just the saved and the lost, we might have an ontological hierarchy; instead of just the ins and the outs, an order of priority. For example, it might be agreed that attributes of properties are ontologically secondary to the objects to which one attributes them, in so far as reference to properties presupposes reference to objects, but not conversely. [...] And in general we might find reason for saying that reference to items of certain types was secondary to, or derivative from, reference to items of other types without there being any implication that the former should therefore be, as it were, expelled from the domain of existence. (p.46)

This hierarchy of concepts and of ontological commitments has an upper limit and a lower limit, both set by our fundamental cognitive structure.

The upper limit of sense is that beyond which terms are non-committing ontologically. Taking these terms otherwise would be engaging in what Strawson calls “myth-making”, or “fairy-tale metaphysics”. (This corresponds to his conception of “empty terms”.)

The lower limit of sense corresponds to the basis of our hierarchically structured vocabulary. Conferring a universal scope upon a language which excludes basic concepts is bound to entail descriptions and explanations that trespass the lower limit of sense, and thus fall into unintelligibility.

This implies that a rationally acceptable language does not admit of terms to which have no good reason to ontologically commit ourselves; and on the
other hand, it avoids excluding fundamental terms, which designate items to
which we are ontologically committed in a fundamental way.

By contrast, Quine’s reductionist doctrine shows a lop-sided concern for
empty terms. The rationally acceptable language it defines might be adequate for
the language of science, but again, that language does not enjoy a universal
scope. As it is deprived of many fundamental concepts, attempts to exhaustively
describe and explain reality with such a language can only fall into unintelligibility.

Strawson’s rationally acceptable ontology results from a critical scrutiny of
our vocabulary, and of our general conceptual structure. Instead of asking «What
are the objects of reference which survive the pressure of critical paraphrase,
conducted on severely Quinean principles?» we should ask: «What are the
fundamental types of individuals, properties and relations which characterize the
structure of our thought and what relations can be established among them?»
(1992, p.46-47; emphasis added) In this way, we shall avoid unwarranted
ontological commitments towards empty terms, as well as the sort of
unintelligibility which follows from the elimination of indispensable terms:

[…] it may seem that the drive towards ontological reduction may be
in part attributable to an unwarranted fear: the fear that the admission
to one’s ontology of certain abstract non-particular things (such as
properties, propositions, etc.) involves the risk of myth-making, of
treating such things as more like particular spatio-temporal objects
than they are. (Think of the common philosophical talk of
overpopulating the universe, etc.) The danger may be real; but it
seems to me that it can be avoided without yielding to reductive pressure. (1992, p.58)

In sum, instead of assimilating rational acceptability to scientific acceptability, we should pay attention to the grammar of our fundamental conceptual structure, and to its implications. Reality is not a set of ontologically objective objects, but a hierarchical structure of ontological commitments, corresponding to the hierarchical structure of our vocabulary, underdetermined as it is by our fundamental cognitive structure.

Strawson argues, moreover, that we have no good reasons to stipulate that the material world as we human beings experience it is solely constituted of ontologically objective items, i.e. items which do not depend, for their existence, on human mental states, as we will see in 1.1.4.. Before that, we must see how Strawson avoids the idea that there is a gap between the perceiving and judging Subject, and the reality about which she judges.

1.1.3. Strawson’s metaphysics of experience

According to Strawson (who follows a Kantian insight), the meaning of terms is determined by their use in judgments, which aim at the truth. Now, the question of how judgments can be true has been, and continues to be, a deep
problem for all those doctrines according to which the relation between Subject and Object — what Strawson calls «the duality of the judging Subject and the Objective Reality about which he judges» (1992, p.53) — involves a gap, logical or metaphysical.

According to Strawson, such a gap, and the epistemological problems with which it is associated, are artifacts of inadequate doctrines, rather than genuine philosophical problems.

On this count, Strawson mostly targets the empiricist tradition. Empiricism not only denies that sense experience is unproblematically dependent on “how things are”, it claims that concepts are empty except in relation to possible experience. This implies that there is a gap between Subject and Object, and entails sceptical doubts about our perceptual judgments:

Let us call the initial duality here “the duality of the judging Subject and the Objective Reality about which he judges”. The risk I am now speaking of, the risk inherent in the empiricist insistence [that concepts are empty except in relation to possible experience], is the risk of losing hold of this duality altogether. In insisting that experience not only bridges the gap between Subject and Object, but also gives the concepts we use all their sense of content, we run the risk of the notion of objective reality being entirely engulfed or swallowed up in that experience. Many idealisms and all phenomenalisms are the results of this engulfing; and the history of much of our epistemology, when it isn’t one of sinking into (or wallowing in) this gulf, is one of struggle to get out of it, to show that various forms of scepticism about
Objective Reality are unwarranted. But it is better to avoid the gulf altogether. (1992, p.53)

By contrast, Strawson argues that sense experience is causally dependent on the way things are. If that is right; and if the meaning of concepts is determined by their place in our overall vocabulary; and if judgments aim at the truth; then, Strawson suggests, judgments, as candidates for truth, purport to bear on the world, not on the content of experience. Given the causal dependence of sense experience of how things are, there is no good reason to deny that in the basic case, sense experience delivers judgments that are true, in a world-involving sense (in which case they “correspond”, in a minimal sense, to “how things are”):

Hence the minimum that seems to be involved in the notion of sense perception generally yielding true judgments about an objective spatio-temporal world is that there should be some pretty regular relation of dependence of the experience enjoyed in sense perception on the way things objectively are. (Otherwise the normal truth or correctness of perceptual judgments would seem to be something inexplicable, an extraordinary coincidence.) […]

Hence any philosophical theory which seeks to be faithful to our general framework of ideas, our general framework of thought, must provide for this general notion of causal dependence. It must, to this extent at least, be a causal theory of perception. (1992, p.60-61)

For Strawson, then, there is no gap between the perceiving and judging Subject, and the Objective Reality about which she judges. To the extent that the
fictional gap plays a crucial role in the genesis of scepticism, the dismissal of the former entails the dismissal of the latter. We may legitimately assume that sense perception delivers perceptual judgments that are generally true. Consequently, we are warranted to take the way things appear in experience as the way things are, unless we have good reasons to suspect that some sort of cognitive failure invalidates the process.

Thus, Strawson’s “anti-gap” metaphysics of experience and anti-sceptical epistemology bring him to conclude that “we normally perceive things as they really are”. However, in line with the “distinct languages and subject matters” argument, we must distinguish two senses of “things as they really are”: that of physical theory, and that of the ordinary outlook.

Pursuing this line of thinking, Strawson argues that there are no good reasons to consider that secondary qualities are not real, as we will see more in detail in the next subsection.

1.1.4. The rational acceptability of secondary-quality terms

According to Strawson, the scientific language which reductionist philosophers assimilate to the very idea of a rationally acceptable language admits only of primary-quality terms, and so excludes secondary-quality terms. The elimination of non-objective vocabularies (including secondary quality-terms) from the scientific language is linked to the in-principle objectivity of the
standards of correctness to which this language is answerable. According to such standards, the truth value of scientifically acceptable statements, and the ontological status of that to which they refer, must be independent from the mental states of particular subjects, and independent from the dispositional constitution of the human species. Thus, in the discourse of physical theory, answerable as it is to the standards of physical theory, «all ascription to things of qualities peculiar to a given sensory mode — as colour, felt texture, etc. — is excluded from the description of things as they really are.» (1992, p.66)

Against that view, Strawson argues for a rationally acceptable language and ontology which admits of secondary-quality terms, and of secondary qualities. His argumentation to that effect is related, not only to antireductionism as such, but also to the anthropocentric view.

That is, ordinary discourse admits of non-objective vocabularies, including secondary-quality terms, which occur in descriptions of the way things appear for human beings: «such ascription [of qualities peculiar to a given sensory mode] is not excluded; it is those ascriptions (of sensible qualities) which are associated with normal conditions of observation that are taken as the standard by which others are corrected.» (1992, p.66; emphasis added)

The standards to which secondary-quality ascriptions are answerable define what normal human responses are in normal conditions, and those normal human responses involve mental states. Such standards, then, do not have the objectivity of scientific standards. Still, secondary-quality ascriptions, answerable
as they are to ordinary, non-scientific standards, may reach beyond individual subjectivities, for two reasons.

For one thing, Strawson suggests that secondary-quality ascriptions remain below the upper limit of sense, so they command an ontological commitment on our part. He describes as «a typically obsessive reaction» the insistence on the idea that the material world as we human beings experience it is solely constituted of ontologically objective items.

Second, according to Strawson, the proper contrast of “subjectivity” is not “objectivity”, but rather “intersubjectivity”. In answer to a virtual objector who advocates the primacy of the scientific worldview, Strawson replies:

We should note, first, what has happened to the word “subjective”: it has lost its contrast with “intersubjective”, which standard ascriptions of sensible qualities typically are. Obsessed with a particular, scientific aspect of our concern with the objective world, the objector has associated the notion of objectivity exclusively with that concern, and so has lost altogether the sense of the part which concepts of the objective play in our lives as acting, social, and intercommunicating beings. (1992, p.67)

The Kantian influence on Strawson’s view is visible in his ultimate rejoinder to the “scientistic” objector:

What he has perhaps more fundamentally lost sight of is the fact that objects must be perceived as bearers of sensible qualities, visual and
tactile, if they are to be perceived as space-occupiers at all. And this is why the ascription to them of sensible qualities, the standard of correctness of such ascription being intersubjective agreement, is something quite securely rooted in our [ordinary] conceptual scheme. (1992, p.67; emphasis added)

Thus, secondary-quality ascriptions are answerable to standards of correctness that are not independent of the (subjective) mental states of individuals, nor of the dispositional constitution of the human species — but they do not need to be. They are rationally acceptable as they are.

1.2. FREEDOM, RESENTMENT, AND NECESSITY

Strawson's anthropology of moral life, as I consider it, is outlined in two essays of his, which I regard as complementary: “Freedom and Resentment” (1968 [1962]) and “Freedom and Necessity” (1992 [1983]).

In both texts, Strawson explores central features of moral life, in answer to the question as to whether the truth of the "determinist thesis"\(^{15}\) could have any

\(^{15}\) Whereas most authors speak of determinism as a "doctrine", Strawson speaks of it as a "thesis", but no theoretical difference is thereby implied. However, it is only for the sake of the discussion that Strawson grants the assumption that the determinist thesis is intelligible, since he presents himself at the outset of ([1962]) as one of those philosophers «[who] say they do not know what the thesis of determinism is» (Strawson 1968 [1962], p.71).
impact on our moral lives. The primary determinism-sensitive issue is the attribution of moral responsibility: in the case where determinism is true, can we still hold others, and ourselves, responsible? For Strawson, a negative answer to that question would indeed undermine our moral lives, as many essential features of moral life (notably moral sentiments, and the judgments that accompany them) depend upon humans being genuinely responsible. A related issue concerns the knowledge we have of the causal circumstances under which we act. Could it be, as Spinoza claimed, that our ordinary sense of a free, responsible self depends upon our being largely ignorant of those circumstances?

Strawson claims that our sense of a free, responsible self is genuine. His argumentation, in both texts, is largely based on an examination, both phenomenological and analytic, of the ordinary circumstances in which we hold others and ourselves responsible, and of those in which ordinary responsibility attribution is suspended. His main insight is that our ordinary practice of holding people responsible, and our experience of moral sentiments, are protectively isolated, naturally and conceptually, from the corrosive potential of determinism and of the growth of scientific knowledge about the causal conditions under which we act.

There are two logically distinct issues here, which Strawson does not clearly distinguish: isolation and justification. Admitting that our ordinary moral experience, and the judgments that accompanies it, are isolated and protected in
the way mentioned above, does not imply that they are justified. Is our ordinary practice of responsibility attribution, together with the sentiments and judgments that accompany it, genuinely justified, for being protectively isolated in those ways? That is an issue which Strawson does not tackle directly. Let us minimally clarify the matter here.

Let us provisionally assume the correctness of Strawson’s claim that our ordinary practice of responsibility attribution is protectively isolated from the corrosive potential of determinism. Then, there are three possibilities, each of which would require some argumentation: (1) the protective isolation vindicates, on purely pragmatic grounds, our practice of responsibility attribution, together with moral sentiments, and the judgments that accompany them; (2) the vindication issue can be dismissed as idle; (3) responsibility attribution, etc., stand in need of being vindicated, which they are not for being protectively isolated from the corrosive potential of determinism; and if it proves impossible to provide for them some vindication independent from the isolation phenomenon, then the protective isolation is shown to be feeble, and can very well crumble.

We will return to that issue in section 1.3. For now, let us see briefly what the two forms of “isolation” are, that might protect our ordinary practice of responsibility attribution from the corrosive potential of determinism.

We have already encountered the conceptual form of isolation: it is, basically, the “distinct languages” argument. As we will see, Strawson suggests that the moral-psychological language that we use to to talk about free,
responsible agency is irreducible to the “scientific language” that we use to talk about causal processes, and their possibly determined character. They are conceptually isolated from one another in the sense that our understanding of the former does not determine our understanding of the latter.

The other form of isolation is a natural one, which Strawson emphasizes in ([1962]). His point, which is reminiscent of Hume, is that we are naturally disposed to hold others and ourselves responsible for what they/we do, in “normal” circumstances; and that this disposition, which is part and parcel of our natural, “participant” attitude, is too deeply rooted in our constitution to be undermined by the truth of a metaphysical thesis like determinism. However, Strawson adds this twist: our participant attitude is not, even in the context of common life, an undisruptable default position, as it interacts with an “objective attitude”, which does involve the suspension of responsibility attribution (more on this shortly).

There is a connection between the natural and the conceptual forms of isolation: the moral-psychological language, in which we articulate our understanding of our free, responsible agency is, to put it roughly, the language of our natural, participant attitude, of which ordinary responsibility attribution is part and parcel. Conversely, the “scientific language” can be associated to the “objective attitude”, although this association is not necessarily exclusive (more on that shortly).
Subsections 1.2.1. to 1.2.4. follow the argumentative thread of “Freedom and Resentment” ([1962]). In that text, Strawson highlights the connection between responsibility and moral sentiments (cf 1.2.2. in particular). That emphasis, however, should not mislead us: the practice of responsibility attribution is a more central issue than moral sentiments as such\(^{16}\).

Subsection 1.2.5. focuses on “Freedom and Necessity” ([1983]), in which Strawson explicitly argues for the compatibility with determinism of our ordinary sense of a free, responsible self.

1.2.1. Optimists, pessimists, and the contrast between “participant” and “objective” attitudes

Concerning the possible consequences on our moral lives of the putative truth of determinism, Strawson considers two positions: the “optimistic”, and the “pessimistic”. He takes them both to be mistaken, and proposes a distinct view.

\(^{16}\) Strawson calls moral sentiments “reactive attitudes and feelings”, in a way which suggests that both expressions (the traditional one and his own) may be used interchangeably, although the latter expression emphasizes that the sentiments in question are patterns of moral and emotional response to the conduct and attitudes of others. I tend to favor the stylistically lighter expression, “moral sentiments.”
The pessimists, as Strawson portrays them, entertain a certain "metaphysical" conception of moral responsibility and of free will, which they take moral sentiments and judgments to presuppose. Therefore, this "metaphysical" form of moral responsibility and free will is at the very foundation of moral life. The pessimists are incompatibilists: for them, the truth of determinism (and its acknowledgment) would entail the generalized and permanent suspension of responsibility attribution, and so the generalized and permanent inhibition of our moral sentiments and of the judgments that accompany them. This would come down to the destruction of our moral lives.

As for Strawson’s optimists, they are consequentialists: for them, moral judgment is a matter of assessing the consequences of people’s conduct. They concede to the pessimists that acknowledging the truth of determinism might inhibit our moral sentiments on a generalized and permanent basis— but they do not take moral sentiments to be essential to moral life. They also seem ready to concede that determinism would entail the generalized and permanent suspension of responsibility attribution as most people understand it — but we would retain a significant, if minimal, sense of free will, and we would remain able to make moral judgments, understood in a consequentialist sense. They are, then, compatibilists.

Strawson rejects consequentialism, and recognizes, with the pessimists, that moral sentiments (and the judgments that accompany them, understood in a non-consequentialist way) are essential to moral life. But he also rejects the "metaphysical" conception of moral responsibility and free will associated to the
pessimists. Crucially, he rejects the premise on which both optimists and pessimists (as he portrays them) agree, i.e. that the truth of determinism would entail the generalized and permanent inhibition of our moral sentiments, as well as the generalized and permanent suspension of responsibility attribution. Strawson, then, is a compatibilist17, and an “optimist”, of a kind distinct from that which he portrays, at the outset of ([1962]).

In sum, Strawson suggests that it is right to be optimistic, but those who are in virtue of consequentialist theories are optimistic for the wrong reasons. The pessimists’s view of the nature of moral life is more adequate, but they are needlessly worried, as responsibility and free will are found to be compatible with determinism.

As mentioned in the introduction of this section (1.2.), Strawson’s main line of argumentation is that our ordinary practice of responsibility attribution is protectively isolated from the corrosive potential of determinism. In ([1962]), he mostly emphasizes the “natural” from of isolation: the main insight is that we are endowed with a disposition to attribute responsibility in “normal” circumstances,

---

17 Strawson’s compatibilism remains largely implicit in [1962]. He explicitly argues for it in [1983].
As for the consequentialist allegiance of the “optimists” he depicts, it is not explicitly stated, but it is quite clearly implied.
and that this disposition is too deeply rooted in our constitution to be disrupted by the truth of determinism.

A complication is introduced by the fact that, as Stawson presents it, our natural, “participant” attitude (in which partakes our disposition to attribute responsibility in normal circumstances) interacts, in common life, with an “objective attitude”. This contrast is not fundamental in regard to Strawson’s moral anthropology as a whole, but it is worth introducing properly.

The nature of the participant attitude has already been clarified: it involves the automatic attribution of moral responsibility in circumstances taken as “normal”, as well as the ordinary experience of moral sentiments, taken at face value.

The objective attitude, by contrast, involves the suspension of moral responsibility, in “special” circumstances. The suspension of moral responsibility can occur in an unthinking way, or it can be deliberate — as in the case of a special set of considerations which involves a “bracketing” of responsibility, rather than its “suspension”18 (more on this in 1.2.2.). Suspension and bracketing alike entail the inhibition of moral sentiments (and of the judgments of praise or blame which accompany them, unless these are construed in consequentialist terms).

Strawson says of those two attitudes that they do not exclude one another, «but they are, profoundly, opposed to each other.» (1968 [1962], p.79) In any case, the gist of Strawson’s thought does not concern the participant/objective contrast as such, so much as the contrast between those normal circumstances

---

18 This terminological nuance is mine, not Strawson’s.
in which we naturally take others and ourselves to be responsible, and those “special considerations” in which responsibility attribution is suspended or bracketed. (This contrast is paralleled by the conceptual form of isolation by which responsibility attribution is protected, viz. the “distinct languages” argument, to which we will return in 1.2.4. and 1.2.5.) Now, one could argue that the suspension of moral responsibility is very much part of our “participant” attitude, and that it does not entail the adoption of a distinct, “objective” attitude. But, as I hope will appear clearly, that would be an incidental point in regard to Strawson’s main thought.

1.2.2. Moral sentiments, the “manifestation requirement”, and its “responsibility condition”

Strawson distinguishes three types of “reactive attitudes and feelings”: the “personal”, the “properly moral”, and the “self-reactive”.

“Personal” reactive attitudes and feelings are «essentially natural human reactions to the good or ill will or indifference of others towards us, as displayed in their attitudes and actions.» (1968 [1962], p.80) These, then, are the reactive attitudes and feelings that one experiences in reaction to the qualities of someone's will towards oneself. Strawson insists on «the very great importance that we attach to the attitudes and intentions towards us of other human beings»
(1968 [1962], p.75) and mentions as prime examples «gratitude, resentment, forgiveness, love, and hurt feelings.» (ibid)

"Properly moral" reactive attitudes and feelings are vicarious and generalized analogues of the personal type: we experience them in reaction to the qualities of someone's will towards other persons. Strawson claims that moral disapprobation (or "indignation") is an analogue of resentment:

Thus, one who experiences the vicarious analogue of resentment is said to be indignant or disapproving, or morally indignant and disapproving. What we have here is, as it were, resentment on behalf of another, where one's own interest and dignity are not involved; and it is this impersonal or vicarious character of the attitude, added to its others, which entitle it to the qualification "moral" (1968 [1962], p.85; emphasis added)

As for "self-reactive" attitudes and feelings, Strawson describes them as a motivational element underlying other-regarding moral duties:

[T]here are self-reactive attitudes associated with demands on oneself for others. And here we have to mention such phenomena as feeling bound or obliged (the "sense of obligation"); feeling compunction; feeling guilty or remorseful or at least responsible; and the more complicated phenomenon of shame. (1968 [1962], p.86)

Strawson mentions that «all these three types of attitudes are humanly connected» (ibid). However, he is mostly concerned with the first two types. Although those moral sentiments come in a great variety of nuances, most of
them can be seen, roughly speaking, as expressions of “praise” (or “approbation”, as Hume would put it), or as expressions of “blame” (“disapprobation”).

The experience of moral sentiments, taken at face value, presupposes three principles, or standards, which (as we will see shortly) Strawson only evokes implicitly. I propose to formulate them as follows.

At basis, moral sentiments and the judgments that accompany them presuppose the following principle of entitlement: “All human beings are, as such, entitled to regard.” I call this the principle of universal entitlement.

Given that principle, all responsible persons are required to minimally manifest the regard to which all humans are entitled. I call this the manifestation requirement\(^\text{19}\). Outstanding cases of fulfillment of that requirement elicit attitudes, feelings and judgments expressing praise. Non-fulfillment of the same requirement elicit attitudes, feelings and judgments expressing blame. That requirement, then, constitutes the standard of correctness by reference to which moral sentiments (and the judgments that accompany them) are justified or not. It is a non-consequentialist standard: what makes expressions of praise or blame justified or unjustified is not the consequences of an agent’s conduct, but rather whether or not she minimally manifests the regard to which all human beings are entitled.

\(^19\) Arguably, such a requirement is also a commitment: all responsible persons are committed to minimally manifest regard...
There is a condition on the manifestation requirement, indicated by the italics in its formulation above: one is required to minimally manifest regard on the condition that one is responsible for one's conduct (and for its immediate effects). I call this the responsibility condition. That condition is the primary determinism-sensitive issue: the fundamental question is whether the truth of determinism entails that we are not genuinely responsible for our character and our conduct. If we are not, then the manifestation requirement does not hold either. If the manifestation requirement does not hold, how could moral sentiments be justified? And how could the judgments that accompany them be justified, otherwise than in consequentialist terms?

Although Strawson himself does not point that out, it appears that the principle of universal entitlement and the responsibility condition, taken together, imply the following two-level notion of personhood.

In a basic, “metaphysical” sense, personhood is universal, unconditional, and entitling: all humans are persons in this minimal sense, and as such are entitled to regard.

Full personhood, however, requires that one be held responsible for one's decisions and performances — hence the importance of correct attributions of responsibility. This is the conditional, and committing, level of personhood.
Moreover, those three principles allow us to specify what it means to say that determinism has a “corrosive potential”: it might undermine the applicability of any of the three principles in question.

Strawson himself evokes the principle of universal entitlement and the manifestation requirement only indirectly, as he briefly discusses the “expectations or demands” that are made on persons:

The personal reactive attitudes rest on, and reflect, an expectation of, and demand for, the manifestation of a certain degree of goodwill or regard on the part of other human beings towards ourselves; or at least on the expectation of, and demand for, an absence of the manifestation of active ill will or indifferent disregard. (What will, in particular cases, count as manifestations of good will or ill will or disregard will vary in accordance with the particular relationship in which we stand to another human being). The generalized or vicarious analogues of the personal reactive attitudes rest on, and reflect, exactly the same expectation or demand in generalized form; (1968 [1962], p.85; emphasis added)

As for the responsibility condition, it is implicitly invoked in Strawson’s discussion of the “special considerations” which are ordinarily taken, in common

---

20 We will see in Chapter Two how Wiggins claims, in a way similar to Strawson’s, that humans are universally endowed with certain “concerns”, “motives”, “purposes”, “needs”, “aspirations”, and “expectations”, some of which are similarly related to the standards by reference to which we morally assess action, emotions and judgments.
life, to make moral sentiments and judgments unjustified. Those considerations pertain to circumstances where the responsibility condition is not fulfilled, or where it is (as I put it) “bracketed”. In such circumstances, the manifestation requirement is suspended, and moral sentiments (and the judgments that accompany them) are inhibited — or ought to be.

Strawson distinguishes three main classes of “special considerations”. The first two classes include considerations which are ordinarily taken to be excusing conditions.

The first class corresponds to the case of offenders who, although responsible agents, are excused in reason of the unintentional character of the injury they caused. The agents concerned are recognized as responsible for their conduct, but not for its harmful — but unintentional — effects.

The second class corresponds to the case of offenders who are not considered as morally responsible agents, either in reason of their immaturity (children), or in reason of significant psychological impairment (due to mental handicaps or disorders, for example).

Does the suspension of moral responsibility (and the inhibition of moral sentiments), in such circumstances, imply that the “objective attitude” has been adopted? It is not clear that it does. But the third class of special considerations does involve, specifically, a deliberate and reflective adoption of the objective
attitude. Strawson points out that the objective attitude may be adopted in order to see the world differently than it appears from the participant attitude:

[t]o adopt the objective attitude to another human being is to see him, perhaps, as an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something [...] to be managed or handled or cured or trained [...] (1968 [1962], p.79)

This is a case, then, where the objective attitude is adopted as an intellectual resource to be used as an aid to policy, or as an intellectual instrument (here we may note the association of the objective attitude to science). Or again, it can be adopted out of intellectual curiosity, or be used «as a refuge, say, from the strains of involvement.» (1968 [1962], p.80)

In such a case, persons are seen from a perspective where moral responsibility appears as an irrelevant feature of human agents. Their responsibility is not denied, but we might say that it is “bracketed”.

Strawson stresses that those three sets of considerations pertain to special circumstances of common life, standing out from the normal background of our natural, participant attitude. They are part of common life, but they have the status of exceptions.
1.2.3. Strawson’s Humean argument: irremovability by nature

In the Conclusion of Book I of the *Treatise of Human Nature*, Hume argues that philosophy drives those who reason correctly to sceptical conclusions, and he remarks that those conclusions, in themselves, have a deeply unsettling character. But he also testifies of the fact that, as soon as he leaves his philosopher’s office, those unsettling conclusions, although theoretically correct, lose their grip on him. Hume’s point is not merely idiosyncratic. Rather, it is a point about human nature: certain features of our natural attitude, and of the ordinary outlook that we entertain in common life, manifest propensities that are too deeply rooted in human nature to be removed by the conclusions of philosophical reasonings, even if those conclusions are correct.

Although Strawson does not mention Hume, he makes a similar point: our disposition to attribute responsibility in normal circumstances, and the experience of moral sentiments, taken at face value (together with a certain acknowledgement, however dim, of the “manifestation requirement” that this presupposes), are too deep features of our nature to be suspended, on a generalized and permanent basis, by the truth of determinism — or by any other metaphysical consideration.

This is the argument to which Strawson accords the most importance in ([1962]). However, Strawson does not clarify the ambiguity that I mentioned at the outset of the present section, regarding the relationship between the
“protective isolation” claim and the justification issue. He does not seem to endorse (3), the idea that our ordinary practice of responsibility attribution, and the moral sentiments and judgments that come with it, stand in need of a justification distinct from “isolation” arguments. But it is not clear whether he endorses (1), the claim that moral responsibility (etc.) is vindicated on purely pragmatic grounds; or whether he endorses (2), the claim that the justification issue can be dismissed as idle.

Strawson somewhat gestures in the direction of (1), however, with what appears like a puzzling variation on the pragmatist thesis of “the primacy of the practical”. It goes as follows.

Suppose that determinism is true, that it implies that the responsibility condition can never be fulfilled by anyone, and that this entails the generalized and permanent inhibition of moral sentiments, and the suspension of the judgments that accompany them. Then, given that moral sentiments are essential to our moral lives, according to Strawson, the practically rational thing to do would be to blind ourselves to that theoretical conclusion. Let us see how he puts the matter.

In Strawson’s own words, let us admit for the sake of the argument that the issue can be formulated as «a question about what it would be rational to do if determinism were true, a question about the rational justification of ordinary inter-personal attitudes in general.» (1968 [1962], p.83) Then, why think that the rational thing to do would be dictated by theoretical considerations on determinism?
[...] if we could imagine what we cannot have, viz. a choice in this matter, then we could choose rationally only in the light of an assessment of the gains and losses to human life, its enrichment or impoverishment; and the truth or falsity of a general thesis of determinism would not bear on the rationality of this choice. (1968 [1962], p.84)

The suggestion, then, is that if we could have a choice in that matter, it would be better, all things considered, to blind ourselves to certain truths. What is puzzling here — independently of the fact, set aside by Strawson’s thought experiment, that the endorsement of judgments as beliefs is a process which, in many cases, is automatic and completely involuntary — is the suggestion that in certain connections, ignorance and false beliefs may serve human flourishing better than truth and knowledge.

1.2.4. The argument from the distinctness of languages (ca 1962)

In ([1962]), Strawson evokes the “distinct languages” argument as he suggests that the “special considerations” motivating the suspension of responsibility attribution (concerned as they are with the intentions of agents, their capacities, appropriate methodological approaches for certain purposes, etc.) express concerns that have nothing to do with determinism. The suggestion is that we deal with concerns about the conduct of free, responsible agents with
an intentional, moral-psychological language; and we deal with concerns regarding the determined character of causal processes with the “scientific language”. He then proposes a somewhat ambiguous qualification, as he recognizes that the growth of knowledge in various domains (which he describes as a growth of «human self-consciousness») entails the expansion of the circumstances motivating the suspension of moral responsibility. He distinguishes three such domains of knowledge growth.

First, knowledge growth in history and anthropology has increased our awareness of cultural diversity, and of «the great variety of forms which these human [reactive] attitudes may take at different times and in different cultures» (1968 [1962], p.95). The point here does not concern the corrosive potential of determinism, but rather the corrosive potential (in the sense specified in 1.2.2.) of the idea that moral sentiments and moral judgments are products of social conventions — an idea that might be suggested by growing anthropological and historical self-awareness.

Second, there is knowledge growth in scientific psychology, about which Strawson merely says that «psychological studies have made us rightly mistrustful of many particular manifestations of the attitudes I have spoken of» (1968 [1962], p.96). However, he returns to this issue in ([1983]), as we will see in the next subsection (1.2.5.).
Last but not least, there is knowledge growth, both in depth and in scope, in the natural sciences. Strawson recognizes that as a consequence, the space occupied in common life by the natural sciences, and by the objective attitude associated to them, tends to expand — and so the contexts in which responsibility attribution is suspended, and moral sentiments inhibited.

As the antireductionist argument goes, the language and standards of science cannot grow to the point where they become universal; conversely, the use of the moral-psychological language, and the contexts where responsibility is spontaneously attributed, cannot shrink to extinction. Now, this is supposed to be for principled reasons, viz. the conceptual and logical limits, and limited scope, of the scientific language. The ambiguity in Strawson’s evokation of the argument is that it is not clear whether he means to make a principled, conceptual point, or a weaker pragmatic point (or even whether he wants to rule something out by fiat):

That prestige [of scientific studies] is great, and is apt to make us forget that in philosophy, though it is also a theoretical study, we have to take account of the facts in all their bearings; we are not to suppose that we are required, or permitted, as philosophers, to regard ourselves, as human beings, as detached from the attitudes which, as scientists, we study with detachment. This is in no way to deny the possibility and desirability of redirection and modification of our human attitudes in the light of these studies. But we may reasonably think it unlikely that our progressively greater understanding of certain aspects of ourselves will lead to the total disappearance of those aspects. Perhaps it is not inconceivable that it should; and perhaps, then, the dreams of some philosophers will be realized. (1968 [1962], p.96)
At the same time, Strawson remains true to his commitment to a rationally acceptable, critical ontology, which excludes “myth-making”. Thus his anti-scientistic plea is immediately followed by an appeal to ontological rigor, against the metaphysics of libertarianism. Having acknowledged the validity of the pessimists’ view of moral life, and having dismissed their anxieties, Strawson commands the optimists’s conclusion and guarded compatibilist ontology: «When we do remember this, and modify the optimist’s position accordingly, we simultaneously correct its conceptual deficiencies and ward off the dangers it seems to entail, without recourse to the obscure and panicky metaphysics of libertarianism.» (ibid; emphasis added)

Towards the end of ([1962]), Strawson makes several programmatic assertions to the effect that the investigation of moral life necessarily involves an investigation of the natural attitude and of human nature — or, as he put it, «the general framework of human life» (1968 [1962], p.84). Well beyond the analysis of ordinary language (still flourishing at Oxford in 1962), we ought to investigate the «general structure or web of human attitudes and feelings» (1968 [1962]), p.94), which constitutes such an essential part of moral life. Strawson calls for a rehabilitation of the notion of moral sentiments: «It is a pity that talk of the moral sentiments has fallen out of favour. The phrase would be quite a good name for that network of human attitudes [..]» (1968, [1962], p.95).
1.2.5. “Freedom and Necessity”: a compatibilist analysis of the free, responsible self

In “Freedom and Necessity” (1992 [1983]), Strawson proposes to invalidate two famous theses of Spinoza, which he formulates as follows:

According to the first of these theses, the sense of freedom of decision and action, which we experience daily, is nothing but illusion, since it implies a belief which is incompatible with the universal reign of natural causality. According to the second thesis, this illusory sense of freedom is itself caused by a combination of two factors: on the one hand, our consciousness of our actions, decisions, and desires and, on the other hand, our ignorance of their causes. (1992 [1983], p.133)

The first thesis can be detailed as follows. Free will and responsibility are incompatible with determinism; determinism obtains; therefore, our ordinary sense of a free, responsible self is illusory.

Against that thesis, Strawson proposes a compatibilist analysis of human agency, which vindicates our ordinary sense of free, responsible agency (as we will see in 1.2.5.1.).

The second thesis implies that if we knew everything about the causal circumstances of our actions, we would realize that our sense of a free, responsible self is illusory.
Strawson’s suggestion is that this presupposes that “the scientific language” has a universal scope. Against this, Strawson asserts, more forcefully than in ([1962]), the “distinct languages and subject matters” argument (1.2.5.2.)

1.2.5.1. Strawson’s compatibilist analysis of the free, responsible self

Strawson’s compatibilist analysis is based on two points (which he makes very curtly)\(^\text{21}\).

The first point is a gloss on Harry Frankfurt’s argument, according to which the experience of second-order desires partakes in the constitution of a genuine self. Following Frankfurt, Strawson claims that our capacity to entertain second-order desires (i.e. desires that bear on one’s own first-order desires) heighten our sense of self. We notably experience in deliberation and decision-making that those desires which have superior status are more closely involved in the definition our self-identity:

That experience [i.e. deliberation — P.C.] heightens our sense of self; in the higher-order desire which determines what we call our choice we identify ourselves more completely; and this is why we call it our choice. (1992 [1983], p.135)

\(^{21}\) As we will see in Chapter Four, the same two points are integrated (among others) in Dennett’s more elaborate compatibilist analysis of the free, responsible self.
Following Frankfurt, Strawson claims that the capacity to entertain second-order desires is a condition of personhood\textsuperscript{22}. The idea, then, is that our sense of self depends on higher-order desires, but that it is independent of, and compatible with, determinism.

The second point consists in a non-formal analysis of the ordinary concept of “could have done otherwise”.

According to its libertarian interpretation, the ordinary idea that one “could have done otherwise” denotes a metaphysically special capacity that humans are somehow endowed with. Against that, Strawson argues that it simply expresses the non-applicability of excusing conditions in regard to condemnable actions, or to one’s failure to live up to certain moral requirements.

Thus, when we say of a certain agent that “he could have done otherwise”, it is hardly plausible that what we mean is equivalent to «"There was no sufficient natural impediment or bar, of any kind whatsoever, however complex, to his acting otherwise”» (1992 [1983], p.136). According to Strawson, what we mean «amounts rather to the denial of any sufficient natural impediment of certain

\textsuperscript{22} Dennett summarizes Frankfurt’s idea as follows:

A person can want one thing but want to want something else — and act on that second-order desire. Such a capacity to reflect on, and endorse or reject, the desires one discovers in oneself is not just a symptom of maturity, Frankfurt claims; it is a criterion of personhood. (Dennett 2004, p.285)
Such impediments, which constitute genuine excusing conditions for failing to satisfy certain moral requirements, would include for example lack of adequate muscular power, or of financial means (ibid). By contrast, character does not count as an impediment of the relevant kind: as Strawson remarks, pleading something along the lines of “It simply wasn’t in his nature to do so” will not inhibit the moral sentiments and suspend the moral judgments of most people.

It appears, then, that the ordinary usage of “could have done otherwise” is indifferent to, and so compatible with, determinism. The correctness conditions to which responsibility attributions are answerable could not be disrupted by the truth of determinism.

True to his critical approach to ontology, Strawson suggests that libertarianism can hardly avoid an unsatisfactory metaphysics, and an unsatisfactory use of concepts, producing «high-flown nonsense»:

> [W]hen those who accept the equivalence [of “could have done otherwise” with a libertarian conception of free will] are invited to enlarge on the question, how a belief in the absence of determining causes explains or justifies the practices and attitudes in question, their answers are singularly insufficient. It is hard to see how randomness, or a belief in randomness, could either explain or justify any such thing; and attempts to formulate the belief in other terms have never resulted in anything but either high-flown nonsense or
psychological descriptions which are in no way incompatible with the thesis of determinism. (1992 [1987], p.137)\(^2\)

In the end, the truth of determinism would be irrelevant to our ordinary sense of a free, responsible self, as well as to our moral sentiments. There is no good reason to suppose that our ordinary sense of a free, responsible self is illusory, or that it implies beliefs incompatible with determinism. Strawson adds, echoing his Humean “irremovability by nature” argument, that «Our proneness to moral attitudes and feelings is a natural fact, just as the sense of freedom is a natural fact.» (1992 [1987], p.137)

That takes care of Spinoza’s first thesis.

1.2.5.2. The argument from the distinctness of languages (ca 1983)

Whereas Strawson merely evokes the “distinct languages” argument in ([1962]), he makes a more forceful, straightforward use of it in ([1983]).

Strawson, then, makes a distinction between a moral-psychological language and “scientific language”. He then suggests that they fulfill different purposes, as the former deals with the rational, intelligible and morally

\(^2\) Strawson seems to presuppose that an event is either deterministically caused, or random. As we will see in 2.2.3., Wiggins, in his own discussion of libertarianism (Wiggins 2002 [1973]), disputes the validity of this dilemma, and proposes that human agency may be located in the “unclaimed” metaphysical ground between causal determination and randomness.
assessable conduct of the free, morally responsible agents that we are, while the latter deals with causal connections between object-involving event-types.

Strawson then points out that as elaborate, precise, and deep as rational-purposive accounts may get, there is nothing about their development that could threaten our sense of a free, responsible self. He also suggests that causality plays no significant role in the descriptions and explanations of ordinary intentional psychology.

He then claims that the scope of scientific theories and of scientific disciplines is restricted by the conceptual and logical limits of the scientific language.

For one thing, human conduct responds to rational principles, which are beyond the grasp of the scientific language. Consider moment-by-moment, putatively exhaustive causal descriptions of the behavior of human individuals, subsumed under causal laws. As Strawson puts it, it would be literally absurd to mistake the latter for descriptions and explanations of agency in the intelligible and moral sense, because science has no access to the principles that could make the behavior thus described intelligible:

The idea [of giving moment-to-moment scientific accounts of every piece of human behavior — P.C.] is absurd; and not because there would not be world enough and time to work out the solutions to such particular problems, as there is not world enough and time to work out the particular causal conditions of every moment of a leaf on the surface of a stream. It is more fundamentally absurd because there is no practical possibility of establishing the general principles on which
any such calculation would have to be based. (1992 [1983], p.140; emphasis added)

Second, Strawson suggests that the two languages are isolated from one another by the impossibility of psychophysical laws. It is worth pointing out that Strawson means to fully acknowledge that significant correlations between certain physical conditions on the one hand, and certain parameters of human behavior on the other, have been established. He also recognizes that some features of human behavior have come to be subsumed under causal laws:

If the fine connections [...] between the language of the exact sciences on the one hand and the language of mind and behaviour on the other are unattainable, grosser connections are attainable. Many general kinds of dependence of the mental and behavioral on the physical are well enough known. (1992 [1983], p.141)

He also recognizes that knowledge of such dependences is already put to different sorts of practical uses (not all therapeutic), as for example certain chemicals can produce predictable modifications in affect, conduct, performance, etc. Moreover, he recognizes that such knowledge and its uses, are likely to undergo considerable expansion in the future.

All that, Strawson admits, can have an inhibiting effect on our moral sentiments. However, that inhibiting effect has limits, for the aforementioned

24 The thought seems close to Davidson’s anomalous monism.
principled reasons. The point about the absence of psychophysical laws seems to be that, if there cannot be psycho-physical laws, then the inhibiting effect generated by the scientific knowledge of “many general kinds of dependence of the mental and behavioral on the physical” cannot be generalized:

[K]nowledge of such dependences, and the availability of techniques for exploiting them, may surely, in certain cases, contribute to inhibiting those personal and moral attitudes and reactions whose basis is at issue, or at least to lessening their force. So why, it may be asked, should this inhibiting effect not be generalized? All the general traits which manifest themselves in particular episodes of human behaviour, however nuancé may be the descriptions we are inclined to give of those episodes, must, we suppose, have a physical base. So why should the inhibiting effect of such knowledge be confined to certain cases? I think the answer […] is to be found in first noting the fact that these are also the cases which we are favourably disposed to regarding as “cases for treatment”. They are the cases in which the traits in question are displayed in a form which, of itself, tends to inhibit ordinary interpersonal attitudes in favour of “objective” attitudes. Another part of the answer consists in the obvious point that the general, and surely justified, belief in the physical basis of mind is something very different from, and falls far short of, such detailed and particular knowledge of psycho-physical correspondences as I earlier declared to be out of the question. It is for these two reasons, then, that out knowledge of general kinds of dependence of the mental on the physical can be said only to contribute, and only in certain cases, to modifying or inhibiting personal and moral reactions and attitudes. (ibid; emphasis added)
In this way, Strawson claims to invalidate Spinoza’s thesis that knowledge of the causes of our action would dissipate, as an illusion, our ordinary sense of a free, responsible self: the latter pertains to a subject matter that is distinct from that to which pertains knowledge of causes.

1.3. STRAWSON IN PERSPECTIVE

I will begin by identifying three strong points of Strawson’s framework and moral anthropology. Then I will identify what I take to be its weaknesses.

First, the thesis according to which our thought is underdetermined by a fundamental cognitive structure is supported by solid empirical evidence, as we will see in 3.4.1.. This helps understand how certain dispositions partaking in human nature, although they do not determine the content of our thought, nevertheless set conditions and constraints on its articulation.

Second, Strawson proposes what I consider to be an adequately balanced view of rational acceptability, to the extent that he rightly criticizes both those conceptions that are overly restrictive (namely physicalist reductionism) and those that those that are overly liberal. However, the antireductionism he defends might not be nuanced enough (more on that shortly).
Third, as regards Strawson’s anthropology of moral life specifically, I contend that Strawson proposes an adequate, if not sufficiently explicit, analysis of the principles and standards presupposed by the experience of moral sentiments, viz. the principle of universal entitlement, the manifestation requirement, and the responsibility condition. Those principles and standards, moreover, contribute in elucidating the notion of personhood (we will see in Chapter Four that Dennett adopts a similar two-level notion of personhood as that suggested by Strawson).

As for the weaknesses, I find deficient Strawson’s view that our ordinary practice of responsibility attribution, together with our ordinary experience of moral sentiments, are protectively isolated from the corrosive potential of determinism. Let us critically review the two aspects of that isolation.

As regards the conceptual form of isolation, Strawson seems to endorse what I have described as the “distinct languages” argument. I contend that the antireductionism supported by that sort of argument is too coarse.

It might be correct to say that the language of the physical sciences, and the moral-psychological language with which we describe and assess the conduct of the rational, moral beings that we are, express distinct concerns and fulfill distinct functions (and in that sense may be said to have distinct “subject matters”). However, we still have to explain how they are related, and how the conduct of rational, moral beings fits in the world described by the physical
sciences. As it stands, and as Strawson apparently endorses it, that sort of argument does not provide a satisfactory answer to the placement question.

The problem is that this sort of antireductionism assimilates “scientific” to “physicalist”, and fails to acknowledge:
a) The significance of the fact that biology requires an irreducible functional level of discourse.
b) The fact that there is a place, within a regimented, rationally acceptable language, for levels of discourse over and above the functional. Thus, Dennett argues that an irreducible *intentional* level of discourse is indispensable to ethology, as well as to a rationally acceptable, and scientifically credible, form of psychology. At that point, we can see that an irreducible moral level of discourse is just as indispensable to make rationally acceptable sense of morality.

In other words, we must reject the idea that the moral-psychological language on the one hand, and the “scientific language” on the other, constitute a dichotomy. If we adopt Dennett’s multi-level view of rational acceptability, it is possible to see how ordinary talk about human affairs is related to the physicalist language. As it stands, Strawson’s advocacy of a rationally acceptable language broader than the “scientific language” of the physical sciences leaves us in the dark as to how the latter is related to ordinary talk.

As regards the natural form of isolation — the Humean “irremovability by nature” argument — we may dispute Strawson’s contention that our natural
disposition to attribute moral responsibility (in normal circumstances) is too deeply rooted in our constitution to be sensitive to the putative truth of the determinist thesis. Moreover, even if we accept this Humean, “irremovability by nature” claim, it remains that Strawson does not satisfactorily tackle the justification issue. Even if it was the case that responsibility attribution is isolated from the corrosive potential of determinism, it would not be vindicated de facto. Strawson’s suggestion that, in the case were determinism would obtain, and where we would have control over our beliefs and our disposition to attribute responsibility, the rational thing to do would be to blind ourselves to the truth of determinism, is insufficient.

Overall, Strawson provides for moral anthropology a strong thesis (moral principles and standards are underdetermined by human nature) and an adequate analysis (moral experience presupposes the principle of universal entitlement, the manifestation requirement, and the fulfillment of the responsibility condition). That is, our human predicament conditions and constrains the standards by reference to which we assess our conduct and our emotional responses; in this way, our human predicament makes it natural for us human beings to endorse those standards and principles in the practice of giving and asking for reasons. But our human predicament does not determine those standards and principles; it does not de facto justifies them; and it does not make idle the issue of their justification.
Now, the theoretical basis that Strawson provides us with stands in need of being developed. I contend that Wiggins and Dennett show us how we can elaborate it along the following lines (which correspond to my idea of what we can extract from the work of those two philosophers).

We are endowed with species-wide cognitive, behavioral and emotional dispositions, which determine “natural attractors” in the space of reasons. Those natural attractors, in turn, *condition and constrain* ordinary standards, by reference to which we assess human conduct, emotions and judgments. In common life, the pull of natural attractors favors consensus upon the propriety, or lack thereof, of certain sentiments, judgments and actions, in given circumstances. In the case where sentiments, judgments or actions that correspond to the manifestation of natural dispositions are being scrutinized, most people will agree about the reasons that justify them. Our natural dispositions, together with the pursuits and practical necessities of common life, set “natural limits” on rational criticism and drive our “why questions” towards certain answers, that seem universally shared across humankind.

However, natural dispositions and natural attractors do not *determine* standards, because humans do have the capacity to call them into question. Rational criticism opens up “the logical space of reasons”, within which everything can be scrutinized, and called into doubt. In principle, rational criticism is radically open-ended: any reason offered in support of any judgment, decision or action can only be another judgment, for which a ground may be demanded in turn. Thus, on the one hand, rational criticism confers autonomy upon the setting
of the standards of correctness; but on the other hand human natural dispositions act as attractors, so that rational criticism tends to confirm the natural dispositions. In sum, natural dispositions and natural attractors in the space of reasons, underdetermine ordinary standards.

As regards specifically the issue of moral meaningfulness, we might say that Strawson argues in favor of a quietist position in regard to the elementary form of nihilism, according to which the irremovability of our standards makes of nihilism a negligible position (as we will see in Chapter Four, Dennett comes to the same view about nihilism, but in connection with moral decision-making, not moral sentiments.)

In sum, we need to push further the critical scrutiny of common moral life. As we will see in the next chapter, that is what Wiggins does.
CHAPTER 2. DAVID WIGGINS: HUMAN NATURE AND MORAL LIFE

This chapter is divided into four parts. Section 2.1. bears on the Wigginsian framework; section 2.2. presents Wiggins’ moral anthropology itself; section 2.3. presents Wiggins’ metaethical view, and the metaphysics of experience which underlies it; it also presents his criticism of rival views. I present a brief assessment of Wiggins’ view in 2.4.

The introduction of this chapter presents Wiggins’ view of the field of moral philosophy in general. That will allow us to locate Wiggins within a certain theoretical and doctrinal landscape, as well as to clarify a few terms.

At the outset of his Twelve Lectures on the Philosophy of Morality (2006) Wiggins formulates as follows the four main clusters of issues with which, according to him, the philosophy of morality is concerned:

(A) It concerns the question of the substance or content of morality, its nature, and its extent.

(B) It concerns the question of the reasons there may be — and the reasons agents may make their own — to participate, persevere, and persist in morality.
(C) Consequentially upon the kind of content and authority that is attributed to it under heads (A) and (B), the philosophy of morality seeks to determine the logical and metaphysical character of the findings of morality. It studies the question of the truth, objectivity, relativity, etc. of its judgments and the logical status of the approbation that it extends to some acts and responses but denies to a host of others.

Among further questions there is a question (D) about the relations of morality, meaning and happiness. (Wiggins 2006, p.9-10)²⁵

In his currently published work, Wiggins has focused on questions (A), (B), and (C). This prepares the ground for an extensive, yet unpublished, discussion of question (D). Basing myself on (2002 [1976]), I assume that Wiggins sees this last topic as bearing principally on “the meaning of life”, i.e. an ordinary, existential, teleological construal of the good as moral meaningfulness: what has intrinsic worth, such that it may constitute a non-derived finality to human existence, etc. Clearly, it is on the basis of his moral anthropology that question (D) is to be discussed.

Let us now clarify the position of Wiggins’ approach to the philosophy of morality, relatively to other, more traditional approaches. What follows is taken to be common lore for moral philosophers, but those points that Wiggins explicitly

---
²⁵ For the rest of this chapter, the name of Wiggins will be omitted from references to his work.
mentions are referenced, as are those junctures where he takes a distinctive stand.

The content of morality (issue (A) above) involves values, and as such corresponds to the valuational dimension of morality, paradigmatically expressed by “is” judgments.

The philosophical investigation of the valuational takes the form of what we can generically call the “theory of value”, or “theory of valuation”. To the extent that the principal values distinguish good from evil, the theory of value can be identified with a “theory of the good”.

The motive of morality (issue (B) above), involves the threefold distinction between prohibited, obligatory, and permissible action. It usually implies that morality is binding, i.e. it makes a motivational claim upon our decisions, actions, policies, general conduct… As such, it corresponds to the deontological dimension of morality, paradigmatically expressed by “ought” judgments.

The philosophical investigation of the deontological domain corresponds to what we can generically call the “theory of right decision/conduct/action”, coupled to the account of moral motivation (how is it that morality makes a motivational claim upon us, if it does?)

For Wiggins, “is” judgments, expressing the valuational dimension of morality, are moral evaluative judgments, included in the broader class of
“evaluative judgments” in general. And “ought” judgments, expressing the
deontological dimension of morality, are moral practical judgments, included in
the broader class of practical judgments in general. Wiggins points out that these
two classes partly overlap. Thus, the distinction between the valuational domain
and the deontological domain is not a dichotomy, but a distinction involving an
intersection between the two distinguished terms.

The broad class of evaluative judgments includes aesthetic judgments.
However, Wiggins does not directly discuss them for the reason that, according
to him, they «have special features that would in the end require special

The class of evaluative judgments also includes affective judgments ("t is
disgusting", “t is funny”, etc.) In Wiggins’ philosophy, the considerations that apply
to moral judgments and properties often apply to affective judgments and
properties, and vice versa (in this, his doctrine manifests a Humean influence).
Thus, in many regards, Wiggins’ philosophy of morality partakes in a broadly
construed “philosophy of value”, even though he rarely discusses the case of
aesthetics (we will return to the distinction between evaluative and practical
judgments in 2.1.3.5.)26.

26 Wiggins uses the term “valuational” as a generic term designating conjointly the moral, the
affective and the aesthetic (which is theoretically taken on board where the special treatment it
otherwise calls for is not required). I will follow that usage.
The field that has come to be known as "substantive moral theory" comprises three main, traditional theoretical families: consequentialist theories, deontological theories, and virtue theories. These are essentially normative, and not descriptive, philosophical theories, according to which genuine moral standards are founded on certain principles (although, in the case of virtue theory, it is not presumed that the foundational principle is independent from our human constitution). Analytic versions of those three kinds of moral theory focus on the analysis of moral discourse, in a way which suggests that morality comes down to moral discourse (and therefore that the investigation of morality comes down to the analysis of moral discourse).

According to consequentialist theories (2006, p.175, 206), the valuational determines the deontological. More precisely, consequentialism proposes a theory of value which identifies the good (what has intrinsic worth) with "states of satisfaction", and upon this theory, right action/decision/conduct is defined in terms of the maximization of the good. Consequentialist theories imply that the aim or point of morality «comes down to the promotion of a single-minded generalized pursuit of happiness or desire satisfaction» (2006, p.264); or, in the case of theories of the "ideal utilitarian" type, that it «comes down to the (however double-minded) pursuit of intrinsic value» (ibid).

According to deontological theories, the deontological determines the valuational. More precisely, deontological theories propose a theory of right action (against the background of the prohibited/obligatory/permissible
distinction) and upon this theory, it determines the good. Most deontological theories — especially of the Kantian sort — imply that the aim or point of morality «comes down to the sovereign will's discovery of maxims it can will universally[.]» (ibid)

As for virtue theory, it states that «the organizing focus of morality is virtue itself.» (ibid) Wiggins continues:

If Kantianism were right, then agents who deliberated well would have constant regard to the sovereign will; if utilitarianism were right, then agents who deliberated well would have regard to happiness or desire satisfaction; if ideal utilitarianism were right, then they would have regard to intrinsic value, etc. These implications are not counterintuitive or unintended. What then if virtue ethics were right? Agents who deliberated well would have regard to virtue. That sounds wrong, however, for it suggests a general preoccupation on the part of agents with virtue as such, a preoccupation that few “virtue ethicists” can have intended. (2006, p.264, n.34)

Wiggins suggests that each of the three classical approaches to substantive moral theory is unsatisfactory, for the reason that it unduly gives priority to one thread of the moral dimension of human existence, at the expense of the others. Not only does this result in lop-sided theories of morality, it also fails to capture, altogether, the significance of such issues as moral meaningfulness (question D).

Against those traditional “moral theories”, Wiggins proposes his anthropology of moral life, from which it differs significantly.
First, the moral anthropology is descriptive before it is normative. Humans share an “ordinary morality” (or “first-order ethic”), and in order to understand it and to critically assess it, we have to investigate those human dispositions which (in my vocabulary) underdetermine them\textsuperscript{27}. Methodologically and doctrinally speaking, the description of this phenomenon is a theoretical, “speculative” task, which leads to a “practical” conclusion as it entails (given an anthropocentric conception of the nature of adequate standards of correctness) the vindication of ordinary morality — rather than a radical revision of it:

Even though our interest in the practical is speculative in this way, let us leave the expression “moral theory” in the hands of those who relish its present connotation of reconstruction or believe that it is the business of moral philosophy to trespass on the preserve of the moralist by remaking first-order morality. Until we attain some speculative understanding, it is too soon for us to commit ourselves to any of that. At the outset, let us bracket all preconceptions and allow the notions we already have to speak to us for themselves. (2006, p.7)

Second, against the specifically analytic version of the three aforementioned theories, Wiggins denies that morality comes down to moral discourse, so that the investigation of the former comes down to the analysis of the latter. In the wake of Aristotle, Wiggins claims that morality is an essentially practical domain of human thought and activity. Although he tends to speak of the first-order ethic as if it corresponded to the linguistic articulation of morality, it

\textsuperscript{27} That is what Strawson calls for in ([1962]).
must be understood that morality, as an essentially practical endeavor, does not come down to moral discourse, and cannot be assimilated to a particular linguistic articulation of first-order morality.

Third, Wiggins claims that the deontological and the valuational dimensions of morality conjointly “arise out of” certain dispositions which are part and parcel of human nature, and mutually determine one another.

2.1. THE WIGGINSIAN FRAMEWORK

This section presents the three doctrinal components which constitute the basis of Wiggins’ moral anthropology, together with the Davidsonian theory on which his framework is centered.

Wiggins speaks of a “thesis of underdetermination” at a few junctures, in a way that remains unclear, for example: «But really there is no such thing as the one good, no such thing as knowledge of it, and nothing fixed independently of ourselves to aim at. Or that is what is implied by the thesis of cognitive underdetermination.» (2002 [1976], p.126) But it will appear in 2.1.1., and
throughout this chapter, that he is pointing to the thesis of underdetermination that I described in the Introduction.

As we will see in 2.1.1., he suggests that all human beings are endowed with a set of dispositions, manifested in action, judgments and sentiments, and which underdetermine the standards by reference to which we assess action, judgments and sentiments.

Considering the nature of morality as a domain of human experience, thought and action, Wiggins adds the following insight: morality is an essentially practical domain, a practical mode of being. This implies that it is a part of common life — part of the “general framework of humanity”, as Strawson put it — and that it does not come down to the moral discourse which conceptually articulates it, even though, of course, moral discourse is a very important part of morality: moral responses (moral sentiments) are usually accompanied by moral judgments.

Wiggins defends an antireductionist view, to which he ties the anthropocentricity thesis (as we will also see in 2.1.1.). Even though he nowhere proposes a canonical formulation of the latter, he clearly states that the standards to which most domains of human thought and action are answerable are anthropocentric, and adequate — including moral standards. Thus, Wiggins’ philosophy of value purports to vindicate the basis of our “ordinary morality”.

Wiggins defends a metaphysics of experience according to which no logical or metaphysical gap intervenes in the Subject/Object relation (2.1.2.). He
does not discuss the experience of the primary qualities of material objects. Rather, he focuses on non-primary properties: secondary qualities, and valuational properties. He argues that in regard to such properties, the perceiving subject, with her dispositional constitution, and the non-primary properties that she perceives, are equal and reciprocal partners, so the experience of such properties is at once world-involving and subject-involving. (2002 [1976], p.106)

Wiggins' point is that responses and judgments that bear on non-primary properties are answerable to species-wide standards; and that non-primary properties depend for their very existence on the dispositional constitution of the human species. Secondary qualities are a clear case of that: «Red is not a relational property. [...] All the same, it is in one interesting sense a relative property. For the category of colour is an anthropocentric category. The category corresponds to an interest that can only take root in creatures with something approaching our own sensory apparatus. (2002 [1976], p.107)»

In the case of valuational responses and judgments, however, and especially in the case of morality, Wiggins recognizes that culture exerts an influence such that it creates standards of its own. Not all moral standards are species-wide, then; but a basic range of them are. Even though all humans might not recognize that a basic range of their moral responses and judgments are answerable to species-wide standards, this is something that is there to be discovered by all, in the intersubjective practice of rational criticism.

Wiggins, then, defends a view of valuational experience according to which the valuational responses of the human subject on the one hand, and the
valuational properties experienced on the other hand, are “equal and reciprocal partners”, due to the fact that human valuational responses manifest dispositions which are naturally, normally attuned to certain properties of the external world, while the properties in question are undefinable in independence of the responses they elicit. Property-attuned responses, and response-dependent properties, are equal and reciprocal partners.

The fact that certain human dispositions, and the patterns of response that manifest them, are attuned to certain valuational properties, explains why these patterns of response feature *an irreducible element of directedness and appropriateness* (or “desert”, or “merit”)²⁸.

Now, if the same human dispositions, attuned to valuational properties and manifested in patterns of response featuring an irreducible element of appropriateness, underdetermine valuational standards, then it is plausible that these standards are at once subject-involving (because they are underdetermined by the dispositions of subjects), *and* world-involving (because these dispositions are partly shaped by the properties which are their equal and reciprocal partners).

²⁸ I take it that the idea of “being attuned” implies the idea of appropriate directedness. Since it is implied, I will not systematically carry the reference to the directedness of dispositions, and of the patterns of response that manifest them.
As we will see, the critical side of this second emphasis is directed against reductionist approaches, which treat intentional states «in independence of what they are directed to» (2002 [1976], p.97).

Wiggins’ metaphysics of experience supports his theory of interpretation and of the critical assessment of discourse — not just moral discourse, but discourse in general (2.1.3.).

Ordinary agents interpret ordinary moral discourse as part of their interpretation of general discourse; and the subject matter of moral discourse, as that of any area of discourse, is a potential object of rational discussion. Wiggins accounts for the interpretation and critical elucidation of discourse with a Davidsonian theory. The first task that he carries off within this framework is, in Wiggins’ view, of crucial importance for the analysis of any area of discourse: the elucidation of our ordinary concept of truth, into five marks of truth, according to which our ordinary notion of truth is both subject-involving and object-involving or ("world-involving"). According to Wiggins’ metaethical analysis, valuational judgments can be true in a sense that is subject-involving and world-involving.

It appears, then, that Wiggins’ Davidsonian theory of interpretation interlocks with the metaphysics of valuational experience. According to the latter, valuational responses experienced by human subjects and valuational properties are equal and reciprocal partners (so valuational experience is at once world-involving and subject-involving). According to the former, the judgments that
accompany those responses can be true in a sense that is subject-involving and world-involving.

Wiggins suggests that this combination of world-involvingness and subject-involvingness is reflected in the logic of our valuational vocabulary. If we «allow the notions we already have to speak to us for themselves» (2006, p.7), it appears that our valuational concepts designate response-dependent properties, where the responses in question are not definable independently of their sense of appropriateness and directedness, which is thus included in the correctness conditions for the use of those concepts. Valuational responses and judgments, then, are answerable to standards of correctness that are, not just intersubjective, but world-involving as well as subject-involving.

2.1.1. Underdetermination, anthropocentricity, and the practical nature of morality

Wiggins claims that the standards to which our thought and action are answerable emerge from the practice of giving and asking for reasons, and that our human constitution sets limits on that practice, and exerts constrains on it, so that its deliverances are conditioned and constrained, but not determined, by “human nature”. He does not give more explicit indications as to the nature of the relation of underdetermination between human nature and morality, but we can elaborate a bit upon the two terms involved in the relation.
To put it as clearly as possible, we may say that dispositions underdetermine the standards by reference to which we morally assess (in terms of “right/wrong”, “good/bad”) conduct, judgments, sentiments.

Now, Wiggins does not always clearly distinguish the dispositions which partake in human nature, on the one hand, and the moral standards that they underdetermine on the other. Rather, he makes great use of a vocabulary which does not lend itself to such precisions, but which is very suggestive; that is, his vocabulary of concerns, motives, purposes, needs, aspirations and expectations. These notions give substantive, phenomenological bite to the theory of human nature, and can be construed in dispositional terms (so they do not alter the dispositional character of the theory). As Wiggins puts it, motives and concerns (etc.) «are translated into settled dispositions and habits of mind» (2006, p.13)\(^29\). But Wiggins also uses those terms as if they were designating features, not of human nature, but of morality itself:

inward ends or purposes or concerns which, in the business of their life at a given time and place, participants in a first-order ethic will steadfastly adhere to as if by second nature, distinguishing readily, however essentially contestably, between these concerns and other concerns that can be abandoned and may have to be. (2006, p.329)

\(^{29}\) The substantiveness and the phenomenological bite characterizes the set of all these terms.

Wiggins’ usage does not imply that he makes important distinctions between each term, individually taken. In fact, he mostly uses the first two terms of the series, but in a way that suggests that the others are taken on board. I will follow that usage.
Wiggins claims that some of those concerns and motives are “unforsakeable”. This announces the vindicatory orientation of his philosophical account of morality. Some moral standards underdetermined by human dispositions are vindicated, to the extent that they are unforsakeable. They are not to be rejected on the basis that they are “merely” human.

The underdetermination thesis, then, is asserted with the greatest clarity in terms of dispositions underdetermining standards. But Wiggins speaks more often in terms of concerns, motives, purposes, needs, aspirations and expectations, and does not clearly indicates whether those that he characterizes as unforsakeable are part of human nature or part of morality. The way he uses those notions suggests that they are actually part of both. Wiggins seems to have decided to sacrifice some clarity to the profit of phenomenological suggestiveness.

Wiggins has found in the philosophy of Hume various manifestations of the underdetermination thesis and of the anthropocentricity thesis.

Whether his reading of Hume is exegetically acceptable or not is of no direct concern to the present project. If Wiggins’ reading of Hume is indeed exegetically acceptable, then his philosophy is doctrinally and substantively indebted to Hume. If it is not exegetically acceptable, then Hume was merely a source of inspiration for him (in which case we could say, as Richard Rorty would put it, that Wiggins is engaged in a “strong misreading” of Hume). Moreover, it is my reading of Wiggins’ reading of Hume that I will present here. Although it is
faithful to Wiggins’ thought, I shape his view with the vocabulary and the articulation of four doctrinal components that I presented in my Introduction.

Thus, Wiggins finds in Hume a moral philosophy shaped by the thesis of underdetermination. He also brings out a strand of Hume’s thought that «has no clear place in his official theory» (2002 [1987c], p.194), and which marks a step out of the “classical” form of subjectivism to which Hume’s philosophy as a whole is often assimilated. This strand of Hume’s thought suggests that valuational responses and judgments are answerable to intersubjective standards of correctness, and that they reach beyond mere subjective mental states (as we will see, Wiggins pulls it in the direction of his own metaphysics of experience).

Wiggins also brings out the vindicatory strand of Hume’s thought, associated with the thesis of the adequacy of anthropocentric thesis, to which Hume came to adhere, according to Wiggins. This vindicatory strand of thought (which will be presented very shortly in 2.1.1.1.) involves a major tension with the reductionist strand of Hume’s thought, which runs throughout his intellectual life, although, according to Wiggins, he abandoned the “classical” form of subjectivism after the Treatise. The reductionist strand will be presented in 2.3.2., together with Wiggins’ exegetical account of how it recedes as the vindicatory strand asserts itself, in conjunction with the intersubjectivist strand.

Wiggins also found inspiration in Aristotle (2.1.2.2.) for the thesis of underdetermination, for the thesis of the essentially practical character of morality, and for the anthropocentricity thesis.
2.1.1.1. Anthropocentricity and the vindication of morality in Hume

Hume’s anthropology of moral life is descriptive before it is normative. That is, morality, with its normative dimension, is accounted for on the basis of a description of human nature.

Schematizing Hume’s view of human nature with a vocabulary not exactly his own, we may say that it includes behavioral, cognitive and emotional dispositions, which, interacting with faculties and capacities, are manifested in patterns of response, accompanied by judgments.

Some of those dispositions are moral, and can be classified along the age-old distinction of “virtues” and “vices”. Hume proposes a distinction between two kinds of virtues and vices: the natural, which are inherent to human nature; and the artificial, which developed in interaction with conventions that arose out of the practical necessities of human sociality (given certain fundamental features of human nature). There is a historical development there, which implies that Hume’s phenomenological descriptions of “passions”, of natural and artificial virtues and vices, relies, as Wiggins puts it, on a “genealogy of morals”.

In Wiggins’ reading, the vindicatory strand of Hume’s thought is progressively asserted in the chronological deployment of his work, as Hume progressively acknowledges that there may not be any good reason to reject anthropocentric standards. In the Treatise, Hume has not resolutely adopted the anthropocentric thesis, and he gives contradictory cues as to whether scepticism
is a definitive conclusion or a reductio ad absurdum. But according to Wiggins, Hume leans towards the anthropocentric thesis from the *Enquiry* on.

In the end, Hume “made room”, Wiggins tells us, for the thought that our ordinary moral standards, as they arise out of unforsakeable human dispositions, are by and large justified (note how Wiggins uses his own vocabulary of “needs”, “aspirations” and “expectations” in his presentation of Hume’s dispositional view of human nature, and note how he suggests that the latter constrains morality):

The morality of human beings is certainly a singular creation. It does not follow that human animals and human morality have themselves to lie beyond the reach of “moral science” as Hume came to conceive it. As his project advances, it is made more and more evident that to look upon human beings in the way in which Hume interprets his Newtonian aspiration is not in itself to transform or denature these things. [H]is method […] makes room for the possibility that a well-observed description and explanation of morals may prove to vindicate morals — not least by showing the multiplicity and stringency of the constraints that lie upon any first-order ethic that will answer to the strictly unforsakeable needs, aspirations or expectations of its participants. (2006, p.35; emphasis added)

The idea is that the standards of morality are binding — they make a motivational claim upon all humans — not in spite of their being “merely” grounded in human nature, but in reason of their being grounded in certain dispositional features of human nature, of which Hume came to see that that they were (in Wiggins’ vocabulary) “unforsakeable”.

104
It should be clear that the bindingness of morality does not depend upon an elaborate knowledge, on the part of people, of human nature. Rather, accounting for the bindingness of morality requires a descriptive account of human nature, in which morality finds its own vindication. Wiggins, then, attributes to Hume this vindicatory view that one need not forget about one's theoretical life in order to save one's moral life. On the contrary, the latter can be enlightened by the former:

One who recognizes the force of the Humean account of why he, the subject himself, feels or thinks as he does need not find that, if he accepts the explanation, this subverts his finding of moral beauty in this or that act or character. The acceptance of the explanation will not undermine the reasons he would give himself to follow this or that line of action. For the explanation, the conviction, and the reasons will coexist happily enough in the agent’s consciousness. (2006, p.36)

In the light of Wiggins’ anthropocentric reading, scepticism is not a definitive conclusion for Hume, but rather a reductio ad absurdum: it tells us that if genuine knowledge requires to be grounded in apodictic truths or in necessary principles, then we do not have knowledge — an absurd conclusion that shows that the premise is false.

We will also see in section 2.2. that, according to Wiggins, Hume came to reject "cosmic" conceptions of rationality, in favor of the ordinary concept of «the
reasonable as a sum of good reasons which agents can appropriate as reasons of their own» (2006, p.66), which he vindicated.\textsuperscript{30}

Wiggins characterizes the Humean moral anthropology as “dispositional”, for the reason that Hume views human nature first and foremost as a nexus of dispositions. He also characterizes it as “formational”, for the following reason: although we are naturally disposed to morality, moral education is required for those dispositions to find their right expression. Morality as such is not innate. This is one of the many significant common traits that Wiggins finds between Hume and Aristotle.

2.1.1.2. Aristotle: morality as essentially practical

For Wiggins, much of Aristotle’s ethics is “fully harmonious” with Hume’s, and enriches it with distinct insights. In his own way, Aristotle also proposed a “dispositional and formational” moral anthropology, framed by the theses of underdetermination and of anthropocentricity.

\textsuperscript{30} Wiggins defines “cosmic” rationality as «rationality for all conceivable agents» (2002 [1976], p.121). As an example of an attempt to make sense of cosmically valid rationality, he points to the conception of rationality that underlies Peirce’s idea of truth as the ideal limit of inquiry — a conception of rationality that Peirce identifies with Science.
Thus the Aristotelian view is that the human species is endowed with a set of dispositions, some of which give rise to morality, although the acquisition of a first-order morality by individuals requires moral education.

Aristotle, then, can enrich the Humean anthropology with the insight that morality is an essentially practical domain of human thought and action, expressing the ability to make the right judgments and the right decisions, to adopt the right policies and the right conduct. As the expression of an ability, morality does not come down to moral discourse. It is a *phronesis* (practical wisdom) that propositionally articulated judgments express, but do not exhaust:

On the formational and essentially practical view, which is the view of Aristotle but fully harmonious with that of Hume, what is the relation of a person’s ethical dispositions to his or her moral convictions? A rough answer might be that, however incompletely or inexpressively, a person’s spoken convictions express the state of being that is realized in these dispositions. An agent’s spoken utterances are a part of his practical response to the situation in which he finds himself. Or else they are a provisional response to the as-if practical envisaging of a whole class of such situations. In the typical case [...] *the person who gives voice to a moral conviction thereby manifests a state of practical understanding of a certain context* — an understanding whose title to correctness does not depend on its degree of articulacy or the immediate availability to the moral subject of propositional grounds adduced for it. Its correctness will depend on the soundness of the subject’s perception, the experience and the practical intelligence applied, and the quality of the ideas and expectations brought to bear in the context. (2006, p.237; emphasis added)
In other words, our *moral dispositions* — concerns, motives, purposes, etc. —, in order to manifest themselves in the right way, involve a certain practical understanding of the circumstances we find ourselves in. Acquiring the capacity of practical wisdom requires a moral upbringing, involving the living, active example of persons who are practically wise. Practical wisdom is not primarily acquired through arguments or proof:

Only by drawing on some such understanding of the “that for the sake of which” [the practically wise person acts] will anyone subjugate the indefiniteness of the subject-matter of the practical or impose upon any part of it a shape that is faithful to the ethical aim. It is understanding of the *spirit* in which the agent is to act.

According to this formational (or, as one might say, mental dispositional) view of morals and morality […], the propositional or declarative element in morality rests in part on the practical. But, once you take the practical fully seriously, the question is not foreclosed whether the correctness or incorrectness of moral judgments can amount to substantial truth or falsehood. (2006, p.238)

In regard to the analytic tradition, the Aristotelian-Humean sort of moral anthropology implies the rejection of the idea that morality — a moral outlook — comes down to moral *discourse*, and that consequently the investigation of the former comes down to the analysis of the latter.

In philosophy as it is, there is a tendency for a first-order morality to be conceived as a structured array of propositions or judgments. But it will cohere much better with [our] purposes […] to conceive of such an ethic more dispositionally, as a nexus of distinctive sensibilities, cares
and concerns that are expressed in distinctive patterns of emotional and practical responses. As instantiated in the formation of those who live by a given first-order ethic and see their way of being and acting as answerable to some particular standard of norm that they hold in common, such a nexus will be conceived by a neo-Humean genealogist or aetiologist as something with a prehistory and a history, as well as a present and a possible future. The dispositions that realize the ethic he will see as arising neither uninvited from the given nature of human beings nor in a manner contrary to that nature [...], but as establishing themselves in creatures who are well enough suited to receive them and are rehearsed, practised, or habituated therein by those who already have and exercise such dispositions. (2006, p.236; emphases added)

Wiggins adopts an Aristotelian construal of morality, according to which (as we will see shortly) it is a practical ability and domain of thought and action, which does not come down to moral discourse, even though moral discourse is of course extraordinarily important as an expression of morality. Wiggins often uses the expression first-order ethic in a way which suggests that it refers to something linguistic. But clearly, the linguistic articulation of morality is not identical to morality.

Moving beyond Aristotle and Hume, Wiggins sets the concern for morality within a view according to which a thorough understanding of our ordinary linguistic practices, and of the “forms of life” to which they are connected (so Wittgenstein said), demands that we investigate the natural history of the natural attitude, and of these forms of life:
[A] morality does not consist of a set of moral propositions, even a very large one. Moral judgments are indeed partial expressions of the findings or demands of some particular mode of being and its associated sensibility. But there is simply no question of collecting up a whole mass of moral judgments [...] in order for these judgments to represent or construct or create an ethic, a morality, or a whole manner of life [...] For (1) judgments are only half of the story (the later half) and (2) moral judgments themselves, even when spelled out, cannot even be understood as they are intended except against the background of a lived understanding that will never be fully articulated. In the absence of such a background, you have no hope of being understood exactly as you intend to be understood even if you say something as simple as “It is wrong to say what is not true/ what you know is not true…” In the absence of such a background, which it would be an endless process, that is impossible, to spell out fully explicitly, you could never keep in balance, as is second nature to us, the conflicting claims of a pair of fully compacted proverbs such as “He who hesitates is lost” and “Look before you leap”. (2006, p.350; emphasis added)

The fact that morality does not come down to moral discourse should not be taken, however, as suggesting that the question as to whether moral judgments are good candidates for truth is irrelevant: «once you take the practical fully seriously, the question is not foreclosed whether the correctness or incorrectness of moral judgments can amount to substantial truth or falsehood.» (2006, p.238)
2.1.2. Wiggins’ metaphysics of valuational experience

Morality arises out of certain dispositions that are attuned to certain properties, in a way such that the responses that manifest the dispositions feature an irreducible element of directedness and appropriateness (or “desert”, or “merit”). That is, we have dispositions to respond to certain properties in certain ways — with certain patterns of response — which are appropriate, on the condition that they are directed to those properties that “normally” elicit them, following the “normal” attunement of human dispositions.

As Wiggins points out, this is an idea that Hume came to in his essay “Of the Standard of Taste” (1757). It is, as Wiggins put it, «a Humean suggestion that has no clear place in his official theory: “It must be allowed that there are certain qualities in objects which are fitted by nature to produce particular … feelings.”» (2002 [1987c], p.194; quoting Hume (1757), paragraph 16).

Echoing Hume, Wiggins proposes that valuational properties (affective and moral) and human responses are “made for one another” (2002 [1987c], p.195 n.16). He also uses the metaphor of “equal and reciprocal partners” to express the same idea:

Surely an adequate account of these matters will have to treat psychological states and their objects as equal and reciprocal partners, and is likely to need to see the identifications of the states and of the properties under which the states subsume their objects as interdependent. (2002 [1976], p.106; emphases added)
In the case where valuational responses are accompanied by judgments that exhibit all marks of truth, we may justifiably consider that we genuinely respond to “real” moral or affective properties, although those properties are response-dependent, and therefore not entirely objective. Still, judgments about valuational properties which exhibit all marks of truth are true in the world-involving sense of the ordinary concept.

Wiggins suggests that this view is merely a philosophical elucidation of our commonsensical realism about valuational properties, which is a weak form of realism which can also be characterized as a “sensible”, world-involving subjectivism, i.e. «a subjectivism of subjects and properties mutually adjusted.» (2002 [1987c], p.199)

As a corollary, a particular judgment can be at once subjective and objective:

For the following possibility has now become visible: that we should characterize the subjective (and then perhaps the valuational) positively, in terms of a subjective judgment’s being one that is however indirectly answerable for its correctness to the responses of conscious subjects; that we should characterize the objective positively, in terms of an objective judgment’s being one that is a candidate for plain truth; and that, having characterized each of these categories of judgments positively and independently, we need to be ready for the possibility that a judgment may fall into both, may both rest upon sentiment and relate to a matter of fact. (2002 [1987c], p.201-202)
As we will see in section 2.2.1., this thesis supports Wiggins’ reading of Hume’ moral anthropology, which includes a similarly weak-realist strand that Wiggins emphasizes. In section 2.3., we will see that the standards of correctness to which valuational responses and judgments are answerable include the “equal and reciprocal partners” relation between response-dependent valuational properties and patterns of response manifesting dispositions attuned to these properties (and so featuring an irreducible element of appropriateness); and we will see how concepts come to designate certain properties, thus attached to the appropriate patterns of responses.

2.1.3. A metalanguage for the interpretation and the critical assessment of discourse

This subsection presents (2.1.3.1.) a minimal summary of Davidson’s theory of radical interpretation, and of the metaphysical and epistemological views that are linked to it, so as to show how Wiggins’ framework is indebted to Davidson’s.

Wiggins, then, adopts a Davidsonian theory, which he proposes to modify at certain junctures.
For one thing, he proposes an alternative “reading” of the biconditional, according to which *meaning and truth mutually elucidate one another* (2.1.3.2.)\(^{31}\)

Basing himself on the idea that meaning and truth mutually elucidate one another, Wiggins offers an elucidation of our ordinary concept of truth, into five “marks of truth”, according to which our ordinary notion of truth is objective and world-involving (2.1.3.3.). The second mark of truth calls for a slightly more elaborate presentation than the others, provided in 2.1.3.4.

Finally, Wiggins distinguishes two kinds of moral judgments: the evaluative and the practical (2.1.3.5.).

2.1.3.1. The Davidsonian theory

As I mentioned in Chapter One, Strawson, as early as 1949, propounded the view that the investigation of ordinary language was to constitute a “commentary” that would stand to ordinary language as a an informal metalanguage would stand to an object-language.

\(^{31}\) This, he claims, goes against the traditional interpretation of the biconditional, from Frege to Davidson, according to which the very concept of meaning, as well as the meaning of sentences, are elucidated on the basis of the concept of truth.
Davidson has proposed a similar (although more formal) view with his theory of radical interpretation. However, the latter does not directly originate from Strawson’s work, but rather from Quine’s theory of radical translation.

Both Quine’s and Davidson’s respective theories can be roughly described as accounts of speech-based linguistic communication. Sentences are a more basic unit of meaning than words, and we cannot assume that speakers of what appears like a common natural language necessarily attribute the same meaning to the same words. Rather, an observer attributes meaning to a speaker’s sentences, by presuming them to be believed by the speaker, and by correlating these sentences with features of the environment presumed to motivate them. This in turn calls for certain methodological principles.

Quine proposes (Quine 1969) to eliminate the notion of meaning, as well as intentional vocabulary. He acknowledges that the intentional area of discourse is irreducible, but considers that it is useless for purposes of “serious, literal” descriptions and explanations. Intentional vocabulary must be eliminated from scientifically credible discourse.

Against Quine’s contentions, Davidson claims that the notion of meaning and the intentional vocabulary, are rationally acceptable and indispensable. The intentional area of discourse is indeed irreducible, but it can be used in a regimented way within a scientifically credible theory, so as to account for the interpretation of speech (and thus for linguistic meaning and understanding generally), and for the learnability of language.
According to Davidson’s theory of radical interpretation, the ordinary interpretation of speech is modeled into a pragmatics and a semantics.

The pragmatics corresponds to a theory of speech acts, and accounts for the interpreter’s recognition of speech acts, notably assertions, in the flow of the speaker’s linguistic and non-linguistic behavior, in the context of its environmental circumstances.

The semantics is provided by Tarski’s truth theory, based on biconditionals that pair sentences in the object-language (the language being interpreted) with sentences in the metalanguage, which state the truth conditions, and therefore the meaning, of the former. The metalanguage is built by the interpreter on the basis of the methodology that I evoked above: uttered sentences are presumed to be believed by the interpretee and they are correlated, with the help of the Principle of Charity, with features of the environment that presumably motivates them.

The meaning of words is a function of the sentences in which they occur — in other words, of how we use them in sentences. The Davidsonian theory of interpretation is therefore compatible with the elucidative conception of analysis, and with what Strawson defends as the “connective model” of the meaning of concepts. This implies that the meaning of terms enjoy a certain plasticity, which it would lack if it was determined by “appropriate causal chains” with items of the extralinguistic world, as stated by the rival “building-block” approach to meaning.
In sum, the theory explains how we are able to solve the equivalent of an equation with two unknowns, by ascribing both propositional attitudes to the speaker, and meaning to her utterances — thanks to the Principle of Charity.

This theory of interpretation is at once descriptive and normative. It purports to describe how ordinary interpretation works — which involves certain methodological principles (which guide and constrain the interpreter’s attribution of meaning to the interpretee’s sentences, in interaction with empirical evidence regarding the interpretee’s conduct and the empirical circumstances of her speech). And it vindicates these principles, and what is presumed to be their ordinary use.

For a long time, Davidson has designated under the heading of “Principle of Charity” two methodological assumptions which he eventually distinguished: the Principle of Correspondence and the Principle of Coherence.

Michael Devitt and Kim Sterelny have proposed to split the Principle of Coherence into two more specific principles, which can be put very concisely. “Coherence-as-rational-belief” «requires the rationality of beliefs given other beliefs; the inferential connections among beliefs must be rational» (Devitt and Sterelny 1987, p.245-246). “Coherence-as-rational-action” «requires that the
connection between an intentional system’s beliefs and desires and its actions be rational.» (ibid)\(^{32}\)

The *Principle of Correspondence* can be described as the assumption that a speaker’s beliefs, taken as a whole, are mostly true.

Some philosophers have proposed modifications to the Principle of Correspondence. Richard Grandy, notably, has proposed to replace it by what he has called the “Principle of Humanity” — which Wiggins adopts, as we will see.

Graham MacDonald summarizes the Principle of Humanity as follows: «the interpreter should not so much maximise agreement whatever the cost, as minimise a certain sort of disagreement, specifically disagreement which we find unintelligible.» (MacDonald and Pettit 1981, p.29) Unintelligible disagreement corresponds to a situation where epistemic disagreement occurs, while no cognitive failure is identified or suspected. This epistemic divergence is unexplained, and ultimately unintelligible. This is the sort of case that the interpreter must minimize as she attributes meaning to the interpretée’s sentences.

This means that the interpreter should expect that most of the statements held true by the interpretée will epistemically converge with her own beliefs, while

\(^{32}\) The generic expression “intentional system” includes, but is not limited to, human agents. We will see in Chapter Three how Dennett uses this notion.
acknowledging that various sorts of cognitive failures may intervene, and disrupt the expected epistemic convergence.

The upshot is that in the case where no cognitive failure is identified or suspected, the interpreter should expect epistemic convergence.

The Principle of Correspondence (or of Humanity) is linked to certain epistemological and metaphysical positions, each supporting the idea that epistemic convergence is the normal aspiration, and an expectable result, of rational discussion.

Epistemologically, the principle implies that skepticism is ill-founded. There is no good principle on the basis of which we could seriously doubt of the truth of all of an agent’s beliefs (as if all of an agent’s beliefs could in principle be simultaneously false). What might appear prima facie as epistemic divergences between speakers, should be considered, in the absence of identified or suspected cognitive failure, as disagreements about the meaning of words.

Originally, Davidson’s metaphysics was minimal. His positions were based on a priori considerations, which emphasized that the Principle of Correspondence was a necessary methodological assumption of interpretation. In later writings, Davidson came to a more metaphysically elaborate view: «the considerations in favor of semantic realism seem to depend in part not on purely a priori considerations but rather on a view of the way people are.» (Davidson 1999, p.194; quoted in Sosa 2003, p.177; emphasis added)
Davidson, then, has come to a metaphysics of experience, according to which no metaphysical or logical gap intervenes in the relation between Subject and Object. Such a metaphysics supports the idea, already implied by the anti-skeptical epistemology, that most of our perceptual beliefs are true; and that our beliefs in general can be taken to be largely justified, and true.

John McDowell has proposed a certain adjustment to Davidson’s theory (McDowell 1998 [1976]). His goal is to draw Davidson’s theory closer to realism, while acknowledging certain features of Michael Dummett’s verificationist critique of realism. McDowell recommends a position of neutrality in regard to the principle of bivalence, and a construal of truth as a limiting case of warranted assertability.33

In regard to the principle of bivalence, McDowell recommends that we hold onto «a refusal to assert the principle, combined with a refusal to deny it.» (McDowell [1976], p.10) His rationale can be summarized as follows.

McDowell draws from the verificationist critique of realism the following idea: in absence of verifying evidence, we have no grounds to assert that a sentence is either true or false. On the other hand, he rejects the verificationist idea that in such epistemic circumstances, we can assert that the sentence is neither true nor false.

33 The principle of bivalence is, in McDowell’s words, «the principle that every significant indicative sentence is either true or false» (McDowell 1998 [1976], p.9)
The move, then, is to detach the notion of truth from the principle of bivalence, and to think of truth as a limiting case of warranted assertability. This way, we have a realist notion of truth-conditions, while acknowledging that «we may not think of truth as independent from verification» (McDowell 1998 [1976], p.10). As we will see, Wiggins adopts this addition to Davidson's theory.


Thus, he agrees with the Davidsonian idea that «the interpretation of what sentences mean must itself go via the propositional attitudes, conspicuous among which is belief» (2002 [1987b], p.148), but he proposes to substitute his talk of concerns, motives, purposes, etc. for the notions of "desires" and "pro-attitudes", which he finds too schematic. As a result, Wiggins’ version of Grandy’s Principle of Humanity reads as follows:

By “making sense of them” would be meant ascribing to the speakers of L, on the strength of their linguistic and other actions, an intelligible collection of beliefs, needs and concerns. That is a collection that diminishes to a bare minimum the need for the interpreter to ascribe inexplicable error or inexplicable irrationality to them (2002 [1976], p.111-112; emphasis added)34

---

34 At one juncture, Wiggins indicates that he takes “judgments”, or “propositions” as the objects of beliefs: «When someone believes truly that p, we may predicate truth of his state of belief, then of
Within Wiggins’ Davidsonian theory, the meaning of terms is determined by the set of all the sentences that they occur in. The inter-relation between ordinary concepts is revealed by examining their occurrences in a variety of sentences: «an elucidation illuminates a concept by employing it in a set of true judgments that involve it revealingly and interestingly with distinct, coeval, collateral concepts.» (2002 [1987a] p.142 n.4). This corresponds to what Strawson described as the elucidative construal of analysis (or the “connective model” of the meaning of concepts). One gathers that Wiggins’ insistence on “coeval and collateral concepts” need not preclude the acknowledgment that, as Strawson remarks, there are also asymmetrical relations between certain concepts, such that some of them appear more basic than others, without for that constituting a “basis of reduction”.

2.1.3.2. The mutual elucidation of meaning and truth

the object of his belief, i.e. the proposition or judgment; and then we may predicate truth of the sentence that expresses or reports the belief.» (2002 [1987b], p.148) In other words, a belief is an endorsed judgment, or an endorsed proposition.
For Wiggins, the investigation or moral discourse requires that we first elucidate our concept of truth\(^{35}\). A trap that we must avoid at the outset is an unexplicative recourse to the notions of “facts” and of “correspondence”. These are legitimate notions, but they stand in need of being explicated, and so cannot help us in elucidating truth:

[W]e must be prepared to postpone the moral or the aesthetic or whatever else entirely until we can reach them by a detour through the concept of truth itself, a detour within which we refrain scrupulously from question-begging use of the idea of fact. What a fact is, is something we may hope in the end to cast light upon, if \textit{en route} we deliberately abstain from relying on the idea in our account of truth. This detour must also be made without any positive reliance upon the idea of correspondence. (2002 [1987b], p.140)\(^{36}\)

Of course, the elucidation of ordinary concepts is a theoretical endeavor, and as such it may involve technical terms. But their role is to contribute in the elucidation of ordinary concepts, not to replace them. On this count, Wiggins remarks that the notion of correspondence can mislead us:

\(^{35}\) Wiggins sometimes calls our ordinary concept of truth “plain truth”, or “truth for the working day”. Cf (2002 [1987d], p.343 n.24)

\(^{36}\) As we will see when we talk about Wiggins’ second mark of truth, this restriction drives him towards the notion of “vindicatory explanation”.
The only way forward is for us to treat is as a methodological assumption, pending disproof, that there is only one relevant sense of the predicate “true”; to take this as the sense of the word in ordinary English; and then as calmly as possible, simply forgetting correspondence, to inquire what the property is like that “true” stands for, admitting technical terms and their associated philosophical preconceptions one by one and only in exchange for good reasons for the preconceptions and bona fide non-creative, non-prejudicial explications (such as have not yet been produced for “corresponds”) of the meanings of all technical terms. (2002 [1987d], p.332)

According to Wiggins, the biconditional, as it occurs in Tarski-like T-sentences, has been traditionally read, since Frege, as if its main purpose was to elucidate the meaning of the object-language sentence cited on its left side by the metalanguage sentence used on its right side, which states the truth conditions, and so the meaning, of the object-language sentence. Wiggins suggests that this traditional reading avails itself of the notion of truth in order to elucidate, not only the meaning of particular sentences, but also the very concept of meaning.

But, Wiggins proposes, if we can take declarative meaning «as already to some modest extent understood» (2002 [1987b], p.141), then the notion of meaning and the notion of truth mutually elucidate one another:

The interest that is normally attributed to such formulations [i.e. biconditionals — P.C.] derives from the promise that they hold out of elucidating meaning in terms of truth. [...] But if we want to understand more about truth and we think we understand more than nothing about what declarative meaning is (and surely we do), then one thing we can attempt is to shift our weight from the right to the left side of the
Fregean biconditional, and look for any elucidations that it will then provide, not on this occasion of declarative meaning by reference to truth, but of truth by reference to declarative meaning, the latter being taken as already to some modest extent understood. We shall not arrive in this way at an “analysis” of truth. But, even if such an analysis were possible, we should not need it for our advertised purpose [i.e. investigating the truth-aptness of moral judgments, which Wiggins tackles later — P.C.] All we need in order to answer our question about ethics is to determine what properties are possessed by every sentence that expresses a truth, or marks (as one might say in Frege’s terminology) of the concept true. (2002 [1987b], p.141-142)

In other words, we start with a minimal understanding of assertive meaning, which allows us to minimally elucidate the concept of truth, which elucidates meaning in return.

Wiggins elucidates our ordinary concept of truth into five “marks of truth”. The notion of “the mark of a concept” is borrowed from Frege. As Wiggins presents it, it is the notion of a necessary condition that would be at the same conceptual level as the concept it is a mark of: «A mark of the concept F is a property that anything has if it falls under F. So the marks of a concept F are concepts of the same level as F.» (2002 [1987b], p.142 n.5). As such, the marks of truth «can help to determine what exactly is at issue when truth is at issue»,
even though «they do not provide [...] any clear effective test (palpable sufficient condition) of truth.» (2002 [1987d], p.340)\textsuperscript{37}

2.1.3.3. The five marks of truth

Wiggins elucidates what he presents as our ordinary concept of truth into five marks, and thus means to show that this concept is at once subject-involving (more exactly, intersubjective) and world-involving.

Briefly stated, Wiggin’s marks of truth are as follows. Mark (1) simply asserts that truth is the primary dimension of assessment for beliefs, judgments, and assertive sentences. Marks (2) states that true judgments are expected to elicit epistemic convergence among participants in the practice of giving and asking for reasons (that is the intersubjective mark of truth). Mark (3) states that a true sentence must be true in virtue of something in the world (that is the world-involving mark of truth). Mark (4) «simply condenses the statement of Marks (1), (2), and (3).» (2002 [1987b], p.152). Mark (5) states that all true statements are expected to cohere.

\textsuperscript{37} In an early version of the elucidation of truth (2002 [1976], p.115), Wiggins uses the expression “truisms of truth” to designate what he calls, from 1980 on, “marks of truth”. The terminological change does not reflect any substantial theoretical change.
There are two distinct formulations, in Wiggins’ writings, of the five marks of truth: one in (2002 ([1976]), the other in (2002 [1987b]) There are no crucial theoretical differences between the two formulations, but placing them side by side helps in understanding what goes into each mark of truth.

Wiggins’ formulations are preceded by the year of first publication, and followed by a short commentary (or further, helpful quotes).

Mark (1)


1987: «If \(x\) is true, then \(x\) passes muster in that dimension of assessment in which \(x\) demands to be assessed by virtue of being the sort of thing it is, thus in the primary dimension of assessment for \(x\);» (2002, p.147)

Somewhat more clearly, Wiggins states that it is «a constitutive norm for beliefs as such to aim at truth» and that it is «a norm for speakers to aim at truth in what they [say]». (2002 [1987b], p.148) About moral judgments in particular, Wiggins emphasizes that «truth is the proper \textit{aspiration} of all moral judgments.» (2002 [1987b] p.141)

Mark (2)

1976: «the answerability of truth to evinced argument that will under favourable conditions converge upon agreement whose proper explanation involves that very truth;» (2002, p.115)
1987: «If \( x \) is true, then \( x \) will under favourable circumstances command convergence, and the best explanation of the existence of this convergence will either require the actual truth of \( x \) or be inconsistent with the denial of \( x \);» (2002 p. 147)

Of the five marks of truth, this is the one to which Wiggins dedicates the most elaborate discussions. The epistemic convergence that Wiggins refers to is, as he makes clear, the sort of intersubjective, substantive agreement that can result from rational discussion, and which presupposes that the meaning of terms is sufficiently stabilized. We will return to this mark of truth in the next paragraph (2.1.2.4.).38

Mark (3)

1976: «the independence of truth both from our will and from our own limited means of recognizing the presence or absence of the property in a statement;» (2002, p.115)

1987: «For any \( x \), if \( x \) is true then \( x \) has content; and if \( x \) has content then \( x \)'s truth cannot simply consist in \( x \)'s being itself a belief, or in \( x \)'s being something believed or willed or... [sic]» (2002, p.148)

In other words, a judgment cannot be true exclusively in virtue of subjective matters (as for example the mental state that the judgment expresses,

38 Wiggins explicitly distinguishes this sort of convergence from that sketched by Peirce, who defined truth as the convergence that science would command, at its ideal limit.
or a desire that the judgment be true, or a mental act of will, or an act of assertion). The first formulation suggests, more strongly than the other, that truth is recognition-transcendent. In any case, true judgments have the potential to elicit epistemic convergence of the right sort, but may in principle fail to actually elicit it (in reason of the cognitive blindness of the participants in the practice of giving and asking for reasons).

This mark of truth implies that our ordinary concept of truth is world-involving, i.e. sentences that are true, in the ordinary sense of that concept, are world-involving.

Mark (4)

1976: «that every truth is true in virtue of something;» (2002 [1987b], p.115)

1987: «Every true belief (every truth) is true in virtue of something;» (2002 [1987b], p.148). Somewhat more informatively, «there must be something distinct from x, namely how things have to be for x to succeed in its aim [of being true] or be correct, if there is to be any interpretation of x as the belief that it is.» (2002 [1987b], p.151)

Wiggins adds: «The statement of Mark (4) simply condenses the statements of Marks (1), (2), and (3).» (2002 [1987b], p.152) As a positive characterization of what truth depends upon, it appears as the positive counterpart of the 1987 version of Mark (3).

Wiggins’ formulations of Mark (4) point in the direction of a certain principle of epistemic objectivity. But the main idea is that in interpretation, beliefs are
presumed to be true \textit{in virtue of “how things are”}, where it is excluded that states of affairs spring from our states of beliefs themselves, or from acts of will.

Mark (5).

1976: «that every plain truth is compatible with every other plain truth.» (2002, p.115)

1987: «If $x_1$ is true and $x_2$ is true, then their conjunction is true.» (2002, p.148)

This is, with epistemic convergence, the other property of well-groundedness that is required for a belief to be true: rational coherence with other beliefs.

2.1.3.4. Mark (2): epistemic convergence and its best explanation

Mark (2) states two conditions: a true judgment — say, the judgment that $t$ is F — is expected to elicit epistemic convergence; \textit{and} the best explanation of this convergence will either require the actual truth of the judgment, or be inconsistent with its denial.

Let us examine these conditions in turn.

As regards the condition of epistemic convergence, it should be clear that it does not require that an actual group of persons converge upon a particular judgment. We may consider the case of that judgment being endorsed, and so
believed, by a single person; and we may ask what best explains that case of endorsement (of belief acquisition, or belief fixation), as we would ask what would best explain the convergence of many persons upon the same judgment. If the best explanation of the acquisition, by one person, of the belief that “t is F” is that there is nothing else to think but that t is F, then we may consider that many persons would converge upon the same judgment, and so that it exhibits the second mark of truth.

In the following quote, Wiggins considers the case of the belief, held by one person, that “the cat is on the mat”:

[Eliminating the reference to such things as facts, let us convey the point by an example [...]]: John’s belief that the cat is on the mat is vindicated as objectively correct if and only if the best explanation of why John has arrived at that belief depends on (and/or is inconsistent with the denial of) the explanatory premiss “the cat is on the mat”. (2002 [1987d], p.345)

As we will see in 2.3.2.3., Gilbert Harman centers his discussion of evaluative moral judgments on a case of endorsement, by one person, of the judgment “that is wrong”.

The second condition of the second mark tells us how the convergence upon the judgment that “item t is F” is to be best explained in order for that convergence to count as a mark of truth
The convergence that is in question — the convergence the lack of which would need explaining — is not a mere intersubjectivity or a chorus of agreement in which each new voice that is added sings the note it joins in upon just for the sake of unison, reinforcing thereby the voices that are already singing that note. If the convergence in the belief that item $t$ is $F$ is to be relevant to truth — if this is to be a case of the interesting, significant convergence that truth commands — then what puts that belief there and holds it there has to be nothing more and nothing less than the fact that the item $t$ really is $F$. Or dispensing, as we must show is possible, with the language of “facts”, we have the truth-relevant sort of convergence where the statement of the best explanation of the agreement in the belief needs a premise to the effect that item $t$ is indeed $F$, and the explanation would be simply invalidated by its absence. Or, relaxing this very slightly, and prescinding from the example, we have the kind of convergence in belief that truth commands where the best explanation of the agreement in the belief that $p$ is inconsistent with any denial on the explainer’s own part that $p$. (2002 [1987b], p.150-151)

This best explanation, Wiggins calls a “vindicatory explanation”: it not only explains the convergence as a phenomenon, but also vindicates it. Wiggins suggests that convergence upon a particular judgment is explained in a way that vindicates the judgment, in the case where this judgment is the only one that reasonably survives critical reflection — when, as he puts it, “there is nothing else to think” but $x$. Wiggins gives few examples. One points to the convergence commanded by an arithmetical judgment:

There is an impressive consensus that $7+5=12$; and, when we rise above the individual level and look for the explanation of the whole
consensus, only one explanation will measure up to the task. There is *nothing else to think* that seven and five add up to. (A full-dress explanation will prove this.) Since any other answer besides “twelve” will induce a contradiction in arithmetic, no wonder we agree. We believe that 7+5 is 12 because 7+5 is 12. We have no choice. (2002 [1987b], p.153)

Wiggins’ reliance on talk of there being “nothing else to think but…” is a suggestive way of obeying the restriction he has cast at the outset upon his account of truth: we must avoid an unilluminating recourse to the notions of “fact” and “correspondence”. Moreover, it implies that, in the case of judgments using a valuational vocabulary, the best explanation of the epistemic convergence cannot dispense with that evaluative vocabulary.

2.1.3.5. Two classes of judgment

In the introduction to this chapter, I have briefly presented Wiggins’ distinction between these two classes of judgments: *moral evaluative judgments* are “is” judgments which express the valuational dimension of morality; they are included in the broader class of “evaluative judgments” in general. And *moral practical judgments* are “ought” judgments, which express the deontological dimension of morality; they are included in the broader class of practical judgments in general.
Wiggins emphasizes that his distinction is not a dichotomy, as the two classes overlap. This reflects the complexity of the relations between the deontological and the valuational domains. If we think of those concerns and motives which Wiggins characterizes as “unforsakeable”, it is most plausible that many of them can be expressed either in the form of evaluative judgments, or in the form of practical judgments.

Wiggins identifies the class of “is” judgments to “facts”. Evaluative moral judgments being a subclass of the class of “is” judgments, Wiggins suggests that values are a subclass of facts. This he makes clear in a footnote:

[...] my own view is that the fact-value distinction is not like a bat/elephant distinction, but like an animal/elephant distinction. On the other hand, [...] the is/must distinction is more like a mammal/carnivore distinction. (1976, p.96 n.7)

Wiggins formulates the distinction between the two classes of judgments in the following quote (note that the class of practical judgments includes prudential judgments). At the end of the quote, Wiggins proposes that we «emancipate ourselves from a limited and absurd idea of what is». This is clearly an objection to the reductionist sort of doctrine according to which moral properties are “ontologically queer” and moral terms are “empty terms”. Consequently, according to that sort of doctrine, values do not exist; and, given a “building block approach” to meaning and to the cognitive status of sentences, those sentences in which moral terms occur cannot be true:
I propose that we distinguish between *evaluations* (typically recorded by such forms as "x is good", "bad", "beautiful", "ugly", "ignoble", "brave", "just", "mischievous", "malicious", "worthy", "honest", "corrupt", "disgusting", "amusing", "diverting", "boring", etc. — no restrictions at all on the category of x) and *directive* or *deliberative* (or *practical*) judgments (e.g. "I must [psy]", "I ought to [psy]", "it would be best, all things considered, for me to [psy]", etc.) It is true that between these there is an important no-man's-land (comprising, e.g., general judgments of the strongly deprecatory kind about vices and virtues, and general or particular statements about actions that it is ignoble or inhuman or unspeakably wicked to do or not to do). But the fact that many other kinds of judgment lie between pure valuations and pure directives is no objection; and it does nothing to obstruct the discrimination I seek to effect between the fact-value distinction and the is-ought or is-must distinction. The unavailability of any well-grounded notion of the factual that will make the fact-value distinction an exclusive distinction can only promote our interest in the possibility of our finding an *ought* or *must* that will not count as some species of *is*. If we then conceive of a distinction between *is* and *must* as corresponding to the distinction between appreciation and decision and at the same time emancipate ourselves from a limited and absurd idea of what *is*, then there can be a new verisimilitude in our several accounts of all these things. (Wiggins 1976, p.95)

Wiggins evokes the content of the intersection of the two classes as including judgments "of a strongly deprecatory kind", etc. I take it that a judgment as simple as "x is right" could count as included in the intersection of the two classes, for the reason that, although it is formally an evaluative judgment, it is
also a deontological (if not necessarily action- \textit{motivating}) judgment, easily translatable as an “ought” judgment.

Wiggins’ metaethical theory investigates whether the two classes of moral judgments can be candidates for knowledge. As we will see in section 2.3., the question whether judgments of the classes in question do or do not exhibit the marks of truth is central.

\subsection*{2.2. WIGGINS’ ANTHROPOLOGY OF MORAL LIFE}

I divide Wiggins’ anthropology of moral life into three parts.

The first part (2.2.1.) shows how Wiggins outlines a theory of human nature out of Hume’s own, and how he acknowledges that the theory of evolution supports, expands and in some regards deepens the Humean theory. In regard to the latter, Wiggins proposes that it «can be cut loose from his philosophy of impressions and ideas» (2006, p.5), and that we can, independently of the “impressions” and “ideas”, focus on those dispositions which (in my interpretation of Wiggins’ moral anthropology) underdetermine the standards of ordinary morality.

Wiggins remains silent on Hume’s “dissolving” analysis of the self. In 2.2.3., we will see that he disagrees with his compatibilist account of free will. In
any case, Wiggins does not simply *adopt* Hume’s theory of human nature. He also *adapts* it to draw it closer to the moral anthropology he envisages.

The second part (2.2.2.) bears on the specifically Wigginsian view of morality. It surveys those concerns, motives, purposes, needs, aspirations, expectations that Wiggins characterizes as “unforsakeable”, and gives a broad substantive view of the “ordinary morality”, which *arises out* of them: «The first-order ethic we are concerned with incorporates a human scale of values and a human deontology. It *arises out* of our human sentiments and predispositions.» (2006, p.242; emphasis added)

Wiggins claims that two broad features of the first-order ethic “follow from” the unforsakeable concerns, in a determinate enough way, following which ordinary morality is “a first-order ethic of solidarity and reciprocity”. He suggests that the deontological and the valuational dimensions of this first-order ethic mutually determine one another.

Wiggins also suggests that unforsakeable motives and concerns are shareable by all humans. The first-order ethic he sketches is in fact a minimal core (standing at the basis of more elaborate, and competing first-order ethics), and this minimal core is universal, and binding for all humans. The thought is *not* that such an ethic is in use in the common life of all people of all cultures. It is, rather, that all humans can discover, in the practice of intersubjective rational criticism, that we share unforsakeable motives and concerns; that these stem from, as Wiggins puts it a common «*given-cum-acquired nature*» (2006, p.234;
emphasis added); and that consequently all human may eventually come to a
common, universal first-order ethic.

The third part (2.2.3.) bears on free will and responsibility, an issue that is
part of Wiggins’ moral anthropology. Morality, on any construal, requires that
humans have free will and be morally responsible in some sense. The main
question is whether this sense must be libertarian, or whether it can be
compatible with determinism. Wiggins defended a libertarian conception in (2002
[1973]). It is not clear whether he still supports it, but what view he otherwise
entertains of free will and responsibility seems lacking.

2.2.1. A Humean theory of human nature

This subsection shows how, in Wiggins’ reading, a strand of thought which
points in the direction of his own metaphysics of valuational experience runs
across Hume’s account of how certain human dispositions (those he groups
under the heading of “natural virtues”, and those he groups under the heading of
“artificial virtues”) underdetermine moral standards. And it presents Wiggins’
opening to evolutionary theory.

Paragraph 2.2.1.1. focuses on those dispositions that correspond to
“natural” virtues.
Paragraph 2.2.1.2. focuses on those dispositions that correspond to “artificial” virtues (one might object that they could be part of “human nature” *strictu sensu*, but in this connection, Wiggins speaks of «our given-cum-acquired nature»). In Hume’s treatment of artificial virtues, Wiggins finds an unsatisfactory subjectivist point. On the other hand, it is a rarely acknowledged feature of Hume’s philosophy, Wiggins claims, that he came to vindicate an ordinary conception of reasonableness, against those “cosmic” conceptions of rationality which a certain exegetical tradition nonetheless attributes to him.

Finally, Wiggins acknowledges (2.2.1.3.) that Hume’s theory requires the support of evolutionary theory, and that they mutually enrich one another, Wiggins clearly endorses a naturalist approach.

2.2.1.1. Natural virtues: benevolence and beyond

From Hume’s account of natural virtues, Wiggins extracts two sets of insights. One is related to the underdetermination thesis, while the other corresponds to the intersubjectivist strand of Hume’s thought, which Wiggins wants to pull in the direction of his metaphysics of experience.

The first set, more precisely, is linked to the thesis of underdetermination, the thesis of anthropocentricity, *and* the metaphysics of valuational experience. Hume’s thought is that morality originates in the sentiment of benevolence, which overcomes the sentiment of self-love thanks to the *intersubjective character of*
language, which entails a common perspective among speakers. The intersubjectivity of language thus reinforces benevolence against self-love, but it also implies that moral judgments are not purely, individually subjective.

However, this insight is merely hinted at in the Treatise, and remains embryonic by comparison with the reductionist, subjectivist strand developed in the same book (as we will see in 2.3.2.1.). The intersubjectivity of language and of standards is more explicitly asserted in the Enquiry.

The second set includes insights which suggest that morality «reaches beyond» (2006, p.50-53) the sentiment of benevolence from which it originates, and therefore beyond purely subjective mental states. It reaches beyond sentiment in three ways, which, as Wiggins presents them, have to do with the following things: the involvement of affections with understanding; the aspiration to universality of valuational standards; and the redirection of the sentiments of self-love and of benevolence by the social rules that emerged at the origins of large, “political societies”. This is a step in the direction of the Wigginsian metaphysics of valuational experience. However, the insight about the redirection of sentiments would have required a development that Hume ultimately fails to provide, as we will see in the next paragraph (2.2.1.2.).

Let us examine these two sets of insights in turn.

According to Hume, we are naturally endowed with the sentiment of benevolence, a disposition associated with the capacity of sympathy (in Wiggins
Hume claims that benevolence has a foundational role in human morality. This foundational role is double: benevolence motivates our conduct, and contributes in fixing the standards to which our moral responses and judgments are answerable. That our moral standards are marked by benevolence is suggested by the empirical fact that the majority of people respond with approbation to those features of characters and actions, called “virtuous”, that manifest benevolence in regard to others — not necessarily themselves. And they respond with disapprobation to those features of characters and action, called “vicious”, that manifest malevolence in regard to others — again, not necessarily themselves.

39 In the order of explanation, Hume takes the human disposition to benevolence as a primitive: «It is needless to push our researches so far as to ask why we have humanity or fellow-feeling with others. It is sufficient that this is experienced to be a principle in human nature.» (Enquiry, V.2 Selby-Bigge p.219-220; quoted in (2006, p.37-38)) But evolutionary theory has showed that it was indeed possible and necessary to push the research further, and has delivered credible answers to the question why humans have altruistic dispositions. Wiggins acknowledges that Hume’s theory of human nature has to be connected to, and revised in the light of, evolutionary theory, as we will see in 2.2.1.3.
Hume acknowledges that we are also endowed with the “sentiment of self-love”. However, he opposes the double claim (made by such philosophers as Hobbes) that self-love entirely, or primarily, determines the motives upon which humans act; and that it determines the moral standards to which our responses, judgments and actions are answerable, as they supposedly come down to sophisticated ways of protecting one’s self-interest (notably one’s long-term interests).

Against this “selfish theory” of motivation and morality, Hume claims that the sentiment of benevolence is clearly irreducible to the sentiment of self-love; and that, as such, it determines the standards of morality and motivates much of human action. Wiggins emphasizes that Hume does not claim that the disposition to benevolence is naturally “stronger” than the disposition to self-love. On the contrary, he claims, in the Enquiry, that benevolence is a comparatively “weak” sentiment, while self-love is comparatively “strong”.

On the other hand, Hume argues that the intersubjective character of language, and in particular of the valuational vocabulary, reinforces (so to speak) our benevolent dispositions, and contributes in making them stronger than the sentiment of self-love, both in fixing our moral standards, and in motivating our action. The argument, as Wiggins reconstructs it, goes as follows: the very “logic” of ordinary valuational terms implies that the judgments in which they occur are answerable to intersubjective standards of correctness; the intersubjectivity of standards implies in turn a commonality of perspective between speakers; this
commonality of perspective in turn contribute in making benevolence stronger than the sentiment of self-love:

In the process of learning the sense of the public language in which there is provision to talk of useful and useless, good and bad, fair and foul, beautiful and ugly, a human being enters into the commitment to learn to depart from his private and particular situation, and see things not only from thence but from the point of view that shall be common between one person and another. (See Treatise, III.iii.1 ad fin. and Enquiry cited below [my next quotation — P.C.]) Simply by virtue of their intersubjective significance, these terms, “good”, “bad”, “fair”, “foul”, “beautiful”, “ugly”, force anyone who will seize their proper meaning to transcend that which is good, bad, fair, foul, beautiful, ugly for him. In learning to speak the public language of approbation and disapprobation, praise and blame, (and arrive at the possibility of agreement with others in judgments), a human being needs to learn to see his responses as answerable to a point of view that lies beyond his own, a point of view that he shares first with others to whose fate he is not entirely indifferent or whose fate he cannot ignore — cannot ignore if he coexists with them or he engages in any cooperative venture with them. (2006, p.45; emphasis added.)

Of course, agents do make speaker-relative judgments, «that arise in the private and particular viewpoint of self-love.» (2006, p.46) However, «as soon as we come to see things from the common viewpoint that is presupposed to the senses of predicates that presuppose it, we are well set to learn a new set of responses and treat our own judgments as answerable to responses beyond our own responses.» (2006, p.45-46; emphasis added.) This insight is mostly
asserted in the *Enquiry*, from which Wiggins quotes the following passage (with his own clarifying additions between brackets):

> When [one] bestows on any man the epithets of vicious or odious or depraved, he then speaks another language [than that of self-love], and expresses sentiments in which he expects all his audience are to concur with him. He must, therefore, depart from his private and particular situation, and must choose a point of view common to him with others; he must move some universal principle of the human frame, and touch a string to which all mankind have an accord and symphony. If he mean, therefore, to express that this man possesses qualities whose tendency is pernicious to society, he has chosen this common point of view, and has touched the principle of humanity in which every man in some degree concurs. (*Enquiry*, IX.1, Selby-Bigge, p.272; quoted in (2006, p.46, brackets original))

According to Wiggins, Hume came to see in the *Enquiry* that the intersubjective character of language, and notably of the standards to which our responses and judgments are answerable, implies that *intersubjective agreement over a certain judgment* counts as evidence of its correctness, while *intersubjective disagreement* counts against according that epistemic status. This does not bar the following possibility, which, Wiggins tells us, Hume did not explicitly acknowledge: «Sometimes, we will stick to our own judgment as sound
and incur the responsibility of explaining the discrepancy between our own response and that of others.» (2006, p.47)40

Wiggins presents as follows the three ways in which morality reaches beyond the sentiment of benevolence from which it originates.

The first way has to do with the involvement of affections with understanding, notably through what Hume called “calm passions”: «the standard of morals is established in the ascent from the level of primitive sentiment to the level of plenary moral thought. […] This is surely the same moment at which the passions Hume calls the calm passions — passions easily confounded, Hume says, with the operation of reason — enter fully in their own.» (2006, p.50)

Hume’s idea of calm passions qualifies the opposition that he otherwise maintains between reason and passion, but this is an insight which, unfortunately, he does not develop. Wiggins remarks that Hume could have drawn inspiration from Aristotle:

Here Aristotelians and others will want to salute Hume’s near-readiness to acknowledge the possibilities which Aristotle marked with the notions of deliberative desire, desiderative reason, and ratiocinative desire […] From the calm passions Hume could surely

40 This possibility is acknowledged in Grandy’s Principle of Humanity. Wiggins seems to be suggesting that Hume’s view on the intersubjectivity of language, properly amended, can be pulled in a Davidsonian direction.
have advanced to something very like the practical reason of Aristotle. […]

Hume does not avail himself of such possibilities. Or rather he does not put these things [i.e. calm passions — P.C.] down to the credit of reason. Continuing his polemic against philosophers such as Samuel Clarke and Ralph Cudworth and notions such as eternal fitness, Hume cleaves here to his habitual opposition of reason and passion. He always persists in the idea that the operations of reason and passion can be isolated from one another, considered separately, and then composed. (2006, p.50-51)

We just saw that Hume acknowledges that the standards of correctness of judgments, in general, are intersubjective. According to Wiggins, Hume also thinks that the limits of this intersubjectivity are those of the human species. This is the second way in which morality reaches beyond the sentiment of benevolence (again, Wiggins attributes the thought to Hume): «The [moral] standard transcends that which is local to the person who applies it. […] The standard transcends the local in the sense that it applies itself to cases far outside the time or place of those who arrive at it. It aspires to be a standard that is universal.» (2002, p.53)

As for the third way in which morality reaches beyond the sentiment of benevolence, it is linked to the social rules which presumably emerged with complex sociality. Hume adhered to the Hobbesian sort of theoretical reconstruction following which the historical transition from primitives tribes to larger, more complex “political societies”, required the establishment of certain
rules (or “conventions”), motivated by self-interest — *at least originally*. Hume’s point is precisely that those rules *redirected* natural propensities to self-love and to benevolence, and brought about new forms of valuation that reach beyond both sentiments. In Wiggins’ words: «the energy that is mustered from fellow feeling and its allies (mustered from these and self-love, rather) can be *redirected* and put at the disposal of concerns that are essentially alien (Hume will insist) to benevolence. (2006, p.54)»

This insight is at the core of Hume’s account of artificial virtues, but in order to be a decisive step in the direction of the equal and reciprocal partners view, it would have required a development that Hume ultimately fails to provide, as we will see in the next paragraph.

2.2.1.2. Artificial virtues and human rationality

Wiggins endorses the core of Hume’s account of artificial virtues, but he also criticizes him for remaining too much of a subjectivist, whereas he could have taken the direction of his own “equal and reciprocal partners” metaphysics of experience. First, I will summarize the kernel of Hume’s account of artificial virtues, which Wiggins endorses. Secondly, I will present what Wiggins presents as Hume’s subjectivist error. Thirdly, I will present how Wiggins proposes to develop the core of Hume’s account of artificial virtues, in the direction of his
equal and reciprocal partners view. Finally, I present Wiggins’ reading of Hume’s conception of rationality, which is different from the traditional reading.

According to Hume, propensities to self-love and to benevolence were redirected by the conventions practically entailed by the emergence of large, complex societies. Thus, “artificial” virtues (justice, promise-keeping, allegiance, loyalty, honesty…) were generated by the valuation of convention observance. This requires to be learned, and so became the object of moral education.

We now come to the development that Wiggins disagrees with. Hume suggests that moral instruction involves the teaching and learning of the rationale that supports the valuation of convention observance, which goes roughly as follows: because social life requires that its members do not act solely upon the motives of self-love and of benevolence, the observance of social conventions is practically necessary, and is to be valued. The result of this moral instruction is that acting upon the motive of convention-observance is valued.

Wiggins objects to the idea that moral education teaches an explicit, willful adhesion to social rules for the reason that they are practically necessary; and that it teaches to approve or disapprove of the motives of agents, according to whether these motives are convention-observing, or convention-infringing.

According to Wiggins, Hume’s idea that moral education teaches an instrumentally rational adhesion to social rules, from which it results that the locus
of artificial virtues is confined in the motives upon which one acts, mistakenly excludes the possibility that action itself might display qualities valued for themselves as genuinely moral qualities (or "properties"); and that it might be to those properties of action to which we are in fact taught to respond by moral sentiments of approbation or disapprobation.

Wiggins does not deny that social agents, upon being given a certain description of the rules they follow, would agree with those rules. Nor does he deny that upon being given a certain description of the reasons why those rules originally came about, they would say that those are good reasons. His point is that moral upbringing does not teach an instrumentally rational adhesion to social rules.

Wiggins’ alternative proposal is that moral education involves the internalization of social rules by moral agents. It shapes the motivation of individuals, and it sets standards of responses and judgments, which come down to approving or disapproving of action (rather than the motives upon which one acts) according to whether it observes or infringes conventions.

Thus, moral education sets standards of correctness for responses and judgments which designate as good properties of action (or conduct) that directly correspond to artificial virtues (i.e. properties of justice, promise-keeping, allegiance, loyalty, honesty…) Convention-observing action is valued for the properties it bears. In Wiggins’ terms: «Such practices and the acts they call for take on a life of their own in the hearts and minds of ordinary agents who are
happy to live under the auspices of the civilization that Hume sees as offering to them the final fulfillment of their natures as reasonable beings.» (2006, p.78)

Wiggins’ point is that moral action and moral responses and judgments are rational, in the sense that they are supported by reasons that humans have given to themselves, and which are good reasons from the “inner” point of view of humankind, the inner, participative perspective of human beings. These new reasons for acting and for believing are not directly derivable from interests, by an algorithm-like form of instrumental rationality (of the sort that utilitarians and other consequentialists will later invoke), nor are they “deducible” by “pure reason”. Only by reference to “cosmic” standards of rationality does it seem that moral action and judgment are not genuinely “rational”.

According to Wiggins, Hume’s overall philosophy points to a conception of rationality as ordinary reasonableness, which partakes in the vindicatory strand of his thought (note in the following quote Wiggins’ evocation of the reasons humans have given to themselves, as a sort of “moral beauty”):

It is an illusion, one might say on Hume’s behalf (and despite his self-imposed abstinence from the language of practical reason), to suppose that there is any well-founded conception of the reasonable that is so independent of human nature, and so independent of the substance of human life as we know it, that it would enable a judgment to be reached on some further question of the reasonableness (or not) of observances which agents can perceive as having the kind of beauty we are here concerned with or as belonging to a system that is “advantageous to the whole and to every part”. (2006, p.78-79)
However, two features of Hume’s philosophy tend to keep this hidden from view. One, mentioned in the previous quote, is Hume’s «self-abstinence from the language of practical reason». The other is his insistence that moral judgments are not answerable to the rational standards (of the “cosmic” sort) that the moral rationalists of his day claimed they were answerable to:

Here surely — in Hume’s reading his opponents as claiming to find some higher or more metaphysical reasonableness than this in morality — is the underlying animus for the denials concerning reason that have resulted in Hume’s being mistaken (perfectly absurdly) for some sort of subversive in relation to morals or read as one contemptuous of the ordinary sorts of reason that morality supplies to human subjects. (2006, p.80)

In the end, the Humean view of rationality, according to Wiggins, is a view of «the reasonable as a sum of good reasons which agents can appropriate as reasons of their own.» (2006, p.66)

2.2.1.3. Evolutionary theory

According to Wiggins, evolutionary theory enlarges upon Hume’s view, deepens it, and generally supports it. Here, Wiggins refers not solely to the work of Darwin himself, but to evolutionary theory generally, including its present-day developments. On this count, he endorses some recent work by Elliott Sober and
David Wilson, who defend in (Sober and Wilson 1998) what they call “the thesis of altruism”, according to which «some people at least some of the time have the welfare of others as ends of themselves» (Sober and Wilson 1998, p.228; quoted in (Dennett 2002, p.194-195)). Wiggins claims that Hume would certainly have endorsed the work of Darwin, as well as that of Sober and Wilson, as it supports his scheme, according to which morality originates from benevolent dispositions inherent to human nature, and reaches beyond it:

If Hume […] had been permitted to read advance copies of Darwin’s *Origin of Species* and Darwin's subsequent writings, he would surely have emphasized — or so one speculates — how close the founding sentiments of human morality are to the maternal and paternal instincts of most mammals that manifest awareness of family or larger groups. It also seems plain, as regards the further reaches of the ethical above and beyond the natural virtues of a parent, that he would have found almost everything he needed in a passage of Darwin’s *Descent of Man* to which Elliott Sober and David Sloan Wilson have recently drawn attention […] (2006, p.55-56)

Sober and Wilson propose to modify, with their thesis of altruism, a hypothesis of Darwin’s, according to which, although a “high standard of morality” (Darwin’s expression) does *not* give much of an adaptive advantage to individuals, it can give such an advantage to groups: communities where high standards of morality are enforced might enjoy a certain adaptive advantage over those where moral standards are not as high. Sober and Wilson investigate how, given the dynamics of the relation between a group and its members, unselfish
behavior can confer an adaptive advantage not only on the group, but on individuals as well. Wiggins commends the insight, and finds plausible Sober and Wilson's advantage-to-individuals hypothesis:

[I]t actively encourages speculation about possible ways in which certain human attributes that confer advantage on a group at the level of ethos, culture, and custom can simultaneously (i) put the group at an advantage over other groups and (ii) confer advantage even upon members of the group who both have and actualize within the group an above average potentiality to develop these attributes. Why cannot such individual members both enlarge by procreation the class of individuals with potentialities for this form of ethical development and enlarge by their conduct and example the chances (on the level of culture and custom) for the further, better, and more general actualization of such potentialities? (2006, p.56-57)

Evolutionary theory, then, shows why and how the emergence of morality might have been a natural, evolutionary development in the beginning of the human species. Far from implying that Hume's theory of human nature is vain speculation, evolutionary theory confirms some of its insights, and shows that his moral phenomenology, his detailed attention to the workings of distinct "passions", and to the standards of morality, is all the more fertile that it is supported by it.
2.2.2. The unforsakeable and the first-order ethic

Wiggins elaborates upon Hume’s view: the responses that manifest our dispositions are answerable to standards which are ultimately agreed upon in the intersubjective practice of rational criticism, because they are conditioned and constrained by unforsakeable concerns and motives.

Paragraph 2.2.2.1. presents Wiggins’ fourfold classification of those concerns and motives that he takes to be unforsakeable. Wiggins does not say precisely what “unforsakeable” means, but his treatment of the matter suggests three things.

First, there is a descriptive, empirical sense of “unforsakeable”, which is the one that Strawson evoked in connection with moral sentiments: they are so deeply rooted in us that they cannot be suspended on a generalized and permanent basis. If they underdetermine a first-order ethic, that is something that, as a matter of fact, we cannot help but live with.

Second, there is a normative sense of “unforsakeable”. Wiggins, as I understand him, suggests that not only is human existence not inconceivable without those concerns and motives characterized as unforsakeable, but that it ought not be; that there is a sense of appropriateness to the experienced irremovability of those concerns and motives; and that we may take this “ought not” and this sense of appropriateness as good reasons in support of the basic core of a first-order ethics. There are competing ethics here, but they share a
common core, which can be sorted out by a limitless process of rational criticism. It is underdetermined because it has no fixed propositional articulation. But it is conditioned and constrained, within a limited space of possibilities, by the unforsakeable motives and concerns that it arises out of. And acknowledging that it arises out of such concerns and motives is sufficient to vindicate that basic core.

Third, “unforsakeable” does not mean “foundational”. The standards of correctness underdetermined by unforsakeable dispositions are not rooted in foundational principles independent of our cognitive and emotional constitution.

In sum, Wiggins suggests that some of our concerns and motives are unforsakeable both descriptively (as an anthropological matter of fact) and normatively (as an anthropocentric but nevertheless genuinely moral, and binding, matter of fact), even though the norms in question are not grounded in foundations. Elaborating upon Hume’s natural/artificial distinction, he distinguishes four categories of such unforsakeable motives and concerns.

Paragraph 2.2.2.2 presents Wiggins’ argument to the effect that the first-order ethic underdetermined by our unforsakeable concerns and motives is an “ethic of solidarity and reciprocity”. In other words, Wiggins suggests that unforsakeable concerns and motives exert such constraints on the first-order ethic that it can only be an ethic of solidarity and reciprocity. Moreover, Wiggins argues that the solidarity requirement presupposes the vindication of our ordinary notions of responsibility and agency. However, I do not find that last argument
satisfactory (hence the examination of Wiggins’ libertarian conception of free will and moral responsibility in 2.3.)

Paragraph 2.2.2.3. presents Wiggins’ claim that the valuational dimension and the deontological dimension of the first-order ethic mutually determine each other.

2.2.2.1. The sphere of the unforsakeable: four categories of motives and concerns.

Wiggins enumerates four categories of motives and concerns, the first two roughly corresponding to Hume’s natural virtues, while the last two roughly correspond to the artificial virtues. The order of their presentation (each category is given a number, from (1) to (4)) must not be confused with what I call their *order of prescriptive priority*, nor with what I call their *genetic order*.

By “genetic order”, I refer to the fact that, following the Humean scheme endorsed by Wiggins, artificial virtues “follow from” natural virtues, in the sense that we have seen: the concerns and motives that give rise to natural virtues are redirected by social rules, in a way which results in distinct concerns and motives, which give rise to artificial virtues.

The order of prescriptive priority follows from the fact that each category of motives and concerns makes distinct claims on action, and it may happen that
they prescribe distinct, incompatible decisions and courses of conduct. In such situations, certain categories have a priority over others.

First, I will present how Wiggins describes the four categories. Then I will indicate what is, according to his phenomenological observations, their order of prescriptive priority.

Category (1) is that of prohibitive aversion (note the reference to the moral phenomenologist as observer of ordinary morality in the following quote):

Is there not something altogether remarkable in the strength and variety of our primitively prohibitive aversions — our aversion to wounding, injury, murder, plunder, or pillage, the abhorrence that we experience against the neglect or abuse of children or other defenceless persons, the horror that we feel at the slaughter of the innocent or the repaying of good with gratuitous evil? Here are acts whose awfulness we take for granted […] Drawing on our intuitive and imaginative shared grasp of moral matters, a moral phenomenologist needs to register the strength and persistence of these responses and the inhibitions associated with them. (2006, p.11)

Category (2) is that of cooperative benevolence:

The obvious candidate for the second place in this preliminary enumeration are the simple and positive concerns that arise in us from primitive fellow feeling and find their expression (independently of self-love and sometimes in opposition to it, or that is how it appears) in our willingness to interest ourselves in the plight of known or identifiable
persons, and our readiness, face to face, to enter into cooperative relations with others. (2006, p.12)

Category (2) merges into category (3), which is that of generalized beneficence, i.e. engagement with the public interest: «Almost imperceptibly, the concerns placed second seem to merge into a third group of divers much more abstract preoccupations, those proper to public spirit, to devotion to the general interest, or to humanitarian causes. Concerns of these varieties combine very naturally with those of the less abstract second kind, which they have seemed to some philosophers to generalize.» (ibid)

Wiggins calls observance category (4). In the following quote, Wiggins evokes the order of prescriptive priority of the four categories, according to which this fourth category (in his order of presentation) occupies the second rank in prescriptive priority, after category (1), but before category (2) and (3) (we shall return to this topic shortly):

Finally, fourth by my count, last but not least, come certain intuitive and special preoccupations to which we are party, less clamant than the primitive aversions we began with but perceptibly more clamant than benevolence in the abstract or the concerns for the general good of which some philosophers have made so much: preoccupations with justice (in a relatively narrow and specific sense of “just”) or veracity or allegiance or fidelity to promises. Preoccupations of this fourth kind keep company with kindred preoccupations and solicitudes which condemn the omission of particular duties or obligations and the commission of particular acts such as stealing, betrayal of trust, lying, treason, the breaking of promises… Here the rights and expectations of particular people and groups of people will loom large. In the ordinary moral consciousness as we know it, such preoccupations and
solicitudes limit or narrowly constrain the working of the second and third kinds of moral concern. (ibid)

These four categories of concerns and motives, then, partake in the constitution of a first-order moral sensibility, which, following Wiggins' thesis of cognitive underdetermination, conditions and constrains the content of a first-order ethic, without determining it. The first-order ethic corresponds to an essentially practical domain of thought and action, and does not come down to moral discourse:

Motives and concerns of the four kinds thus provisionally enumerated, once they are translated into settled dispositions and habits of mind, permeate and condition the whole mentality of a human being. [...] Together, one might say, they help to make up an ordinary morality or a first-order moral sensibility — a state of being that is beyond finite descriptions in words and affords an endless fund of responses, verbal and non-verbal responses, to whatever we encounter. However inaccurate and incomplete our fourfold attempt may appear, it is scarcely to be questioned that some such system of ideas and practices is ours to deploy. For the appearance is this: that we deploy this system constantly in act and feeling without recourse to calculation; that it oversees and regulates the business of everyday life; that it forms a whole framework within which to deliberate; that it excludes all sorts of act from the space within which we want to rehearse or refine the practical possibilities that are open to us; that it frequently narrows those possibilities to unity. (2006, p.13)

Wiggins establishes the order of prescriptive priority of the four categories by making explicit features of what he takes to be a shared, ordinary moral
outlook. The order is as follows. First, comes the category of prohibitive aversion (1), which is prior to that of observance (4), which is in turn prior to those of cooperative benevolence (2) and of generalized beneficence (3):

[I]n ordinary ethical thought, there is a marked tendency for concerns belonging to category (1) not only to trump the concerns belonging to category (2) and category (3), but also to constrain the preoccupations that go with artificial virtues such as loyalty, veracity, or fidelity to promises. These last were assigned to category (4), remarking also that there is a tendency for concerns of category (4) to trump those of categories (2) and (3). To default in loyalty, veracity or fidelity to promises moves us towards something in certain ways analogous to the kind of criminality from which we are deflected by aversions falling under category (1). (Wiggins 2006, p.246-247)

Now, according to Wiggins, the way categories (3) and (4) — corresponding to artificial virtues — “genetically” follow from categories (1) and (2) — corresponding to natural virtues — entails concerns of solidarity and reciprocity, which cross over all four categories, and thus occupy a central place in the first-order ethic.

2.2.2.2. The solidarity requirement, and the vindication of our notions of responsibility and agency.
Wiggins’ proposal about how unforsakeable concerns and motives underdetermine a first-order ethic of solidarity and reciprocity is in line with insights by Hume, Kant — and, I add, Strawson. These considerations bring Wiggins to suggest that the nature of ordinary morality, as an ethic of solidarity and reciprocity, “presupposes” that our ordinary concepts of responsibility and agency (following which we humans are morally responsible, free agents), are vindicated.

Wiggins (as we have seen) endorses Hume’s thesis that human nature includes benevolent dispositions and a capacity of sympathy, which together entail that most humans approve of that which manifest benevolence in regard to others, and disapprove of that which manifests malevolence in regard to others. Wiggins is also ready to endorse the idea that Kant’s second formula of the moral law (the Formula of Humanity as End in Itself) is at the basis of human morality, on the condition that the prescription to take persons as ends in themselves be interpreted not as the conclusion of a transcendental argument, but (rather more simply) as an ideal action-guiding principle (2006, p.109-117; 231-233).41

41 To this I add that Wiggins most probably endorses the point that Strawson makes in (Strawson 1962): we are naturally endowed with this double disposition to consider that a minimum of regard is due to human beings as human beings and to expect that this regard be somehow manifested in everyone’s dealings with other people.
In line with those insights, Wiggins adds that our unforsakeable concerns and motives entail a requirement of solidarity among human beings; and that our natural attitude includes an expectation of mutual recognition between humans, and consequently an expectation of, and concern with, reciprocity. These two thoughts were inspired in him by Philippa Foot and by Simone Weil, respectively.

In (Foot 1985), Philippa Foot suggests that a morality «which refuses to sanction the automatic sacrifice of the one for the good of the many» (Foot 1985, p.86; quoted in (2006, p.240)) entails a requirement of solidarity among human beings. Wiggins concurs, and remarks that the concerns and motives of categories (1) and (2) condition the same sort of morality, and entail the same solidarity requirement:

That which we are to discover need not of course be something that the adherents of a first-order ethic will call by the name “solidarity” or some synonym. It will be enough and more than enough if, in trying to understand the spirit that needs to animate a way of being that is sustained and perpetuated in the manner we have already described [i.e. the construal of morality as essentially practical — P.C.], we find, as I shall declare [...] that we do, that the idea or expectation of solidarity must either be at work there — or else placed there in some temporary and special abeyance. (2006, p.242)

Wiggins proposes to buttress the idea of a solidarity requirement by this phenomenological point that it is part of the “natural attitude” to acknowledge the presence of another human as such (by contrast, the presence of animals is not
acknowledged in the same way as that of human beings). We naturally expect our own humanity to be recognized as such by other humans, as we recognize theirs.42

According to Wiggins, this natural mutual recognition between human beings, together with our natural dispositions to benevolence and with the workings of “sympathy”, entails the equivalent of Foot’s solidarity requirement, and this requirement makes a binding claim on all humans:

We should see Foot’s solidarity requirement as spelling out one non-negotiable corollary of the transition that human beings make from mutual recognition under the aspect of human beings to a social morality that requires personal beings to live out that recognition in a solidarity that is all of a piece with the recognition that requires each personal being to recognize the other as participating in that morality and equally requires the other one to recognize him or her as a participant in it. (2006, p.245)

In the end, Wiggins claims, the first-order ethic conditioned and constrained by our unforsakeable concerns and motives can only be a “social morality” of solidarity and reciprocity.

42 This point has been made by many phenomenologists and social thinkers. Wiggins distinguishes as of particular depth and insight Simone Weil’s discussion of it in her text “The Iliad, or the Poem of Force”, a passage from which he quotes (2006, p.243)
Wiggins claims that the solidarity requirement vindicates our ordinary sense of agency and responsibility. The thought is that the solidarity requirement entails that the intention with which one acts is important, and that attributing intentions to an agent requires that this agent be taken as responsible, in the ordinary sense:

[…] I think that next door to the expectation of solidarity we shall find some of the other ideas that we need in order to vindicate our original notions of agency and responsibility. If solidarity conditions our whole outlook, then we shall want to know whether the intention with which a fellow being acted did or did not bespeak the outlook towards other rational beings that is proper to one who does not fail us in solidarity. And once this idea of intention with which is added to the idea of the quality of a thing done, we cannot help but find here the responsibility of a person for a particular outcome that can be fairly attributed to his or her agency. (2006, p.248-249)

A specific critical comment is in order here. Designating some of the concerns and motives with which humans are presumably endowed as unforsakeable, and claiming moreover that morality arises from them, does seem to presuppose that humans satisfy our ordinary notions of free agency and responsibility. The critical point is that this appears as a theoretical requirement that must be met with, more than as an entailment to which we may avail ourselves. In the case where our ordinary sense of free will and agency is not independently vindicated at the outset, the move which consists in characterizing some of our concerns and motives as “unforsakeable” — including in the normative sense — seems weakened.
As we will see in the next subsection (2.2.3.), Wiggins did propose, in (2002 [1973]), such an independent vindication of our ordinary sense of free will and responsibility, in the form of a “reasonable libertarianism”. It is not clear whether Wiggins means his “argument from presupposition” to replace his libertarian thesis of 1973, or whether it is supposed to complete it.

2.2.2.3. The mutual determination of the deontological and the valuational

The first-order ethic which, according to Wiggins, arises out of our unforsakeable concerns and motives has a content (a valuational dimension) and it makes a motivational claim upon us (the deontological dimension). According to Wiggins, the valuational dimension and the deontological dimension are irreducible, and interdependent, as they mutually determine one another. That is, the motivational claim that morality makes on us is determined by its content; and conversely, the content of morality is determined by the motivational claim it makes on us (hence Wiggins speaks at some junctures of the “motive-determined content” of morality, and of the “content-determined motive” of morality).

I opened this chapter with Wiggins’ “classical formulations” of the issues that are traditionally considered as central in moral philosophy. Now, he proposes two distinctive, supplementary formulations of these issues. Both express his view that the valuational dimension and the deontological dimension of morality
are irreducible, and mutually determine one another. The first formulation (the first pair of statements in the following quote) is “general”, while the second formulation is tailored to the way Plato raises those issues in the Republic (since Wiggins endorses both formulations, the slight variations between them help in clarifying his overall view — note that the second formulation permutes the order of presentation of the issues):

(A’) What is the substance or content (and what is the nature and extent) of that morality, if there is one, which there is reason for us (as under (B’) and any positive answer there may be to that question) to adhere to?

(B’) What reasons are there (if any) for us to adhere to and to persist in a morality that has the content alluded to under (A’)?

Alternatively, in a spirit that would have us begin from a position slightly closer to Republic book II, we may ask:

(B’’) What sort of motive can there be for us to care about justice/morality?

(A’’) How do we need to think of justice/morality, and what content must we think that it has, if we are prepared to see it as furnishing us with any motive at all to persist in any of its distinctive works or aims? (2006, p.25)

Wiggins’ positive view is somewhat clarified by his critique of consequentialism, more precisely his critique of the consequentialist thesis that the valuational determines the deontological. In fact, Wiggins endorses W.D. Ross’ critique of that very thesis, in The Right and the Good (1930). Ross
famously summarized his own view with the dictum that «right does not stand for a value at all».

Following Ross, Wiggins claims that the valuational “presupposes” the deontological. This, however, does not exclude that the deontological may reciprocally presuppose the valuational. The following quote expresses his critique of the consequentialist thesis only, not his positive view (which is that the two domains mutually determine one another):

[T]he duty of a parent with respect to their [sic] children is not a duty to promote the intrinsic value that consists in parents' providing for their children or a duty to promote the intrinsic value that is augmented by more and more parents’ cultivating and exercising the virtue of which looking after their children is a characteristic exercise. It is the duty that any parent has with respect to their own children. Indeed, so far from being its foundation, the value attributed to this virtue’s being cultivated and exercised seems to presuppose the said duty. (2006, p.219-220)

In the following passage, Wiggins vaguely evokes his positive view, but he mostly reiterates the critique of the consequentialist thesis of the priority of the valuational. Wiggins’ assertion of “the distinctiveness of the deontological itself” is clearly an assertion of the irreducibility of the deontological. It does not imply that the deontological asymmetrically determines the valuational (a view he rejects, even though this is not expressed in the following quote); nor does it imply that the deontological is entirely autonomous in regard to the valuational. Note Wiggins’ claim that the right approach to the issue is phenomenological:
Whatever truth lurks in Ross’ dictum “right does not stand for a form of value at all”, a similar truth will lurk in the contention that “wrong does not stand for a form of disvalue”. *This is not a plea to the effect that no disvalue resides in acts of criminality, only to the effect that the original and first work of “right” and “wrong” (as of “must”, “must not”, “ought”, “ought not”) lies at some distance from value.* It lies within the deontological as such. The distinctiveness of the deontological itself will appear much less strange altogether if it is approached first, as here, from the side of the phenomenology of that which is utterly forbidden — the original paradigm of the utterly forbidden or *nfastum* being taken to be that which menaces the very fabric of the ethical by threatening to destroy the basis of the ethical in solidarity. (2006, p.247-248; emphasis added)

Wiggins does not elucidate the precise nature of the relationship between the valuational and the deontological. “That is work for another occasion”, as he puts it in (2006), alluding in passing to the complexity of the relationship in question (the following quote begins by Wiggins’ evocation of what he sees as the necessity of widening the conception of “the forbidden”): «In widening this conception — as it will have to be widened — it would be a delicate matter to maintain the distinctiveness of the deontological while tracing its relation to the simply valuational. But that is work for another occasion.» (2006, p.248; emphasis added)
2.2.3. A libertarian conception of free will

In (2002 [1973]), Wiggins objects to compatibilist conceptions of free will, and specifically directs his criticism at the metaphysical contention upon which Hume based his own position on the matter, i.e. the contention according to which every event (or “state of affairs”) is either causally determined or random. Wiggins remarks that nothing in our current knowledge of the physical world confirms the validity of that speculative thesis. Denying that an event is random, then, does not imply that it is causally determined; conversely, denying that is causally determined does not imply that it is random. According to Wiggins, there is an «unclaimed ground between the deterministically caused and the random.» (2002 [1973], p.291)

Wiggins also rejects what he presents as a traditional strategy of libertarianism, which consists in declaring that this metaphysical no-man’s-land is the locus of metaphysically special abilities, such that they involve the suspension of causality between events.

Instead, he proposes that, given the unsatisfactory character of “Hume’s fork”, and pending a convincing argumentation on the part of the reductionists, we can simply, rationally admit that a) the area of ordinary language that bears upon agency and moral responsibility is irreducible to physicalist discourse; and b) the irreducibility of the language of agency and responsibility does not imply any “placement problem” (in other words, there is no reason to consider that the
placement, in the natural order of the physical world, of the subject matter of our ordinary discourse about moral agency and responsibility, poses a metaphysical problem).

In a way reminiscent of Davidson’s thesis of anomalous monism, Wiggins claims that the two areas of discourse (that about the physical, causal order of the world, and that about agency and responsibility), and their respective subject matters, can cohabit pacifically, without one risking engulfment by the other:

Why should there not be actions that are not even in principle uniquely or deterministically derivable from laws and antecedent conditions but can be fitted into practically meaningful sequences? We need not trace free actions back to volitions construed as little pushes aimed from outside the physical world. What we must find instead are patterns that are coherent and intelligible in the low level terms of practical deliberation, even if they are not amenable to the kind of generalisation or necessity that is the stuff of rigorous theory. On this conception the agent is conceived as an essentially and straightforwardly enmattered or embodied thing. His possible peculiarity as a natural thing among things in nature is only that his biography unfolds not only non-deterministically but also intelligibly; non-deterministically in that personality and character are never something complete, and need not be the deterministic origin of action; intelligibly in that each new action or episode constitutes a comprehensible phase in the unfolding of character, a further specification of what the man has by now become. (2002 [1973], p.292-293)
Ultimately, Wiggins wants to make the point that physicalistic descriptions «fail to characterize the world completely», and that human action is «causally underdetermined»:

[I]t may not matter if the world approximates to a world that satisfies the principles of neurophysiological determinism, provided that this fails in the last resort to characterise the world completely, and provided that there are actions which, for all that they are causally underdetermined, are answerable to practical reason, or are at least intelligible in that dimension. Surely these are not random. They are the mark left on the world by conscious agents who have freedom. (2002 [1973], p.294)

Arguably, in order to make this point, one needs to invoke the existence of the ground between causal determination and randomness, which the dismissal of Hume's fork leaves unclaimed.

Now, one might object that it is unconvincing of Wiggins to officially ban the libertarian strategy of declaring that this metaphysical no-man's-land is the locus of metaphysically special abilities, while relying on our incapacity to positively describe the geography of this metaphysical non-man's-land. Is Wiggins capitalizing on some current incapacity in order to indulge in crypto-mysterianism? Is he asserting an antinomy? Or does he mean to defend a credible, assertive metaphysical thesis?
Wiggins looks in the direction of Strawson’s view in ([1962]), according to which the participant perspective is (in my word) irremovable, and “the language of agency and responsibility” irreducible, ineliminable, and just as irremovable as the participant perspective, of whose conceptual articulation it is an essential piece — never mind metaphysics in general and determinism in particular.

Wiggins commends the claim to irreducibility: «What Strawson maintains is that, from the nature of the case, our ordinary ways of talking about responsibility, agency, human ability, human character and all the rest are best left unreplaced and unreduced. They require no justification at all beyond their manifest viability, and their proven capacity to animate the practices of everyday life.» (2002 [1973], p.299)

However, Wiggins expresses doubts about the claim to irremovability, and the supposed indifference to metaphysics of the participant attitude:

Throughout his article Strawson proceeds as if it were simply obvious […] that, if the use of the ordinary language of responsibility, agency, human ability, character, etc. requires no justification, then the untruth of sentences in the form “he could have helped it” (“he could then have done otherwise then”) will never represent any threat at all to the sorts of things we express in that sort of language. […] The language of action and responsibility is not something one needs reason to opt into. Certainly. We just use it. But it does not follow that nothing could count as a reason to opt out of it. (1987-VIII, p.299; emphasis added).

If Strawson means to vindicate our moral sentiments, of our ordinary sense of free, morally responsible agency, together with the standards to which
they are answerable and with “the ordinary language of responsibility (etc.)” which contributes to the conceptual articulation of the participant attitude, then his position is insufficient and unconvincing. If he means to simply dismiss the issue of vindication, his move is unsatisfactory. If we want to straightforwardly appraise the reasons that might count to opt out of the language of responsibility and action, we have to elaborate a deeper anthropology of moral life — as Wiggins has done. Presumably, Wiggins’ own vindication of the language of responsibility and action relies on the claim that some of our concerns, motives, etc. are unforsakeable.

2.3. METAETHICS AND CRITICISM

Wiggins defends a limited moral cognitivism, which I present in 2.3.1. The question as to whether valuational judgments are good candidates for truth is answered from within the Davidsonian theory. As Wiggins puts it, the practice of giving and asking for reasons (which, within his framework, partakes in the construction of a metaethical theory), is what allows us «to pass a verdict upon the epistemological or metaphysical pretensions of ethical discourse» (2006, p.323). More specifically, we must see whether it is plausible that moral judgments may exhibit all marks of truth, in the sense of the ordinary concept of truth elucidated in 2.1.3.
Wiggins argues that moral evaluative judgments (and valuational judgments in general) make good candidates for truth. By contrast, practical moral judgments do not make good candidates for truth. Nothing bars them in principle from being candidates for truth, and Wiggins mentions that there is nothing unsound about the idea of true practical judgments (although he barely supports this claim). But, Wiggins tells us, few practical moral judgments can exhibit the second and the fifth marks of truth.

Considering the restricted character of his cognitivism, Wiggins rejects the label of “realism” for his position, and characterizes his position as «the weakest recognizable form of cognitivism» (2002 [1987b], p.141).

Wiggins’ weak, restricted cognitivism is supported by a genealogical account of human experience and language, which reconstructs the history of the interrelations between response-dependent properties, shareable responses (or patterns of response) attuned to these properties, and concepts designating these properties. This genealogical account is shaped by Wiggins’ “equal and reciprocal partners” metaphysics of the experience of non-primary properties.

I present in 2.3.2. a survey of theories of valuational responses and judgments, all related to the reductionist approach, which are the main targets of Wiggins’ criticism.
2.3.1. Weak moral cognitivism and the genealogical account of human experience

Wiggins’ analysis of evaluative moral judgments is as follows: “t is good if and only if t is such as to make appropriate, or deserve, or merit a sentiment of approbation”. (2006, p.371). This is not an analytic definition of the concept “good”, but rather a statement of the truth conditions of the judgment “t is good” — the standard of correctness to which particular instances of “t is good” are answerable.

In Wiggins’ view, the same analysis is valid for judgments of affective properties, e.g. “t is disgusting”. It is moral evaluative judgments that are of special interest for the present dissertation, but since they are a subclass of the class of valuational judgments in general (as moral terms, or predicates, are a subclass of the class of valuational terms/predicates), I shall speak in what follows of valuational judgments.

According to Wiggins’ analysis, the standard of correctness to which valuational judgments are answerable is at once subject-involving and object-involving, as it includes the three following, tightly interrelated elements: a response-dependent property (goodness, borne by some item t, e.g. conduct, state of affairs, character, etc); a shareable response, or pattern of response, which is intrinsically related to that to which it is directed by an irreducible sense of appropriateness, desert or merit (viz. appropriate, deserved, merited
approbation); and a concept, “good”, designating the response-dependent property.

Wiggins’ position is cognitivist, as it claims that valuational judgments are candidates for truth, and that in many instances they actually fulfill their truth conditions. At the same time, it is subjectivist, as the properties designated by valuational concepts are not definable independently of the shareable patterns of response that they call upon. According to Wiggins, it is difficult to see how we could truthfully call good something that does not elicit a sense of appropriate approbation.

The genealogical account of human experience and language, which supports Wiggins’ analysis of valuational judgments explains how response-dependent properties, and patterns of response attuned to them, came to mutually interlock, and how this interlocking got expressed in the concepts.

This genealogical account thus brings together Wiggins’ metaphysics of experience and his elucidation of the ordinary concept of truth into five marks. That is, the reconstruction of the historical linkages between response-dependent properties and patterns of response attuned to these properties, and of the linkage of concepts to response-dependent properties, purports to establish both the view that valuational responses and valuational properties are equal and reciprocal partners, and the cognitivist thesis that valuational judgments may plausibly meet all marks of truth.
Although his historical reconstruction does not explicitly trace the natural history of Homo sapiens (as part of the general evolution of species), Wiggins does allude to the fact that the history of the interrelations between response-dependent properties, patterns of response and concepts has its roots in the natural, evolutionary history of the human species:

value terms have their sense by being annexed to properties in objects of our attention that call for certain shareable responses — the responses to these objects that the objects make appropriate. The fact that these responses are appropriate is owed to a scale of values whose recognition is of course downwind of the fabric and constitution of conscious human subjects. (2006, p.372; second emphasis added)

It is worth pointing out that from the point of view of evolutionary theory, primitive responses (manifesting primitive dispositions) were answerable to a primitive standard of correctness: pragmatic success, increasing chances of reproductive success (animals who “got things right” survived and enjoyed better reproductive success than animals who did not). In this way, certain properties in the environment of organisms got to be matched, by the process of natural selection, with the “right” responses” on the part of organisms. I take it that nothing bars Wiggins, who (as we have seen) acknowledges the indispensable contribution of evolutionary theory to the understanding of morality, to endorse this evolutionary preamble: as a rule, for every species, primitive <property, response> pairs were borne out of the interaction of its members with its evolutionary environment; and each species inherited many such pairs from its
evolutionary ancestors. So it is with the human species, thus endowed with an evolutionary heritage of its own.

With the development of language, valuational concepts (including specifically moral ones) are used to designate properties that call upon certain patterns of response. Valuational judgments come to accompany the responses in question.43

In Wiggins’ historical reconstruction, the development of language is bound up with a collaborative process of mutual instruction and improvement in the use of concepts, which causes a refinement of the <property, response> pairs. Capturing properties with language tends to elicit finer perception, and finer patterns of response.

[A]fter the process has begun, those who participate in it may report not only that they discriminate more keenly, make more decisions and are better satisfied with the classifications and subclassifications they now effect, but also that they get more and more cognitive-cum-affective satisfaction in their own responses. Finer perceptions can both intensify and refine responses. Intenser responses can further heighten and refine perceptions. And more and more refined responses can lead to the further and finer and more variegated or more intense responses and perceptions. (2002 [1987c], p.196)

43 This is the territory that Strawson explored in “Freedom and Resentment”
Wiggins suggests moreover that, even at this linguistic stage of what we might now call *human history* (i.e. the history of language-using human beings, by contrast with the natural (pre)history of Homo sapiens, emerging from the evolution of species), a process similar to natural selection is involved in the linkage of <property, response> pairs:

When this point is reached, a system of anthropocentric properties and human responses has surely taken on a life of its own. Civilization has begun. One may surmise that at any stage in the process some <property, response> pairs will and some will not prove susceptible of refinement, amplification, and extension. One may imagine that some candidate pairs do and some do not relate in a reinforceable, satisfying way to the subjectivity of human life at a given time. Some pairs are such that refinement of response leads to refinement of perception and vice versa. Others are not. Some are and some are not capable of serving in the process of interpersonal education, instruction and mutual enlightenment. Those pairs that do have this sort of advantage, we may expect to catch on and survive, and then evolve further, generate further <property, response> pairs and make room for the discovery of yet further properties that lie at a progressively greater distance from specific kinds of affect. Those pairs that do not have this sort of viability will no doubt fall by the wayside. (2002 [1987c], p.196-197)

Examining more precisely the linkage of <property, response> pairs, Wiggins suggests that it follows a two-step process, at the end of which speakers acknowledge a) that, say, *t* is good if and only if *t* is such as to make appropriate, or deserve, or merit the sentiment of approbation; and b) that any particular such
judgment is essentially contestable. The process “fixes” the meaning of valuational terms (in the sense of stabilizing it, while allowing for possible revisions), and conceptually captures the <property, response> pairs by both ends (so to speak). Let us examine the two steps in turn.

The first step consists in acknowledging that the standard of correctness to which a given pattern of response, together with the judgments that accompany it, is answerable, implies as a requirement the presence of the property to which the patterns of response in question is attuned. Conditions for the presence of a property are identified, and get included in the standard of correctness to which responses are answerable. At this point, language users acknowledge that valuational terms do not refer to mere mental states considered independently of what they are directed to, but rather that they designate properties which, as such, call for certain patterns of responses:

Suppose that a point has been reached where a <property, response> pair is well established, the response is corrigible by reference to the question whether whatever is required for the presence of the property is present, and various supplementary considerations have become available that make possible the criticism, explanation, and vindication of attitudes and responses to a given thing. Suppose therefore that we are past the stage at which the critics of classical subjectivism like to see the position as stuck, where the non-accidental occurrence of some simple response is seen as simply sufficient for the presence of the property. (2002 [1987c], p.197)

The second step consists in identifying the conditions under which the manifestation of the pattern of response is appropriate. This has the effect of making evaluative judgments essentially contestable.
Instead of fixing on an object or class of objects and arguing about what response or responses they are such as to evoke, we can fix on a response [...] and then argue about what the marks are of the property that the response itself is made for. And without serious detriment to the univocity of the predicate, it can now become essentially contestable what a thing has to be like for there to be any reason to accord that particular appellation to it, and correspondingly contestable what the extension is of the predicate. (2002 [1987c], p.198; emphasis added)

The essential contestability of evaluative judgments thus arrived at means that the cognitive status of particular evaluative judgments (and the propriety of particular responses) is assessable in the practice of rational criticism:

If a property and an attitude are made for one another, it will be strange for one to use the term of the property if he is in no way party to the attitude and there is simply no chance of his finding that the item in question has the property. But if he is no stranger to the attitude and the attitude is favourable, it will be the most natural thing in the world if he regards it as a matter of keen argument what it takes for a thing to count as having the property that the attitude is paired with. (2002 [1987c], p.199)

The essential contestability of evaluative judgments presupposes that speakers have come to agree upon two closely related things: the standards of correctness to which valuational responses, together with the judgments that accompany them, are answerable; and the meaning of evaluative terms. Both are revisable, but Wiggins suggests that the two-step process just sketched stabilizes them sufficiently for speakers (and participants in the practice of giving and
asking for reasons) to genuinely assess the appropriateness of responses and
the cognitive status of evaluative judgments, by reference to agreed upon
standards of reference, without confusing such cognitive assessment with
disagreements about the standards of correctness or about the meaning of
words. Conversely, contestation of standards of correctness and of the meaning
of terms is not confused with the assessment of the appropriateness of particular
responses and of the cognitive status of particular judgments.

We can now take stock. First, a reminder is in order regarding the
irreducibly of valuational terms. Then, we will see how the reconstruction
supports the claim that valuational judgments may meet all marks of truth,
including notably the second and the third, and therefore be true in a sense that
is both subject-involving and world-involving.

According to the elucidative construal of analysis to which Wiggins
adheres, the analysis of evaluative terms cannot fail to lead us to other evaluative
terms (the meaning of terms is elucidated by their place in conceptual networks).
The irreducibility of the evaluative vocabulary is linked to the irreducible sense of
appropriateness in human valuational responses, which is included in the
standards of correctness to which these responses, together with the judgments
that accompany them, are answerable.

Wiggins anticipates the criticism that this would elicit on the part of
reductionist philosophers: such elucidative analysis is viciously circular, and can
only deliver explanations of the *virtus dormitiva* sort. But, as Strawson explained in his *Analysis and Metaphysics*, circularity is not necessarily a problem for an approach based on an elucidative conception of analysis. This point is reiterated by Wiggins, who points out that the idea of “reducing” evaluations to a “sentiment of approbation” devoid of any sense of appropriateness, desert, or merit hardly makes sense:

> [O]n a proper understanding of the point of subjectivism and its having no need to supplant valuational by non-valuational language, the circularity is benign […] In any case no *new* circularity has been introduced. There was already a kind of circle […] What after all is a sentiment of approbation? […] Surely a sentiment of approbation cannot be identified except by its association with the thought or feeling that *x* is good (or right or beautiful) […] and with the various considerations in which that thought can be grounded, given some particular item and context, *in situ.* (2002 [1987c], p.188)

In other words, a genuinely valuational vocabulary is necessary to conceptually articulate the standards of correctness to which valuational responses and judgments are answerable, and to actually assess particular valuations.

According to Wiggins, his reconstruction shows that valuational judgments meet the second mark of truth, i.e. that they may elicit epistemic convergence among speakers, where the convergence is best accounted for by a vindicatory
explanation. As an illustration, Wiggins takes a bunch of <affective property, affective response> pairs:

\[\text{genuinely } [\text{funny/ appalling/ shocking/ consoling/ reassuring/ disgusting/ pleasant/ delightful/...}] \text{ things are things that not only } [\text{amuse/ appal/ shock /console /reassure /disgust/ delight/...}] \text{ but have these effects precisely because they are } [\text{funny/ appalling/ shocking/ consoling/ reassuring/ disgusting/ pleasant/ delightful/...}] \] (2002 [1987c], p.199-200; brackets original)

He then analyzes the “because” (in the preceding quote) as introducing a vindicatory explanation:

[T]his “because” introduces an explanation that both explains and justifies. (In something like the way in which “there is a marked tendency for us all to think that 5+7=12, and this tendency exists because there is really nothing else to think about what 7+5 is” explains a tendency by justifying it. On request, the justificatory aspect can be made yet more evident by filling out the explanation with a calculation, or a proof. […] A similar complement for the valutational case could consist in an argued vindication of the claim that x is indeed funny, shocking….) (ibid)

In other words, the way properties and responses link up (e.g. the response-dependent property of amusingness and the experience of amusement), and the way concepts and properties link up (the concept of “amusing” and amusingness), imply that in the case where the judgment that t is
amusing elicits convergence, this will be best explained by the fact there is nothing else to think but that it is amusing.

The same reconstruction shows, Wiggins suggests, that valuational judgments meet the third mark of truth, according to which a judgment cannot be true exclusively in virtue of subjective matters (as for example the mental state that the judgment expresses, or a desire that the judgment be true, or a mental act of will, or an act of assertion); consequently, true judgments have the potential to elicit epistemic convergence of the right sort, but may in principle fail to actually elicit it. That is, the reconstruction shows that the linkage between properties and responses, and then between concepts and properties, is not a subjective matter; and that true judgments are world-involving.

Finally, one gathers that for Wiggins, nothing stands in the way of the fifth mark of truth being met, following which «every plain truth is compatible with every other plain truth» (2002 [1976], p.115)

In sum, valuational judgments may exhibit all marks of truth. Thus, valuational judgments may be true in the ordinary sense of the concept, which is subject-involving and world-involving.

As for practical judgments, although it is not impossible that they may satisfy the second mark of truth, such occurrences are bound to remain exceptional, according to Wiggins: «Pure specific evaluations […] are closely
associated with special sensibilities for which there can be something
approaching an agreed test. But many moral judgments appear to lack this
feature, and especially perhaps practical judgments lack it.» (2002 [1987b],
p.163; emphasis added)44

What Wiggins particularly doubts is not that practical judgments may elicit
epistemic convergence; it is that they may elicit convergence of the right sort,
best accounted for by vindicatory explanation, where “there is nothing else to
think but [the propositional content of the judgment]”.

Wiggins suggests that the fact that practical judgments may rarely elicit the
right sort of epistemic convergence is accounted for by a deeper fact, which
implies that it is also difficult for practical judgments to satisfy the fifth mark of
truth (which is the claim that all true beliefs about the world are reconcilable):

As we are reminded by the fifth mark of truth, everything true must be
consistent with everything else that is true. The pieces fit together
because they are made for one another. [...] every fact about what
stands in front of us must be reconcilable with [...] every other fact that
stands in front of us. In contrast, surely the relation of considerations
for and against a course of action is unlike the relation between
evidence and judgment. We must not expect the discovery of a good
course of action to explain away the claim of all considerations that
support other courses of action. [...] If there is no one answer to a

44 Wiggins uses the expression “strictly (or “purely”) valuational predicates” to designate what are
otherwise known as “thick” moral predicates, which contrast with “thin” moral predicates. A “pure
specific evaluation” is an evaluative judgment in which a purely valuational predicate occurs.
practical question whose authoritative revelation can dissipate the force of all other answers, then it is hard to deny that there will be room for different kinds of answers, each answering to a different personal choice of the kind of life to lead, or each determining a distinct valuational emphasis. (2002 [1987b], p.175-176)

Again, this does not preclude altogether the possibility that some practical judgments may satisfy the second mark of truth (without generating the sort of irreconciliability that would bar them from satisfying the fifth as well): «however strong the convergence requirement is, it may be that there are certain practical judgments, especially judgments of the morally prohibitive kind, that pass it easily.» (2002 [1987b], p.176-177) But these would be the exception rather than the rule.

2.3.2. Criticism of rival views

Wiggins directs most of his criticism at three main targets: subjectivism of the Humean sort (of which he distinguishes three varieties), utilitarianism, and noncognitivism, in a broad sense that includes John Mackie’s error theory, as well as distinct views fostered by various philosophers, including notably Gilbert Harman.

All these theories partake in what Wiggins presents as a common, wrongheaded project viz. the reductionist attempt to account for the valuational
domain (valuational responses and the judgments that accompany them, values in general) in non-valuational terms. This reductionist goal favors a distinct but related mistake: the severance of valuational responses (and the judgments that accompany them) from what they are directed to. The former are construed as psychological states devoid of an inherent, irreducible normative ingredient, viz. a dimension of appropriateness. And what those states are directed to is construed as featuring non-moral properties. In this way, the idea that there could be such things as genuine moral (or affective) properties is ruled out on the count that such items would be, in John Mackie’s famous expression, “ontologically queer”.

As for moral evaluative judgments, the theories in question propose distinct analyses of them, but they all concur on this negative point: such judgments cannot be true in the world-involving way defined by Wiggins.

Wiggins’s discussion of Humean subjectivism is presented in 2.3.2.1. His critique of utilitarianism is presented in 2.3.2.2., while his critique of noncognitivism is presented in 2.3.2.3.

2.3.2.1. The reductionist strand in Hume’s analysis of valuational responses and judgments

The reductionist strand of Hume’s thought exhibits the features of reductionism presented above. Thus, Hume believes that a scientifically credible
account of human thought and action requires that valuational responses and judgments be accounted for in non-valuational terms. This reductionist requirement leads Hume to analyze valuational responses, as mental states, independently of what they are directed to. Finally, the idea that these responses could be directed to genuine valuational properties is ruled out, to the profit of a hypothesis according to which responsive mental states are accompanied by a sort of projective illusion, which induces the mistaken idea that we are responding to genuine moral properties.

How does this reductionist strand of Hume’s thought square with his own vindicatory strand? To answer that question, we have to reconstruct Wiggins’ chronological reading of Hume. Again, I am not concerned with Humean exegesis. But the reductionist strand of Hume’s thought has served as a blueprint for the sort of doctrine that is the main target of Hume’s criticism, so a brief examination of (Wiggins’ reading of) the former will help understand (Wiggins’ conception of) the latter.

In the *Treatise* (1739-1740), some elements of Hume’s thought suggest (as we have seen in 2.2.1.1.) that morality reaches beyond the sentiment of benevolence which is at its origin. However, those elements remain undeveloped, and are outdone by the reductionist strand, which weaves a “classical” form of subjectivism that can be roughly summarized as follows.
Moral sentiments come down to sentiments of approbation and disapprobation (which are themselves distinctive forms of pleasure and pain), resulting from the interaction of sensory impressions with our natural dispositions and capacities. By reference to the contrast that Hume makes between “reason” and “passions” (a contrast importantly qualified by the existence of “calm passions”), moral sentiments belong to the category of “passions”.

Systematic (or quasi systematic) correlations are observable between, on the one hand, certain features of items (like characters, actions and consequences of action); and on the other hand, moral sentiments. It seems, that is, that certain kinds of things systematically arouse sentiments of approbation, others of disapprobation. However, Hume suggests that valuational responses (e.g. moral sentiments) come down to projections of our own feelings onto certain items (they are not discriminatory responses to genuine moral properties) and that moral judgments simply express moral sentiments. This implies that neither moral sentiments, as valuational responses, nor the judgments that accompany them, are answerable to public, intersubjective standards of correctness. Rather, they are incorrigible: «There is just so much vice or virtue in any character as every one places in it, and ’tis impossible in this particular we can ever be mistaken. [Footnoted:] This proposition must hold strictly true, with regard to
every quality, that is determin’d merely by sentiment.» (Treatise, 3.2.8., p.350; emphasis added) 45

The word “merely” in the bit just quoted alerts us to Hume’s aspiration to account for valuational responses and judgments in non-valuational terms.

Thus, as regards valuational responses, he accounts for them in terms of specific sorts of pains or pleasures, considered in independence of what they are directed to — objects taken to feature non-valuational properties exclusively. If we have the impression that those responses are directed to something that does feature valuational properties, it is because we have a propensity to project our responsive attitudes and feelings upon objects, and to mistakenly attribute to them valuational properties that are in fact non-existent. As for valuational judgments, if they cannot fail to be “true”, then that is not in the world-involving sense of “true”.

According to Wiggins, the subjectivism defended in the Treatise comes quite close to what he calls “vulgar subjectivism”, according to which moral evaluative judgments are to be analyzed into one or the other of the two following options: «(a) “x is good” as said by me, means no more than “I like it” or (at best)

45 In other words, Hume extends his analysis of moral responses and judgments to valuational responses and judgments in general, on the basis that they are “determin’d merely by sentiment”. Wiggins quotes this line, without the footnote, in (Wiggins 2006, p.371)
"we like it".» (2006, p.372) For Wiggins, vulgar subjectivism is what «has given subjectivism a bad name»:

This is all the associations that are commonly imputed to subjectivism with the thought that evaluative judgments may be assimilated to "mere responses", and with the idea that the responses the subjectivist interests himself in are autonomously inner states — or states such that, if you are in one of them, you can tell and tell infallibly, without looking outward, whether you are in it or not. (These would be states that “have no reference to anything beyond themselves”.) (2002 [1987c], p.208)

In the *Enquiry* (1751), as we have seen in 2.2.1.1., Hume asserts the *intersubjective* character of the standards of correctness to which valuational responses and judgments are answerable. Wiggins suggests that this weakens the subjectivist conclusion of the reductionist strand.

In his essay “Of the Standard of Taste” (1957), Hume asserts more fully the doctrinal shift initiated in the *Enquiry*, although he maintains his reductionist aspiration to account for valuational responses and judgments in presumably non-valuational terms.

---

46 This is equivalent to the definition of subjectivism given by Ayer in *Language, Truth and Logic* (Ayer 1946 [1936], p.104).
First, Hume explicitly rejects the “vulgar” sort of subjectivism which, as Wiggins points out, he came close to propound in the *Treatise* (in the following passage, the lines between quotes are citations from Hume’s essay):

Hume contends [in “Of the Standard of Taste”] that for a subjectivist simply to declare that “all sentiment is right because sentiment has reference to nothing beyond itself” — which might seem a little close to the position he had himself taken up in the *Treatise* — is for him to invite charges of advancing an “extravagant paradox, or rather a palpable absurdity.” (2002 [1987c], p.191)

Second, Hume proposes to elaborate the point that valuational standards are intersubjective by calling upon an analogy with standards of sensory taste. Sensory responses (gustatory, auditory, etc.) are answerable to standards of taste determined by well-functioning sensory organs. On this count, we may assume that most people are endowed with similarly sound sensory organs. From this assumption we may infer that humans share *common* standards of taste. In the case where some persons would enjoy sensory organs that would be significantly more sensitive and sounder than those of most people, those persons would be distinguished as “good judges”, and their sensory responses and judgments would enjoy superior authority. In such a case, the sensory taste of good judges would determine standards of taste. These standards would still be objects of discussion, however, and as such they would be intersubjective. Thanks to the good judges, it would *not* be the case that each person’s sentiment
would have no reference beyond itself, so that each individual would be trapped within his/her own subjectivity.

This is what happens with valuational responses — including moral sentiments — and the judgments that accompany them. They are answerable to standards of correctness determined by a sound sense of what is to be approved or disapproved (we will see shortly what goes into this “sound sense”), but not all people are on an equal footing in this connection, as some persons enjoy sounder discriminatory sense in valuational matters. Those persons are good judges, and their judgment determines the standards of correctness to which everybody’s valuational responses and judgments (including the specifically moral) are answerable. These standards are intersubjective, and therefore reach beyond the sentiments that give rise to them.

Here, Hume seems to suggest that discriminatory sense in valuations is a matter of “judgment”, in a sense which acknowledges the mutual involvement of affections and understanding, and therefore qualifies the reason/passion opposition. He also suggests that the standards established by the judgment of “good judges” stimulate the practice of giving and asking for reasons, within which particular valuational responses and judgments are assessed (again, in the following passage, the lines and expressions between quotes are citations from Hume):

[T]o the extent that there is a standard of correctness in morals, this is determined by the verdict of whoever judges “most coolly”, “with the least prejudice”, and on the basis of the fullest information — all of which, “if we consider the matter aright”, is “a question of fact, not of
sentiment”. When men dispute and argue in valuation, and when they succeed in instructing one another, what they are really seeking to do is to approximate the verdicts of that [good] judge. (2002 [1987c], p.190)

For Wiggins, “Of the Standard of Taste” is Hume’s ultimate step in the right direction, although it remains unsatisfactory on the following counts.

First, Hume sticks to his reductionist aspiration, and as a corollary fails to acknowledge that a valuational vocabulary is necessary to conceptually articulate the standards of correctness to which valuational responses and judgments are answerable (and to actually assess particular valuations). In other words, he fails to acknowledge that, as mentioned in the previous subsection (2.3.1.), «a sentiment of approbation cannot be identified except by its association with the thought or feeling that x is good (or right or beautiful) and with the various considerations in which that thought can be grounded, given some particular item and context, in situ.» (2002 [1987c], p.188) The notion of the “good judge” cannot be defined, and good judges cannot be recognized, without a recourse to valuational language.

Secondly, and relatedly, Hume fails to acknowledge that moral sentiments cannot be adequately analyzed in independence of what they are directed to — that, as Wiggins puts it, such mental states, and that to which they are directed, are equal and reciprocal partners.
Clearly, the reductionist strand of Hume’s thought involves a considerable tension with the anthropocentric thesis and the vindicatory strand of his thought. Wiggins, of course, advocates the rejection of the reductionist aspiration: «[L]et us now abandon Hume’s aspiration to secure the standard of correctness in valuation from outside the domain of values, or by sole reference to the qualified judge. And let us restore to its proper place the ordinary idea in its ordinary construal that the criterion for a good judge is that he is apt to get things right.» (1987c, p.194; emphasis added)

2.3.2.2. Utilitarianism

Utilitarianism proposes a certain theory of value which identifies “the good” (that which, having intrinsic, non-derivative worth, is desirable for its own sake) with pleasure or, as most recent versions have it, mental states of “gratification”, or “satisfaction” — the satisfaction, that is, of other mental states viz. desires. Utilitarianism then defines right action (right decision, right conduct) as that which, as a consequence, maximizes the good overall (hence it is a variety of consequentialism).47 Utilitarianism contends that, even though satisfaction is an

47 As Wiggins puts it, «the rightness of acts (their obligatoriness or permissibility) is explained in terms of the good, which turns out to be the goodness exclusively of outcomes as more or less pre-ethically described.» (2006, p.148)
eminently subjective state, it is in principle possible to measure it (at the scale of individuals, as well as at the scale of masses of individuals) and to calculate its maximization. The maximization of the good is the standard to which both evaluative moral judgments and practical moral judgments are answerable. Utilitarianism is not concerned with affective properties, but aesthetic values and aesthetic judgments fall within its scope.

Utilitarianism lends itself to reductionism, to the extent that states of satisfaction are definable in non-valuational terms. Thus, Wiggins finds in utilitarianism the same basic problems that he finds in vulgar subjectivism, as well as in the forms of subjectivism defended by Hume.

First, it fails to acknowledge that the valuational vocabulary is necessary to conceptually articulate the standards of correctness to which moral judgments are answerable.

Second, against the view that moral sentiments — or, more broadly, the mental states associated with value experience — and that to which they are directed are equal and reciprocal partners, it assumes that states of satisfaction can be considered in independence of what they are directed to. In the following quote, Wiggins alludes to the fact that it is on this dubious basis that utilitarianism unconvincingly dismisses the “higher form” of nihilism (that which follows from the nonexistence of anything intrinsically worthwhile, apt to constitute a non-
derivative goal to human action and existence and thus confer upon them an
ultimate sense of point), evoked by the Myth of Sisyphus:

Finally, nor has nineteenth- or twentieth-century Utilitarianism much to
fear from this manner of fable-telling [i.e. the myth of Sisyphus —
P.C.]. For the *locus* or origin of all value has been firmly confined [by
utilitarianism] within the familiar area of psychological states conceived
in independence of what they are directed to.» (2002 [1976] p.97)

Wiggins remarks that investing mental states of gratification with an
intrinsic value, in independence of what they are directed to, gives rise to what
appears, from the point of view of the ordinary moral outlook, as «an
incoherence» (2002 [1976], p.99), or «a curious instability» (2002 [1976], p.104),
which he describes as follows.

In common life, we experience states of gratification *that would simply not
be* if we severed them from what is at their origin; and we consider that what
value these states of gratification have derives from what we take to be the value
of what causes them (which might be an intrinsic value, or which might depend
upon the value of some further finality). As Wiggins puts it, «the value of the state
depends on the value attributed to the object.» (2002 [1976], p.105 n.19) Now,
utilitarianism claims that, as it stands, our ordinary moral outlook is upside down.
If we see clearly, we will acknowledge that states of gratification have intrinsic
value, and that the value of what causes them derives from them.

The result is the “incoherence”, the “curious instability” that Wiggins points
to — a clash of outlooks, which he evokes in a way that clearly indicates where
he stands (note the suggestion, in the quote that follows and in the next, that the experience of aesthetic values is bound up with that of moral values):

[F]rom the inside of lived experience, and by the scale of value that that imposes, the shape of an archway or the sound of the lapping of the sea against the shore at some place at some time may appear to be of an altogether different order of importance from the satisfaction that some human being once had from his breakfast. (2002 [1976], p.105)

For Wiggins, our ordinary outlook (or, as he calls it, the “inner view”, or “inner perspective”) is sound, and utilitarianism is an unsatisfactory doctrine, sustained by philosophers for whom accounting for the valuational in non-valuational terms is an adequacy condition on any theory of value. But the theories that satisfy such an inadequate adequacy condition can only be impoverished to the point of absurdity. That is, they condition a worldview following which human experience appears absurd:

For no appetitive or aesthetic or contemplative state can see its own object as having a value that is derivative in the special way that is required by the thesis that all non-instrumental value resides in human states of satisfaction. […] [I]n this conception of such states we are entitled to complain that nothing remains that we can recognize, or that the inner perspective will not instantly disown. [Footnote 19:] […] A man comes at dead of night to a hotel in a place where he has never been before. In the morning he stumbles out from his darkened room and, following the scent of coffee out of doors, he finds a sunlit terrace looking out across a valley on to a range of blue mountains in
the half-distance. The sight of them — a veritable vale of Tempe — entrances him. In marvelling at the valley and mountains he thinks only how overwhelmingly beautiful they are. The value of the state depends on the value attributed to the object. But the theory I oppose says all non-instrumental value resides here in the man’s own state, and in the like states of others who are actually so affected by the mountains. The more numerous such states are, the greater, presumably, the theory holds, is the “realized” value of the mountains. (2002 [1976], p.105; emphases added)

Following utilitarian analysis, evaluative moral judgments can be true. However, they cannot be true in the world-involving sense defined by Wiggins (by his third mark of truth in particular). Utilitarianism may claim that “x is good” is true, but that would be in virtue of its definition of that which has intrinsic worth, that is, states of satisfaction considered in independence of what they are directed to. Utilitarianism may claim that “x is right” is factually true, i.e. that it is true in virtue of a fact, namely that x maximizes states of satisfaction. But then again, the fact in question is not world-involving: it brings us back to states of satisfaction, considered in independence from what they are directed to.

2.3.2.3. Noncognitivism

The expression “noncognitivism”, in its narrow sense, designates those theories according to which valuational and deontological judgments are incapable of being true or false, for “logical” reasons. In that sense,
noncognitivism mostly designates the “emotivist” (or “expressivist”) sort of theory
defended by logical positivists, and the prescriptivist theory defended by, inter
alia, Richard Hare.48

48 In early twentieth century, logical positivism is the first school of thought to be based, in a self-
conscious and programmatic way, on reductionist analysis, itself based on the idea of a criterion
of cognitive meaningfulness. In this doctrinal context, the old, historical project of accounting for
the valuational and the deontological in non-valuational, non-ethical terms is annexed to the
reductionist program, and logical positivists discuss whether valuational judgments and
deontological judgments can be reduced to judgments about experience or about physical things.

The vast majority of logical positivists agree that the only way that so-called moral
judgments could satisfy the criterion of cognitive meaningfulness is to interpret them as empirical,
sociological statements about that which is found desirable by certain members of a community.
But moral evaluative judgments cannot state anything about values (about what is desirable for its
own sake), for the reason that such talk cannot satisfy any credible criterion of cognitive
meaningfulness (in short, such statements are incapable of being confirmed or disconfirmed on
the basis of sense experience). Moral judgments, both evaluative and practical, are understood
as vocalizations exhorting others do certain things and not others; but they cannot have any
cognitively meaningful content.

Schlick thinks that it is possible to revise utilitarianism in accordance with the tenets of
logical positivism. This revision implies the rejection of the only bit of normativity that remains in
classical utilitarian theories: that is, the idea that psychological states of satisfaction are desirable
for their own sake. «The question whether something is desirable for its own sake», Schlick
writes, «is no question at all, but mere empty words.» (Schlick 1959 [1930], p.257) Thus, the
definition of the good as what maximizes that which is desirable for its own sake (incidentally
states of satisfaction) has to go. However, if we can analyze “that which is desirable for its own
sake” as “that which is actually desired by certain members of a community”, or as “that which is
Wiggins, however, uses the expression “noncognitivist” in a broad sense, including all theories according to which, as it turns out, evaluative moral judgments cannot be true in a world-involving way. The main reason for such views is ontological rather than logical: we have no good reasons to consider that moral properties exist, so moral predicates are empty concepts. This notably found desirable by certain members of a community”, then utilitarian theories can be analyzed as empirical theories about what actions are found to be right in certain communities.

This proposed analysis of utilitarian theories, as well as subjectivist theories, is scrutinized by Ayer in *Language, Truth and Logic*. According to Ayer, although «it is possible to invent a language in which all ethical symbols are definable in non-ethical terms», he claims that neither subjectivist theories, nor the revised, frankly reductionist variant of utilitarianism are acceptable as analyses of actual ethical statements:

> [W]hat we are denying is that the suggested reduction of ethical to non-ethical statements is consistent with the conventions of our actual language. That is, we reject utilitarianism and subjectivism, not as proposals to replace our existing ethical notions by new ones, but as analyses of our existing ethical notions. Our contention is simply that, in our language, sentences which contain normative ethical symbols are not equivalent to sentences which express psychological propositions, or indeed empirical propositions of any kind. (Ayer, 1946 [1936], p.105)

According to Ayer, the fundamental ethical concepts are unanalyzable for the reason that they are mere pseudo-concepts. It follows that ethical sentences do not meet the criterion of cognitive meaningfulness that logical positivists favor (i.e. testability), and consequently do not express genuine propositions. Rather, they are non-propositional expressions of emotions.
includes the “error theory” propounded by John Mackie, as well as the view of Gilbert Harman.\footnote{49 “Error theory” is not linked to the logical positivist doctrine of cognitive meaningfulness, but it is linked to physicalistic reality criteria. It recognizes that moral evaluative judgments are “logically” capable of being true or false, but (given the reality criteria recognized by the theory) it concludes that they are systematically false, for they are about things that do not exist. According to error theory, value terms (occurring in value judgments) are empty terms, i.e. they fail to refer to things that satisfy what it takes as credible reality criteria.}

Wiggins criticizes the analysis of evaluative moral judgments that Gilbert Harman proposes in his book \textit{The Nature of Morality} (Harman 1977). By reference to Wiggins’ second mark of truth, we can say, very curtly, that Harman recognizes that evaluative moral judgments may satisfy its first condition (epistemic convergence), but denies that they satisfy the second (vindicatory explanation of the convergence).

It will be remembered from 2.1.3. that the fulfillment of the second mark of truth does not necessarily require that a group of persons actually converge upon a particular judgment. We may consider the case of that judgment being endorsed, and so \textit{believed}, by a single person; and we may ask what best explains that case of endorsement (of \textit{belief acquisition}, or \textit{belief fixation}), as we would ask what would best explain the convergence of many persons upon the same judgment. If the best explanation of the acquisition by one person of the belief that, say, “\(x\) is wrong” is that there is nothing else to think but that \(t\) is
wrong, then we may consider that many persons would converge upon the same judgment, and so that it exhibits the second mark of truth.

Harman centers his discussion of evaluative moral judgments on the case of a person (e.g. the reader) endorsing the judgment “that is wrong”. By reference to Wiggins’ analysis of truth, Harman’s position comes down to denying that evaluative moral judgments may satisfy the second condition of the second mark of truth, i.e. an evaluative moral judgment like “t is wrong” might elicit epistemic convergence, but that would not be best explained by the fact that there is nothing else to think but that “t is wrong”.

As a set up for his argumentation, Harman conjures up a case where the judgment “this is wrong” noninferentially occurs in a single person:

If you round a corner and see a group of young hoodlums pour gasoline on a cat and ignite it, you do not need to conclude that what they are doing is wrong: you do not need to figure anything out: you can see that it is wrong. But is your reaction due to the actual wrongness of what you see or is it simply a reflection of your moral “sense”, a “sense” that you have acquired perhaps as a result of your moral upbringing? (Harman 1977; quoted in Wiggins 2002 [1987b], p.156)

What, Harman asks, best explains the occurrence of the observer’s judgment, that the action she sees is wrong? The observation of real moral properties, or the observation of non-moral properties (non-moral salient features of the situation) interacting with certain non-moral dispositions, all describable in non-evaluative terms?
Wiggins summarizes Harman’s noncognitivist position as follows, merging Harman’s original question of what best explains the occurrence of the judgment into a consideration of what best explains the virtual convergence of more than one (virtual) agents upon it:

Harman will say this explanatory claim is wide open to the objection that the best explanation of our seeing the episode in that way can dispense altogether with the claim that the act was wrong, or callous and gratuitously cruel. For Harman says that our reaction to our perception is to be explained by our “psychological set”. The act had some non-evaluative property φ, say, and φ is a property that observers like you and me have been schooled to abhor. So the putative moral qualities of the treatment of the cat are simply irrelevant to why we saw the action as we did. Our seeing it in that way bears only upon the state and character of your and my moral sensibility. (2002 [1987b], p.156-157)⁵⁰

Wiggins disputes Harman’s contention that the explanation of the occurrence of the judgment (or of epistemic convergence upon it) can be put in

⁵⁰ Harman remarks that, by contrast, the observation of physical items and phenomena too small to see with the naked eye does not pose any problem. Regarding the case of a scientist who, observing a vapour trail in a cloud chamber, thinks “there goes a proton”, Harman agrees with pretty much everybody, including Wiggins, that, as the latter puts it, «[t]he scientist counts as observing a proton because the best explanation of his visual state is that there is a vapour trail, and the best theoretical explanation of the vapour trail itself is a proton.» (2002 [1987b], p.156)
non-evaluative terms, and can dispense altogether with the very content of the judgment. Three arguments are presented against Harman’s contention.

As a preliminary, in order to simplify and clarify the case, Wiggins proposes to substitute the judgment “that is callously cruel” for “that is wrong”. The reason is that “wrong” is not a “pure, relatively specific evaluative predicate”. For one thing, the judgment “that is wrong” is included in the intersection of the class of evaluative judgments with the class of practical judgments; moreover, the predicate “wrong” may occur in a broad range of contexts (compare “a wrong action”, “the wrong way”, “a wrong number”, “the wrong person”, “a wrong note”, “the wrong explanation”, etc). By contrast, “cruel” is indeed a “pure, relatively specific evaluative predicate”, which makes of “x is cruel”, or of “that was callously cruel”, clear, simple examples of moral evaluative judgments.

As a first argument, Wiggins remarks that invoking schooling does not settle the matter, but only pushes the explanation further back, as the way one is schooled requires to be explained in turn. Assuming convergence about the way “we” were schooled, Wiggins suggests that the best explanation of schooling could be a vindicatory explanation: “we” were schooled to respond in certain ways to certain situations, and certain judgments (like “that was cruel”) accompany our responses, because “there is nothing else to think” in such situations:

Harman’s reference to how we were schooled raises the question why we were so schooled, and whether the mores and institutions that make up the context of the schooling are not themselves a response
to something that is simply there. (There to be found by anyone, or by anyone who is sufficiently attuned to what bears upon the matter.) Compare the answer to the question why we are schooled that 7+5=12 (1987b, p.157)

Secondly, Harman’s noncognitivism implies that no particular occurrence of (or convergence upon) the judgment “that is callously cruel” could be explained by there being nothing else to think but the fact that that is callously cruel. This demands a psychological theory that best explains that all occurrences, in response to distinct states of affairs, of the judgment “that is callously cruel” must be given the same generic explanation (a psychological mind set, schooled in a certain way, responding to non-valuational properties of a certain states of affairs). But, Wiggins claims, the construction of such a psychological theory begins with radical interpretation: the attribution of beliefs and concerns to interpretees, in order to make their overall behavior (including verbal behavior) intelligible. In this context, Wiggins identifies the sought-for psychological theory with the “descriptive anthropology” which (in the Tarskian semantics that Wiggins adopts) the radical interpreter uses in conjunction with a theory of truth-in-L and with the Principle of Humanity; and he points out that the radical interpreter (and eventual psychological theorist) helps herself to the Principle of Humanity, and interprets the meaning of the interpretees’ utterances accordingly. This, Wiggins claims, entails convergence (between interpreter and interprettee) upon a judgment presumed to be true:
Such a theory is in one way terribly easy to imagine, if we will dignify with the name “theory” the descriptive anthropology that we can sketch of ourselves for ourselves. For this is something we already have. But if this is the theory that the psychologist uses to account for our response to the hoodlums’ action, then he is badly placed to deny that what provokes our responses is simply the callous cruelty of the act. (2002 [1987b], p.158)

Wiggins’ third argument is specifically directed against the error theory, which claims that our moral judgments are systematic errors generated by a psychological process of projection. Considering any particular occurrence of a moral judgment, the error theorist will say that «there are distinct, complex, scientifically palpable properties here each of which would suffice to bring down that particular “projection”. It may even be suggested that the scientific theory has no need to categorize reality in a way that corresponds at all to the moral agent’s own way.» (2002 [1987b], p.159)

Now, given that, according to the theory of radical interpretation, judgments attributed to the interpretee are presumed to be true, unless the interpreter can offer an explanation for their falsity, Wiggins proposes to compare the plausibility in the long run of these two options: on the one hand, the interpretee’s evaluative judgment are true; on the other hand, they are false and their falsity is explained as a sort of systematic error in reaction to episodes that are truly describable in non-evaluative terms (e.g. by invoking psychological processes of projection). It is mere prejudice, Wiggins claims, to think that the latter option is more plausible than the former. Wiggins casts the suspicion of
prejudice on those who require, as an adequacy condition on accounts of human conduct, that they account for the valuational in non-valuational terms (in passing, he seizes Mackie’s claim that moral properties are “ontologically queer”, and redirects it against error theory):

What I question is whether one can imagine a psychological theorist’s dispensing entirely with value-properties not just in the case of non-standard reactions, but in every case (even in the cases where there are standard shared responses to supposedly standard moral qualities), and whether one can imagine him dispensing altogether with predicates that possess non-accidentally the same extension as his subject’s evaluative predicates do. That is what will be required by thorough-going denial of substantial truth to pure value judgments. What stands in the way of taking such a tour de force seriously is the sheer unimaginability of someone’s supplanting by clinical modes of explanation most if not all of our everyday modes of rendering actions and beliefs intelligible; and the prejudicial conceptions of explanation — queer conceptions, I yield to the temptation to say — that are almost always called upon when philosophers seek to interest us in the possibility that all valuation rests upon some sort of error or illusion or when they declare, less strikingly but equally prejudicially, that a philosopher falls into illusion if he supposes that moral subjects’ idea of correctness in moral judgment might coincide with plain truth. (1987a, p.159-160)

This last argument takes us back to Wiggins’ fundamental opposition to reductionism. The constraints of reduction are untenable, and are imposed by the prejudiced project of purifying the language of literal descriptions and explanations of all concepts, and areas of discourse (including valuational
discourse), that do not behave according to the reductionist standards, which follow from the idea that science and logic enjoy (quasi) universal scope and authority.

For Wiggins, there is no credible way to account for valuational responses and judgments (including moral judgments) in non-valuational terms. There are no good reasons to deny that some evaluative moral judgments are indeed true; there are only prejudiced reasons, reflecting the prejudiced standards of reduction.

2.4. WIGGINS IN PERSPECTIVE

As elaborated by Wiggins, the underdetermination thesis relates a relatively developed theory of human nature, linked to a libertarian conception of responsibility and free will, and a relatively elaborate description of ordinary morality.

Following the thesis of underdetermination, human dispositions condition and constrain, but do not determine, the standards of correctness by reference to which we morally assess conduct, emotions and judgments. The vindication of those standards is grounded in Wiggins’ characterization of certain concerns, motives, purposes, needs, aspirations and expectations as unforsakeable. In virtue of the anthropocentricity thesis, the unforsakeable character of those
concerns and motives is sufficient to vindicate those ordinary moral standards that are rooted in them.

I have emphasized that Wiggins’ use of that vocabulary is vague, as it seems to designate items which partake simultaneously in the dispositional constitution of human beings and in the moral manifestations of this dispositional constitution. On the other hand, this vocabulary is suggestive, and adds substantive, phenomenological bite to the idea of underdetermination. Moreover, I have tried to limit the vagueness of the idea of the “unforsakeable” by proposing an elucidation of that notion in three points (2.2.2.), sorting out its descriptive and normative (vindicatory) dimensions, and its non-foundational character.

In sum, the idea of “unforsakeable” concerns, motives, etc. partakes at once in the description of human nature, in the account of the relation of underdetermination which connects it to the first-order ethic, and in the description and the vindication of that ethic.

That is the part of Wiggins’ moral anthropology which I take to be the most important. I think that its overall scheme, as just described, is adequate.

For one thing, I regard as highly significant Wiggins’ claim that the Humean theory of human nature on which his own theory is based demands to be completed by evolutionary theory. In this way, Wiggins explicitly calls upon a naturalist interpretation, not only of Hume’s theory, but of his own philosophy of value, which had up to that point (2006) remained mute on its relation to naturalism.
More substantively, I consider as highly plausible the idea that the concerns and motives which belong to categories (1) ("prohibitive aversion") and (4) ("observance") are indeed unforsakeable, in the sense which I elucidated, following which human existence is not and *ought not* be conceivable without them. Just as there is an irreducible element of appropriateness to the approbation expressed by certain valuational judgments, there is a sense of appropriateness to the experienced irremovability of those concerns and motives. Thus, to the *de facto* irremovability remarked by Strawson, Wiggins' idea of unforsakeability adds the idea of self-conscious vindication, on the ground just mentioned: not only we cannot get rid of them, but human existence is not and *ought not* be conceivable without them.

I find adequate Wiggins' "no priority view" of the deontological and the valuational within morality. His argumentation is somewhat vague (and there is business left "for another occasion" as he puts it), but the consequentialist thesis of the determination of the deontological by the valuational, and the deontological thesis of the determination of the valuational by the deontological seem less satisfactory than Wiggins's view. Moreover, Wiggins is right to claim that there is something problematic with this claim of virtue theory that «the organizing focus of morality is *virtue itself.*» Overall, Wiggins proposes a philosophy of morality that is more satisfactory than classical forms of moral theory. At this point in the
dissertation, this appeal of Wiggins’, quoted at the beginning of the chapter, might sound more profoundly right:

let us leave the expression “moral theory” in the hands of those who relish its present connotation of reconstruction or believe that it is the business of moral philosophy to trespass on the preserve of the moralist by remaking first-order morality. Until we attain some speculative understanding, it is too soon for us to commit ourselves to any of that. At the outset, let us bracket all preconceptions and allow the notions we already have to speak to us for themselves. (2006, p.7)

As I mentioned in the general Introduction, I find no aporia in Wiggins’ framework. It makes sense to claim that philosophers, like ordinary agents, may identify in the practice of giving for reasons, on a piecemeal basis, whatever concept is “unacceptable”, or unsuitable for realist interpretations. One gathers that when intelligible cognitive divergence occurs, in reason of identified or suspected cognitive failure, the use of certain terms may seem to be related to said failure: certain terms might be confusing or unmanageable, or may be conducive to cognitive failure in various ways. These then might be excluded as “unacceptable”, or (to use expressions of Richard Rorty), “more trouble than they’re worth”, or “better done without”. This without specifying “at the outset”, outside of that piecemeal practice, any kind of unacceptable concepts.

Moreover, Wiggins’ “equal and reciprocal partners” metaphysics of valuational experience provides an answer to the placement question. Moral properties might be attributed an ontological status similar to that of secondary qualities: real, but not entirely objective, because depending on the dispositional
constituent of the human species. Judgments considered true in the practice of giving and asking for reasons are taken to engage with “reality”. This is a part of “reality”, however, that is not entirely objective. Wiggins is adequately cautious with the notion “real”. Wiggins’ metaphysics of experience seems to me all the more credible in that it lends itself to an evolutionary interpretation — which unfortunately Wiggins does not tackle.

In regard to the issue of moral meaningfulness, Wiggins suggests, in ([1976]), that the most significant point that we can make about it is that evaluative moral judgments are candidates for truth. This suggestion is accompanied by the invalidation of the “reductionist” family of doctrines and theories presented above (2.3.2). For one thing, they are too impoverished conceptually to adequately capture moral features of the world as we humans relate to it. Secondly, they construe the experience of non-primary properties in terms of subjective states cut off from what they are directed to, which also inadequate.

Reverting to the thesis of underdetermination and the antireductionist cum anthropocentric view, I take this same part of Wiggins’s moral anthropology to be wanting on a number of points.

First, Wiggins exaggerates the polarization between reductionism and the antireductionist view. This polarization is well expressed by his spatial metaphor, following which the reductionist-cum-“cosmic standard view” doctrines constitute
an “external standpoint” on the world and the ordinary outlook; while the antireductionist view-cum-anthropocentric thesis constitutes an “inner perspective”. We will see in Chapter Three what is problematic about this polarization, and how Dennett avoids it.

Second, it is not enough to claim, as Wiggins does, that the theory of human nature demands to be completed by evolutionary theory. The completion must be accomplished. Dennett delivers.

Third, I find insufficient Wiggins’ argument that our ordinary morality presupposes our ordinary notion of free, responsible agency, in virtue of the fact that it is a morality of solidarity and reciprocity. As long as the grounds of the inference are not clarified, an independent defense of our ordinary sense of free, responsible agency is required. As we have seen in 2.2.3., Wiggins proposes one, but I find it unsatisfactory, as it relies on metaphysical considerations that lack clarity, and which are probably irrelevant anyway, as we will see with Dennett, whose compatibilist conception of responsibility and free will I endorse.

On this count, if we refer to those standards analyzed by Strawson, which I called the “commitment-responsive standard”, and the “responsibility-attributive condition”, it seems that Wiggins’ idea of the unforsakeable might vindicate the former; Wiggins does not simply point to the irremovability of certain dispositions, he proposes to vindicate (satisfactorily in my view) the standards that they underdetermine. On the other hand, to the extent that his vindication of
responsibility and free will is unsatisfactory, the “correctness condition” (which is the more fundamental one, on which the other depends) is left pending.

Fourth, whereas I find highly plausible the claim that those concerns and motives which belong to categories (1) (“prohibitive aversion”) and (4) (“observance”) are indeed unforsakeable, the claim that the concerns and motives belonging to categories (2) (“cooperative benevolence”) and (3) (“generalized beneficence”) are just as unforsakeable seems to me to require an argumentation that Wiggins does not provide.

It will be remembered from paragraph 2.2.2.1. that Wiggins establishes the following order of prescriptive priority between the four categories: the category of prohibitive aversion (1) is prior to that of observance (4), which is in turn prior to those of cooperative benevolence (2) and of generalized beneficence (3). These last two categories being weaker in the order of prescriptive priority, haven’t they fallen out of the sphere of the unforsakeable?

Phenomenologically speaking, it is a highly plausible claim that for anyone with so much as a moral sense, not only the prohibitive aversions subsumed under category (1) have priority over other concerns and motives, but human existence is not, and ought not be, conceivable without them. Placing category (4) (observance) in second position might seem surprising, as it corresponds to a set of artificial virtues, which is thus presumed to trump what Wiggins (following Hume) describes as our natural disposition to benevolence. But, again, the genetic priority of natural virtues over artificial virtues does not imply anything
regarding the order of prescriptive priority. Arguably, the claim that concerns and motives with justice, veracity, promise-keeping and the like trump concerns of benevolence is a plausible one.

The claim that the concerns and motives belonging to categories (2) and (3) are also unforsakeable is sensible, but it seems to me to require argumentation.

In spite of those shortcomings, the depth of Wiggins’ moral anthropology is remarkable. Each quotation in this chapter testifies to the insightfulness of his reflection, and of his reading of other philosophers. Now, I think that Wiggins' remarkable anthropology of moral life ought to be articulated in Dennett’s rationally acceptable language.

As I have mentioned in the Introduction, I find Dennett’s framework more satisfactory than Wiggins’, principally because I share his view of a rationally acceptable language. In my view, we can specify at the outset that participants in the practice of giving and asking for reasons should come to recognize that there is a certain kind of concepts that we are better off without (namely “skyhooks”, which will be described in 3.1.1.); that we should also do without essentialist thinking; that the placement question calls for the characterization of the relations between ordered levels of discourse; and that non-referential terms are acceptable when used correctly.

Moreover, Dennett’s doctrinal framework is clearer than Wiggins’, and it provides more straightforward answers to many important questions.
As for Wiggins’ “weak cognitivism”, I find no aporia within it, but again, under Dennett’s influence, I see the issue of the truth-aptness of valuational judgments as of lesser importance.
CHAPTER 3. DENNETT’S FRAMEWORK AND ACCOUNT OF HUMAN NATURE

For starters, let us rehearse the four qualities which I take to characterize the “rationally acceptable” language, distinct from ordinary language, which Dennett indirectly defines for purposes of “serious” descriptions and accounts.

First, it excludes a kind of concept which Dennett calls “skyhooks”.

Second, it excludes “essentialist thinking”, which is related to skyhooks, and substitutes for it “gradualist thinking”.

Third, it is regimented into four irreducible levels of discourse: the physical; the functional; the intentional; and the personal/moral. The interrelations between the four levels of discourse, and the relations between their respective subject matters, are characterized sufficiently to account for their irreducibility, and to provide credible answers to the placement questions which concern those subject matters, i.e. how they relate to the physical world, described as the subject matter of the physical sciences.

Fourth, it admits of “non-referential terms”, i.e. terms which do not refer to material objects. Moreover, it admits of non-referential vocabularies, and considers as acceptable that a whole level of discourse might be non-referential.
in this sense. Thus, the intentional and the moral levels of discourse may not be referential, and yet they may be rationally acceptable (and irreducible).

Dennett’s rationally acceptable language is associated to an approach of ontology which I characterize as “supple”, as based on “good reasons”, and as “comparative”.

It is also associated to an epistemology which may be characterized as “pragmatic" (or “pragmatist”), in that it downplays the significance of truth, and emphasizes the significance of warranted assertability, predictive success being designated as the most significant property of well-groundedness. I also contend that Dennett endorses what I have described as the “anthropocentricity thesis”, i.e. he recognizes that standards grounded in nothing “deeper” than the dispositional constitution of our species may be adequate.

Dennett’s rationally acceptable language, together with its associated ontology and epistemology, constitute the framework of an evolutionary account of human nature, linked to a metaphysics of experience not unlike Wiggins’, and to a theory of interpretation not unlike Davidson’s (which does not play the same theoretical role in Dennett’s framework as in Wiggins’, as we will see). Since Dennett’s metaphysics of experience is not crucial to his moral view, it will not be discussed in this dissertation.

Dennett proposes a “gradualist”, anti-essentialist, skyhook-free account of the emergence in Homo sapiens of language and of language-produced
phenomena like rational criticism, and of how this capacity brings about free will and morality. This, I contend, is Dennett’s own version of the underdetermination thesis. Although he does not put it in these terms, it appears from his account of the emergence of moral life (or “ordinary morality”) as a trait of the human species, that the dispositional constitution of our species conditions and constrains, but does not determine, the basic features of moral life (or ordinary morality).

This chapter presents, first of all, the Dennettian framework (3.1.) Then it presents the evolutionary account that Dennett articulates within this framework, which I divide in three steps. First, the emergence of basic phenomena like agency, normativity and intentionality, and the evolution of agents, and of kinds of agents (and at some point in the process, “kinds of minds”) from self-replicators to primates (3.2.). Second, the evolution of proto-morality in pre-linguistic Homo sapiens (3.3.). Third, the emergence of language, and of language-produced phenomena (3.4.). This last step corresponds to what I call “Dennett’s linguistic extension” of evolutionary theory.

Chapter Four will present Dennett’s account of morality, which, I contend, is a distinctively Dennettian moral anthropology. As we will see, it focuses on responsibility, free will, and moral decision-making, in connection with which Dennett indirectly singles out certain concerns and motives, and gives them a role similar to that which Wiggins gives to those concerns and motives which he characterizes as “unforsakeable”. At this point, it will appear that Dennett’s moral
anthropology is highly congenial with that of Wiggins, although their respective emphases are different — and complementary.

Dennett’s account of language-produced phenomena, including responsibility, free will and morality, is partly revisionary and partly vindicatory. This comes from the nature of his framework, which I describe as being poised between the two traditional doctrinal currents of analytic philosophy.

On the one hand, Dennett’s framework is different from the antireductionist frameworks of Wiggins and of Strawson (and from the “Oxonian” kind of framework in general). The main difference is that Dennett imposes, for purposes of “serious” descriptions and accounts, a “rationally acceptable language”, which satisfies the constraints described at the outset of this chapter. Descriptions and accounts couched in the rationally acceptable language have revisionary implications in regard to certain areas of ordinary language and of the ordinary outlook. For example, Dennett’s theory of consciousness, articulated at the functional level, is largely revisionary in regard to our ordinary conception of it.

On the other hand, Dennett’s framework is not “reductionist” in the sense criticized by Wiggins, for the reason that the rationally acceptable language it defines and imposes admits of non-referential terms and vocabularies, and recognizes the irreducibility of no less than three levels of discourse over and above the physicalist level. Dennett’s rationally acceptable language, then, is not overly restricted. Moreover, as Dennett adheres to the anthropocentricity thesis, accounts couched in Dennett’s rationally acceptable language may vindicate
certain other features of the ordinary outlook. Thus, his moral anthropology, articulated at the moral level of discourse, is largely vindicatory.

In Chapter One, we saw that Strawson’s view of rational acceptability, while it rightly rejects both the overly restricted and the overly liberal conceptions of the same topic, nevertheless stands in need of being refined, for the reason that it relies on a “distinct languages” argument that implies a false dichotomy between a “scientific language” and a moral-psychological language.

In Chapter Two, we saw that Wiggins’s version of that dichotomy is the opposition between an “outer view” constituted by reductionist doctrines, and an “inner view” purportedly constituted by the Davidsonian framework of radical interpretation, which Wiggins further elaborates.

Now we will see that Dennett proposes a regimented and multi-level view of rational acceptability. Like Strawson’s view, it rejects both those reductionist approaches that are overly restricted and overly revisionary, and those antireductionist approaches that are overly liberal. But its multi-level, gradualist character invalidates the outer/inner scientific/moral-psychological dichotomy that Strawson’s and Wiggins’s respective doctrines sustain.

3.1. THE DENNETTIAN FRAMEWORK
Skyhooks and the sort of cosmology induced by essentialist thinking (the “Cosmic Pyramid”) are described in 3.1.1.. The characteristics of Dennett’s rationally acceptable language, and of the gradualist accounts it articulates, are described in 3.1.2.. Dennett’s ontology is described in 3.1.3., his epistemology in 3.1.4.. The sorts of theories that Dennett’s criticism targets are described in 3.1.5..

Subsection 3.1.6. presents the essential of Dennett’s conception of the intentional level of discourse, both in its ordinary and “rationally acceptable” forms. That will help to understand the specificity of the personal/moral level of discourse, in Chapter Four. Subsection 3.1.7. presents the rationally acceptable functional level of discourse.

3.1.1. Essentialism, the cosmic pyramid, and “skyhooks”

Before Darwin, it was a widely shared idea (among philosophers and the reflective part of humankind) that the human mind and the phenomena that seemed to depend upon it (as we will see in 3.2., these include agency, interests, reasons, normativity and intentionality) constituted a special metaphysical realm, isolated from the rest of nature by a metaphysical gap which made it difficult to account for their mutual relation. It seemed that the human mind had a special essence, which it shared with mind-dependent phenomena. Explaining the very
existence of this special metaphysical realm without appealing to God was
difficult. Atheist, materialist thinkers had no satisfactory alternative.

For Descartes, there were two distinct metaphysical realms, corresponding
to two different substances: mind and matter — the latter including live
organisms, which he construed as complex mechanisms. As Dennett points out
(1995, p.64-68)⁵¹, philosophers who rejected substance dualism nevertheless
subscribed to a cosmological view which traded in metaphysical realms and
mysterious metaphysical gaps no less than substance dualism. Dennett calls this
view the “Cosmic Pyramid”.

The Cosmic Pyramid is a generic cosmological view according to which
the world (the “cosmos”) is structured along a scale of increasingly complex
metaphysical realms, each isolated from its neighbors by metaphysical gaps,
which make their mutual relations difficult to explain⁵².

The number of rungs (i.e. metaphysical realms) in the scale varies
according to versions. Dennett suggests (1995, p.64-65) that a “complete” pre-
Darwinian version of the Cosmic Pyramid would go as follows.

---

⁵¹ In this chapter as in Chapter Four, the name of Dennett will be omitted from the references to
his works.

⁵² Arguably, the Cosmic Pyramid could be imputed to some cosmological quarters of a broadly
shared ordinary outlook, entertained by lay persons since ancient times in many cultures and
civilizations (not solely Western). But Dennett is mainly concerned with its place within modern,
Western philosophy.
The abstract notion of Nothing (or “Nothingness”, for good philosophical measure) can be considered as the first rung of the scale, above which is Chaos (matter behaving in a disorderly way); then Order (matter behaving in an orderly way); then Design (which comes with Life, since the design in question is that which characterizes the living world); then Mind (namely the human Mind). Above Mind is God.

The higher the realm, the more “special” it seems, the more difficult it is to explain its very existence without appealing to God, and the more puzzling its relation to the realm below appears. Over Nothingness, the existence of Chaos (matter as such, independently of order) might not require much explanation, and invoking God might not be necessary (although Aquinas thought it was). Divine intervention might be more readily invoked to explain Order, although not with much insistence if no difference “in essence” is postulated between matter behaving chaotically and matter behaving in orderly fashion (in which case no metaphysical gap isolates Order from Chaos).

But the existence of the next realm up (Design and Life) was hardly conceivable, before Darwin, without some kind of divine intervention: it seemed that Design required a Designer, that life involved a special sui generis essence, and that a puzzling metaphysical gap isolated it from Order. As for Mind (and mind-dependent phenomena), it seemed to presuppose a very special sort of design, to involve a very special sui generis essence, and to be isolated from Design by a metaphysical gap that was downright mysterious.
Neither values in general, nor morality in particular, are mentioned in Dennett’s brief description of the Cosmic Pyramid. Arguably, there are various possibilities, including the two following: values may be integrated to the realm of Mind, as a mind-dependent phenomenon; or they may be taken to partake in a distinct metaphysical realm, presumably not inferior to the realm of Mind (thus creating another metaphysical gap).

The growth of scientific knowledge put increasing pressure on the Cosmic Pyramid. As it made progressively more precise the picture of an entirely material universe, it became increasingly clear that explanations that invoked God were unsatisfactory appeals to a *Deus ex machina*. As for the essential properties invoked to characterize and account for Design (and Life), and for Mind, the progress of the materialist conception of the universe made them seem increasingly "special", metaphysically speaking, and therefore increasingly suspect. Before Darwin, however, no one could provide a satisfactory, thoroughly materialist account of Design and of the human Mind. The theoretical void left by God’s early dismissal could only be filled with vague thoughts which did not provide satisfactory answers to some deep questions: how come there is order in the material world? How come there is design in the world, namely that of living things? How come there are beings with minds, namely human beings? How does our mind fit in the natural world?

According to Dennett, Darwin’s theory revealed that it was essentialist construals of design and of mind that made problematic the explanation of their
existence, and of the phenomena associated to them (life and mental
phenomena).

Dennett claims that thinking of life and of mental phenomena in
essentialist terms generates a faulty alternative: «Philosophers often maneuver
themselves into a position from which they can see only two alternatives: infinite
regress versus some sort of “intrinsic” foundation — a prime mover of one sort or
another.» (1998 [1994], p.361) Infinite regresses being an unattractive option, the
typical outcome of the faulty alternative is the postulation of a “prime mover” of
sorts.

Dennett calls “skyhooks” concepts which purport to designate items
fulfilling a fallacious “regress-stopping” (sometimes foundational) explanatory
role. Skyhooks are interpreted essentialistically and realistically. Dennett
condemns the essentialist interpretation, and considers that the realist
interpretation is unsupported. The problem with skyhooks does not reside in their
“non-referentiality”, but rather in the fallacious character of their regress-stopping
role. Skyhooks get their name from the fact that the regress-stopping role they
are supposed to play in regard the domain of life, or of the domain of mind, is
similar to that of «miraculous lifters, unsupported and insupportable.» (1995,
p.75) The category of “skyhooks”, then, does not correspond to the traditional
reductionist category of “empty terms”. What makes skyhooks rationally
unacceptable is not that they fail to refer to anything that satisfies credible reality
criteria, but that their fallacious regress-stopping role implies illusory
metaphysical gaps between what are mistakenly taken as metaphysically distinct
categories of items and phenomena. Generally speaking, skyhooks block inquiry
and constitute an obstacle to making the world more predictable and intelligible.
To put Dennett’s idea in somewhat Wigginsian terms: the concerns, needs, and
aspirations that we have as human beings are much better served without any
skyhook.

Some skyhooks are cousins of the category mistakes that Ryle warned us
against. As Ryle explained in *The Concept of Mind*53, the correct use of the
expression “pair of gloves” does not introduce any distinct item over and above
the two gloves individually designated. Mistaking the “pair of gloves” for a distinct
entity over and above the two individual gloves is a category mistake which
introduces into our ontology a non-existent item, thus inducing confusions and
mystery-mongering (if, for example, we attribute to this mythical entity the
property to keep hands warm, and propose that the existence of the two
individual gloves is merely a “necessary condition” for the existence of the
mythical entity). Ryle’s main point in *The Concept of Mind* is that the correct use
of mental terms is that of dispositional terms; and that mistaking them for terms
referring to material items inserted in the causal order of the world is a category
mistake similar to that which consists in mistaking the expression “pair of gloves”
for a term referring to an entity distinct from the two individual gloves, and
invested with properties of its own. Similarly, certain terms have a legitimate use
and interpretation, and become skyhooks when used and interpreted incorrectly,

namely when they are interpreted essentialistically and realistically, in order to play an improper regress-stopping role. This is what happens, according to Dennett, with our intentional vocabulary (“beliefs”, “desires”, etc.), and of our moral vocabulary (“responsibility”, “free will”, “person”, “wrong”, etc.)

As we will see in section 3.4., Dennett takes the theory of the mind that he calls “Cartesian materialism” to make a mistake akin to the postulation of the mythical, substantiated “pair of gloves” over and above the two individual gloves.

The link between essentialism and skyhooks is not logically necessary, but the former has a strong tendency to induce the latter. This is illustrated by «the well-known but ill-understood fallacy, the sorites.» (1984, p.84). This fallacy consists in taking a certain kind of thing (say, “B”) to have an essence such that there is a metaphysical gap between it and other kinds of things (say, “A”), which rules out the possibility that things of the B kind gradually emerged out of things of the A kind. Consequently, its origins can only be construed along one of two possibilities: either there is an infinite regress, or there is a skyhook.

Dennett’s favorite example of the sorites fallacy (which, he mentions, he has adapted from an article by David Sanford54) is the postulation of “the prime mammal”. The fallacious reasoning starts with a premise that is correct, but which becomes misleading if interpreted in essentialist terms: every individual mammal must have a mammal for a mother. The essentialist error is to think that if we

found nonmammalian ancestors to mammals, then there couldn’t be any real mammals. Now, since there are real mammals, either there has always been mammals (which is absurd) or else there must have been a First Mammal — a regress-stopping skyhook:

Philosophers tend to like the idea of stopping threatened regress by identifying something that is — must be — the regress-stopper: the Prime Mammal, in this case. It often lands them in doctrines that wallow in mystery, or at least puzzlement, and, of course, it commits them to essentialism in most instances (The Prime Mammal must be whichever mammal in the set of mammals first had all the essential mammalian features. If there is no definable essence of mammal, we’re in trouble. And evolutionary biology shows us that there are no such essences.) (2003, p.127)

We will see in the next subsection (3.1.2.) what “evolutionary biology shows us”. Right now, it is worth mentioning already that for Dennett, much of past and present philosophy of mind, and much of past and present moral philosophy, is faulted by essentialist thinking and the skyhooks it induces. In the following quote (which is the continuation of the passage first quoted in this paragraph), he attacks the idea of “intrinsic worth”, and the idea of “original intentionality” as skyhooks. And he points to an alternative:

For instance, it has seemed obvious that for some things to be valuable as means, other things must be intrinsically valuable — ends in themselves — otherwise we’d be stuck with a vicious regress (or circle) of things valuable only as means. It has seemed similarly obvious that although some intentionality is “derived” (the aboutness
Thus, Dennett excludes from his rationally acceptable language skyhooks, essentialist thinking and the Cosmic Pyramid.  

3.1.2. Gradualist accounts and the rationally acceptable language

According to Dennett, few philosophers have acknowledged the deep philosophical implication of evolutionary theory: as far as design and life, and mind and mental phenomena, are concerned, there are no essences, no “marked foundations or thresholds”.

55 Dennett defends a radical anti-essentialism, which seems to imply that no natural kind could be said to have an essence (either real or nominal). Such a position is certainly more difficult to hold in regard to chemical elements and compounds than in regard to the phenomena and categories of the living world. However, I assume that Dennett's anti-essentialism about chemical elements and compounds, and his anti-essentialism about the living world (including the mind and mind-dependent phenomena), are logically independent. Only the latter interests us here.
Dennett stresses that Darwin’s theory was a major contribution to the materialist understanding of the universe. It explained how species with increasingly complex designs and abilities gradually evolved out of the sieve of natural selection, right up to and including the humans species; and it made it possible to envisage how the most primitive organic designs — how life itself — could have emerged from mere inanimate Order. This, Dennett claims, was a quantum leap in the materialistic understanding of the universe, which definitively refuted the theses that underlay the Cosmic Pyramid. Moreover, Dennett insists, the gradualist character of Darwin’s showed how to account for Life and Design, as well as for the human Mind (and eventually for free will, personhood, moral responsibility, and moral properties) without calling upon essences. More precisely, it showed that neither the groupings designated by taxonomical categories (species, genera, phyla, etc.), nor life itself, nor the human mind, have essences — neither “real”, nor “nominal”. Overall, the theory showed that there were no metaphysically isolated realms.

In the wake of Darwin, the idea of God became theoretically useless. In the wake of further scientific progress — notably the “neo-Darwinist” synthesis of evolution theory and molecular biology — essentialist construals of Design and Life collapsed. We should remember that until the theory of DNA-based reproduction and evolution was firmly established (around the middle of the 20th century), some philosophers were ready to call upon notions of “vital forces” (construed as sui generis, essential, metaphysically special properties) to characterize and account for Life. But an empirical theory that shows that life is a
material phenomenon (and which greatly clarifies the role played by genes in natural selection, and thus in evolution) leaves little doubt about the possibility of life arising from inanimate matter, and it shows that it is a mistake to think of Design as a metaphysical realm isolated from mere Order by a metaphysical gap, to be characterized and accounted for by skyhooks. Once Crick and Watson have shown how DNA is structured and how it functions, invoking “vital forces” to account for the phenomenon of life was no longer an option. Bergson’s “élan vital” was stopped by the science of life.

Dennett claims that essentialist, skyhook-involving accounts of mind-related phenomena are bound to collapse under the pressure of the evolutionary theory, just like the skyhook account of life collapsed under the pressure of the neo-Darwinian synthesis of evolutionary theory and molecular biology.

Dennett emphasizes that we now have at our disposal all the elements of an empirically based evolutionary theory of the mind, constituted on the basis of a multiplicity of empirical findings (as well as of other sources of evidence, as we will see in 3.3.1., when we discuss game-theoretic models), which show how the various kinds of minds that are to be found in the living world, and the phenomena that are associated to them, were gradually built in the course of evolution by natural selection.

Dennett’s philosophical and evolutionary theorizing unites those elements into a coherent picture that takes us from the very beginnings of life up to the apparition of language-using Homo sapiens. Along the way, it accounts for the
gradual deployment of phenomena which are often mistaken for specifically human phenomena (notably interests, intentionality, normativity and practical rationality). Dennett’s point is precisely that these phenomena are not specifically human — only their linguistic version is.

The emergence of language marks a turning point in history. Beyond that point, the history of language-using Homo sapiens is accounted for by what I call Dennett’s "linguistic extension" of evolutionary theory. Dennett uses certain insights fostered by Wilfrid Sellars and by psychologist Julian Jaynes to build this linguistic extension, which accounts for the specifically human, linguistic version of the phenomena accounted for in the evolutionary theory; and it accounts for the language-involving production of certain phenomena which (by contrast with normativity, practical rationality and intentionality) are indeed specifically human phenomena. These notably include the attribution of selves as the “centers of narrative gravity” of individual biographies; and rational criticism, which in turn gives birth to values.

Such a gradualist account is to be couched in Dennett’s rationally acceptable language, which articulates four irreducible levels of discourse, each having its own indispensable use and specific scope. Their interrelations are determined by how we, language-using human beings, endowed with the dispositional constitution that we have inherited from the evolution of species, relate to their respective subject matters. That is how Dennett accounts for the relation between the physical and the functional, the functional and the
intentional, the intentional and the personal/moral. Those topical domains, as should be clear, do not involve distinct essences, and their relations do not bridge any *metaphysical* gaps. However, they involve other sorts of gaps.

The descriptions and explanations articulated at the *functional* level are irreducible to the *physical* level in virtue of the fact that from the physical level, it is impossible to capture the *function* of a structure (anatomic, physiological or mental), and to describe the *functional behavior* of that structure.

That is, more precisely, because functional concepts, and the rules of their use (including the heuristic rules that guide and constrain function attribution), imply *the selection, compression and schematization of information which, couched in physicalist language, would be intractable*. In a physical description of the behavior of a structure performing a certain function, the information which corresponds to the description of the performance of this function is literally drowned in intractable masses of information that are not relevant to it. Thus, capturing functions, and producing functional descriptions and explanations, implies, for starters, the neglect of vast amounts of physicalistic information. Furthermore, the physicalistic information that is relevant to the performance of the function is still intractable; it must be compressed and schematized, in just the way implied by the functional vocabulary and the rules of its use, in order to be appear as a tractable description.
As a result, the functional description/explanation, is not translatable at the physical level — it is *irreducible* to it, in reason of the “heuristic blindness” and informational intractability which cripple the lower level.

So there *is* a gap between the physical level and the functional level. And there is a sense in which it might be legitimate to say that functional phenomena are located at a “higher level of material organization” than the merely physical phenomena that we observe in the cosmos. But the gap is informational, heuristic and conceptual, and occurs between levels of discourse and discovery; and the complexity of organisms and phenomena which makes the functional level of discourse indispensable does not involve any “emerging” property, nor any metaphysical essence, and therefore no metaphysical gap between distinct domains.

In fact, in the case of a mechanical structure performing an extremely simple function, it might be possible to do away with the functional vocabulary and its associated theoretical apparatus, and still to succeed in describing, in physical terms, the behavior of the structure. However, the slightest increase in the functional complexity of the structure entails heuristic blindness and the intractability of information.

Climbing from the physical to the functional level of description and explanation involves a certain tradeoff between on the one hand exactness, precision, and reliability, and on the other hand predictive power. That is, if we abstract from the intractability problem, physical descriptions and explanations can be much more exact, precise and *in principle* reliable, than functional
descriptions and explanations could ever be (one of the reasons for that is that the former but not the latter include nomological statements, in the form of mathematical formulas). However, given that there is a problem of informational tractability at the physical level, we must climb to the functional level, and so forsake a measure of exactness, precision and in-principle reliability, in order to obtain workable descriptions and explanations, and actual predictive power.

One might insist that the functional level of discourse is in principle reducible to the physical. However, when we talk about the function of a biofunctional structure (either organic or cognitive), there are no necessary and sufficient conditions to identify that function. It is attributed on the basis of hypotheses as to what it was selected for, and whether it was or not exapted at some point of its history.

Descriptions and explanations articulated at the intentional level are irreducible to the functional level for similar reasons: the increase in complexity of functional design of organisms translates into such complexity of their behavior, that the attribution of propositional attitudes (principally beliefs and desires) to them, as agents, becomes practically necessary in order to capture their behavioral patterns, and to make their behavior minimally predictable and intelligible. We do not infer the existence of propositional contents in the brains of the agents whose behavior we anticipate and make intelligible. According to Dennett, there is no fact of the matter concerning propositional contents borne by the central nervous system of agents. Moreover, the informational gap between
the functional and the intentional levels is deepened by *indeterminacy*, of the same sort which, according to Quine, characterizes “radical translation”. How is it, then, that human beings are responsive to meanings, and appreciate meanings? Dennett’s answer, as we will see in 3.4.2.3., is that the brain is a syntactic engine which *approximates* a semantic engine.

Finally, why is moral discourse, including attributions of responsibility, and evaluative and practical moral judgments, irreducible to the intentional level of discourse? Because human beings have appropriated the reasons why they act as they do and hold the judgments they hold. The assessment of these reasons involves the use of concepts of responsibility and personhood, which are not part of the intentional vocabulary. This we will see in Chapter Four.

In this way, most subject matters are clearly related to the subject matter of the physical level of discourse, i.e. the physical world, as describable by the physical sciences (most but not all, e.g. if mathematics deal with abstract entities, then they are not thus related). Of an item that cannot be designated by a physicalist description, we should be able to say something about how it relates to those items that can, and Dennett’s rationally acceptable language allows us to do that. Whatever topic it bears upon is “placed” within the world described and explained by the physical sciences.
3.1.3. Dennett’s ontology

I describe Dennett’s approach to ontology as “supple”, based on “good reasons”, and “comparative”. What is that to mean?

First of all, in admitting that non-referential terms may be acceptable, Dennett takes up certain insights of the instrumentalist tradition, following which the rational acceptability of terms does not depend upon whether they may be interpreted as “real”, but rather on whether they contribute, as they occur in various judgments and theories, to gains in informational tractability, in predictive power, and so in the intelligibility of the world.

Dennett introduced the expression “non-referential” in his very first book (1969) to characterize those terms, and families of terms — areas of discourse — which we have no good reasons to interpret as “real”, but which may nonetheless afford gains in informational tractability, predictive power and general intelligibility. Upon the condition that they do, those terms, or families of terms, are meaningful (and rationally acceptable).

Dennett came to endorse Reichenbach’s tripartite distinction between “illata”, “abstracta”, and “concreta”. These designate three classes of ontological items, according to their ontological statuses. Concreta are real, material entities. Illata are posited theoretical entities, whose reality is a matter of probability. “Abstracta” are another sort of theoretical entities: «calculation-bound entities or
logical constructs» (1987 [1981], p.53), of which centers of gravity are his prime example.

Although Dennett’s and Strawson’s respective ontologies are different (the former is not based, as the latter is, on the view that concepts command certain ontological commitments, depending on their position of in the conceptual hierarchy), they both substitute a plurality of ontological possibilities for the real/unreal dichotomy. It is in this sense that Dennett’ ontology is supple⁵⁶.

Dennett advocates an approach of ontology that is based on “good reasons”, and that is “comparative”. That means three things.

First, the ontological interpretation of particular terms, or of families of terms, is decided in the practice of giving and asking for reasons. No firm, formal reality criteria can be derived a priori from “first principles”, independently of the practice of rational criticism.

Second, what is most significant in regard to the ontological interpretation of terms is not the ultimate verdict that one reaches so much as the reasons that

⁵⁶ Dennett seems to have dropped the term “non-referential” after he adopted Reichenbach’s threefold distinction. I retain it to mean “not referring to material objects”, i.e. concreta. It need not be confusing that terms that are "non-referential" in this sense might refer to abstracta or illata. Note that in the traditional reductionist use of Reichenbach’s threefold distinction, illata and abstracta are the referents of theoretical notions that are used parsimoniously, in the strictly regimented context of scientific theories. Dennett makes a more liberal use of the Reichenbachian notions.
one provides in its support. That does not imply that the ontological verdict is ultimately irrelevant (as we will see shortly, Dennett advocates caution in the attribution of the “reality” status). The point is rather that the rationale supporting an ontological verdict is at least as significant as the verdict itself.

Third, the significance of a particular ontology does not reside in its being a tentative list of the world’s furniture, so much as in the similarities and differences that it establishes between the verdicts reached in regard to various terms, and families of terms (and so in regard to various ontological items and kinds of items). Dennett suggests that most philosophers have come to such a comparative view, which incidentally is no different from the ordinary “ontological attitude” that we entertain in common life. Evoking the disagreements that his ontological interpretation of propositional attitudes elicits, Dennett writes:

[…] I will assume that these disagreements all take place within an arena of common acceptance of what Arthur Fine calls NOA, the natural ontological attitude (Fine, 1986 […]). That is, I take the interest in these disagreements to lie not in differences of opinion about the ultimate metaphysical status of physical things or abstract things (e.g., electrons or centers of gravity), but in differences of opinion about whether beliefs and other mental states are, shall we say, as real as electrons or centers of gravity. (1998 [1991], p.99)

57 This last point, Dennett remarks, is argued by Arthur Fine in The Shaky Game: Einstein, Realism, and the Quantum Theory. Chicago U.P. 1986.
Three related principles are associated to Dennett’s “good reasons”, comparative approach to ontology.

First, the ontological interpretation of terms, or of families of terms should be approached on a piecemeal basis: «Instead of a general argument there must be detailed investigations of individual words and families of words, and we must be alert to the possibility that only a partial case can be made.» (1969, p.13)

Second, as we assess the ontological interpretation of particular terms, or of families of terms, we must approach them with an attitude of ontological neutrality, and carry off different sorts of test and investigations, which will provide us with reasons to be weighed in rational discussion. Decisions regarding the ontological interpretation of terms are reached by weighing reasons, pros and cons, and they are revisable. Dennett gives examples of the questions that we should ask in this connection, and proposes how the answers should count in our ontological verdicts. Thus, the realist interpretation of a term commits one to answering a few questions about its referent:

Answering yes or no to an ontological question only begins to have some point when we have decided that granting the existence of something licenses us to ask (and expect answers to) certain very general questions about it, e.g., what sort of thing is it?, does it exist in time?, and especially, is it identical with x? (1969, p.11; emphasis added)

The last question in the quote above serves to introduce the idea that the identity of a term with a physicalist description constitutes an overwhelmingly good reason to interpret this term realistically. Conversely, the impossibility of
establishing such an identity does not bar realistic interpretations, but it suggests caution.

The third principle is suggested by the list of questions which, according to Dennett, one is committed to give an answer to in the case of realist interpretations: Dennett advocates parsimony in the attribution of the reality status. Withholding the status “real” does not automatically disqualify a term as meaningless and unacceptable, it simply categorizes it as non-referential. Short of such conservativeness, we might commit the sort of category mistake mentioned in the previous subsection. Unjustified realist interpretation of a concept, in combination with its assignment to an essentialistically construed regress-stopping role, is bound to make a skyhook out of what might otherwise be a useful, even indispensable concept.

However, being cautious is not being over-cautious. Dennett rejects classical, across-the-board instrumentalism on the count that it bars the realist interpretation of certain terms that we have overwhelmingly good reasons to interpret realistically:

Classical instrumentalism was an all-embracing view, a complete rejection of realism; one was instrumentalistic not only about centers of gravity and parallelograms of forces and the Equator, but also about electrons, cells, planets — everything but what could be observed with the naked senses. […] From the outset I have maintained a contrast between my realism about brains and their various neurophysiological parts, states, and processes and my “instrumentalism” about the belief-states that appear as abstracta […] (1987b, p.71-72)
Reflecting on the confusions that were caused by having let his own position be characterized as “instrumentalism”, Dennett states the comparative character of his ontology, and emphasizes the special importance it takes when it comes to intentional vocabulary, and the propositional attitudes this vocabulary denotes. (Besides, the suppleness and comparativeness of his ontology have epistemological implications, which are alluded to in the following quote, and to which we will return in the next subsection, 3.1.4.):

My *ism* is whatever *ism* serious realists adopt with regard to centers of gravity and the like, since I think beliefs (and some other mental items drawn from folk psychology) are *like that* — in being *abstracta* rather than part of the “furniture of the physical world” and in being attributed in statements that are true only if we exempt them from a certain familiar standard of literality. (1987b, p.72)⁵⁸

Finally, it will be useful to briefly survey Dennett’s take on three cases of meaningful but non-referential terms, two of which he takes from (Quine 1960)⁵⁹.

---

⁵⁸ Even more precisely put: «Beliefs and desires of folk psychology (but not all mental events and states) are abstracta» (1987 [1981], p.53). That is, they have the same ontological status as centers of gravity.

⁵⁹ Dennett shares Quine’s conservativeness regarding the ontological interpretation of terms, but considers, contra Quine, that non-referential terms may be rationally acceptable.
One case concerns terms that are narrowly “locked into” certain particular idioms. Quine points out that when we say that, e.g. we do something for someone’s sake, the word “sake” does not refer to anything. Dennett emphasizes that the idioms in which the word “sake” occurs are not analyzable into elementary components: «The contexts maintain their significance but are not subject to further logical analysis.» (1969, p.14)

Another case discussed by Quine is that of nouns for units of measure (such as “mile”). These are already more problematic, but Quine considers that «they are best viewed as integral parts of a small group of idioms rather than as full-fledged nouns which pick out distinguishable items in the world.» (1969, p.7-8). Again, Dennett points out that these terms are non-referential, and yet they are meaningful.

Dennett then discusses an example of his own, that of the noun “voice”. He claims that this term cannot be identified to physical descriptions; that it is difficult to say precisely “what sort of thing it is”, “whether it exists in time”, etc; and that, although there are some good reasons to interpret “voice” realistically, the case for realism is in the end too weak. (1969, p.11). Dennett suggests that we may treat the sentences in which the term “voice” occurs as a class of idioms, somewhat as we treat sentences featuring words for units of measure. “Voice”, then, is non-referential. Yet, it is meaningful and the sentences in which it occurs may be justified. We can talk meaningfully about voices, without committing ourselves to the view that the term refers to concreta.
Again, what is most significant in Dennett’s reflection is not his verdict that “voice” should be treated as non-referential, but rather the principle of approaching a particular area of discourse with an attitude of ontological neutrality, and then the nature of the rationale offered in support of the ontological interpretation, including notably comparisons with other kinds of terms, other kinds of items.

At this point, we can see that the intentional and the moral areas of discourse need not refer to *concreta* in order to be rationally acceptable, and meaningful. The constraints that the rational acceptable language exerts on psychological and moral ordinary discourse comes down to the exclusion of skyhooks. The placement question in regard to their respective objects will be answered by evolutionary theory and its linguistic (and Dennettian) extension, and by the adequate regimentation of the rationally acceptable language in which it is couched.

3.1.4. Dennett’s epistemology and the anthropocentricity thesis

Dennett proposes a view of scientific knowledge according to which science, acknowledging the cognitive limitations of human beings, does not aim at attaining the truth so much as attaining predictive success. This suggests that Dennett accepts the principle that adequate standards of correctness might be
grounded in the dispositional constitution of the human species, rather than in extra-human principles — in other words, that he adheres to the anthropocentricity thesis.

This chimes with the generally pragmatist tone of his epistemology, which downplays the importance of truth, and emphasizes the importance of warranted assertability. Properties of well-goundedness are, generally speaking, properties of descriptions and accounts such that they contribute in making the world more intelligible. Among them we may count general coherence, “epistemic fertility”, etc. But the property of well-groundedness to which Dennett accords the greatest importance is predictive success.60 Let us take these points in turn.

I have already mentioned in 3.1.2. that in order to gain actual predictive power, it is necessary to forsake some precision, exactness, and in-principle reliability.

Dennett suggests that something similar happens in ordinary perceptual experience, as our perceptual and cognitive systems automatically schematize masses of data that would otherwise be intractable (1998 [1991], p.99-104). Thus, in order to extract significant patterns in the conduct of agents, and in the situations that we find ourselves in, we unthinkingly focus on certain salient

---

60 As we will see in 3.1.6., Dennett claims that predictability is much more than a property of well-groundedness: it is an important component in the very intelligibility of agents, and of the world in general.
features of conducts (and of situations), and barely notice — or simply ignore — the rest. This form of data schematization is part of our “natural attitude”. Our ordinary, natural cognitive attitudes are naturally poised to strike a tradeoff between on the one hand the predictive reliability that more precise and exact information affords in principle, and on the other hand the schematization which makes information tractable, and in this way affords actual predictive power.

Dennett remarks (1998 [1991], p.104-105) that science does not proceed otherwise, and suggests that the aim of scientific models and theories is not to attain at the truth, so much as to strike a tradeoff between on the one hand exactness and precision (and consequently the reliability of predictive success that such qualities afford in the ideal case where intractability is not a problem), and on the other hand schematization, informational tractability, and a predictive power that may be diminished in comparison with the ideal, but which has the advantage of being actually attainable.

Scientific models and theories capture patterns in the behavior of things, which they must extract from noise-laden empirical data, at the cost of a certain measure of idealization. The higher the idealization, the higher the lost in accuracy of empirical data, and therefore the higher the loss of in-principle reliability. On the other hand, idealization affords a gain in the compactness of the descriptions that “feed” the models, i.e. in informational (or “computational”) tractability; and this gain in tractability translates into a gain in actual predictive
power (without the compactness, the descriptions would be intractable, and so would afford no actual predictive power):

The ubiquitous practice of using idealized models is exactly a matter of trading off reliability and accuracy of prediction against computational tractability. A particularly elegant and handy oversimplification may under some circumstances be irresistible. The use of Newtonian rather than Einsteinian mechanics in most mundane scientific and engineering calculations is an obvious example. A tractable oversimplification may be attractive even in the face of a high error rate; considering inherited traits to be carried by single genes “for” those traits is an example; considering agents in the marketplace to be perfectly rational self-aggrandizers with perfect information is another. (1998 [1991], p.105)

The notion of “accuracy”, mentioned above, implies the notion of truth. Although Dennett downplays the role of truth, he does not whisk it out of the picture either. However, he emphasizes that the truth to which we have access is mostly approximate truth.

The issue of the relationship between warranted assertability and truth (for example, on what conditions the former can be taken to be conducive to the latter) is a deep one; besides, the role of approximate truth in regard to knowledge is a standard issue in epistemology manuals.61

---

61 For example, see (Audi 1998, p.255-256)
Dennett, however, does not usually engage in “pure” metaphysics, semantics and epistemology. The epistemological question that is of most concern to him is how we should assess the cognitive status of sentences in which propositional predicates occur, given the ontological interpretation he favors in their regard. That is his concern in the following quote, which can nevertheless be taken as a statement of his general epistemological position. This is the continuation of the passage quoted in the previous subsection (3.1.3.), in which Dennett rejects instrumentalism, which he tackles here from the epistemological side:

Some instrumentalists have endorsed fictionism, the view that certain theoretical statements are useful falsehoods, and others have maintained that the theoretical claims in question were neither true nor false but mere instruments of calculation. I defend neither of these varieties of instrumentalism; [...] people really do have beliefs and desires, on my version of folk psychology, just as they really have centers of gravity. Do I then grant that attributions of belief and desire under the adoption of the intentional stance can be true? Yes, but you will misunderstand me unless you grant that the following are also true:

(1) The gravitational attraction between the earth and the moon is a force that acts between two points: the two bodies' centers of gravity.

(2) Hand calculators add, substract, multiply, and divide.

(3) A Vax 11/780 is a universal Turing machine.

(4) SHAKEY [a robot whose “vision” Dennett discusses in many writings — P.C.] makes line drawings even when its CTR is turned off.
It is arguable that each of these is a useful, oversimplifying falsehood; I would rather say that each is a truth one must understand with a grain of salt. I have no official, canonical translation of that familiar phrase, but I also do not see the need for one. I would rather make my view as clear and convincing as I can by explaining why I think all belief talk has the same status (veritas cum grano salis, to put it technically) as (1-4) (1987c, p.72-73).

According to Dennett, in the ordinary practice of giving and asking for reasons, we attach great importance to the predictive reliability of our experience and judgments, while we take truth with a grain of salt.

Citing the predictive success of a particular judgment (or description, explanation, etc.) is bound to be considered by most participants as a very good reason to take it as well-grounded. We are not indifferent to truth, but we are mostly concerned with reaching an effective balance between informational tractability and reliable, attainable predictive success. As examples of this, he points to attributions of propositional attitudes, and to approximate measurement statements. According to Dennett, we assess both sorts of judgments in the same way: we care mostly about reliable predictive success, but we accept to regard as true (with a grain of salt) well-grounded examplars of both sorts of judgments.

However, there is an important divergence between the natural epistemic attitude on the one hand, and science and Dennett's epistemology on the other. The former considers that everyone enjoys a privileged access to one's own
consciousness, and that consequently the “first-person point of view” enjoys an epistemic privilege in regard to the description of those phenomena which seem to occur within what appears to us like the “inner space” of consciousness. Science and the Dennettian epistemology acknowledge that in some sense, everyone does have a privileged access to one’s own mental states, but denies that there is an “inner space” of consciousness within which various phenomena occur, in regard to which the “first-person point of view” would enjoy an epistemic privilege.

Finally, it should be noted that as we climb the scale of levels of discourse in the rationally acceptable language, the scientific approach appears decreasingly effective in buying predictive power (this is something that Dennett himself does not directly discuss, but which I take to be implied by his framework as a whole).

At the physical level of discourse, the physical sciences afford much more predictive power than folk physics, so they must be closer to the truth (with a grain of salt). There are considerable differences between physical science and folk physics, and the former enjoys primacy over the latter. It is at this level that the difference between the “manifest image” and the “scientific image” is most telling. There is no point in trying to replace altogether folk physics by physical science, but the latter certainly has revisionary implications in regard to the former.
Biology makes a scientific use of the functional vocabulary, but before Darwin, the main difference between the biological use and the ordinary use of that vocabulary (and so between scientific and ordinary functional thinking) was relatively modest. Darwin’s theory of evolution by natural selection increased the distance between the ordinary use and the scientific use of the functional vocabulary, in the following way. The ordinary use of the functional vocabulary comes with a “folk functional theory” which tends to assume that a functional design presupposes a designer. Darwin's theory showed that this presupposition, in its application to the natural world, was erroneous. Precisely, it is a skyhook. Evolutionary biology is more fertile epistemically, and affords more predictive power, than pre-Darwinian biology. Its difference from ordinary functional thinking is also more significant, because it involves the removal of a major skyhook, which is something of an “automatic” assumption in ordinary functional thinking. If we revise our ordinary functional thinking in the light of its scientific counterpart, we may emancipate it from this skyhook. As we will see in 3.1.7., rationally acceptable functional thinking includes the philosophical doctrine of functionalism, as used in the philosophy of mind and in the cognitive sciences — all disciplines which allow us to get rid of many basic skyhooks.

Is there a scientific use of the intentional vocabulary? Not really. The ordinary use of the intentional vocabulary, in ordinary psychology (mostly in ordinary intentional ascription), is quite effective, and achieves significant predictive success. The rationally acceptable use — and therefore scientifically acceptable, but not “scientific” strictly speaking — of the intentional vocabulary
simply makes explicit the rules for the attribution of propositional attitudes, and advises against skyhook-inducing moves, such as the realist interpretation of intentional states, which tends to imply such skyhooks as the idea of “original intentionality”, or “original understanding” (which will be refuted in 3.4.2.). Dennett organizes the rationally acceptable use of the intentional vocabulary into a theory, which he calls the “theory of intentional systems” (to which is adjoined “notional world psychology”, as we will see in 3.1.6.). But ordinary psychology, with its ordinary use of the intentional vocabulary, is the source of those rationally acceptable intentional theories, and the latter are not fundamentally different from the former. In fact, rationally acceptable intentional theories vindicate ordinary psychology to a certain extent. At the same time, it guards us against realist interpretations of intentional states.

At the personal level of discourse, the rationally acceptable use of the moral vocabulary provides a largely vindicatory account of ordinary moral life. That is, it phenomenologically describes the essential features of moral life, genealogically explains their origins, informally analyzes the ordinary use of certain ordinary moral expressions — in other words, it constitutes an anthropology of moral life. Its revisionary aspect, again, corresponds to the elimination of whatever skyhook might be involved in ordinary moral thinking, and ordinary moral life.

I contend that Dennett implicitly endorses what I have described as the thesis of anthropocentricity. Dennett’s ontology and epistemology vindicate (while
refining) much of our ordinary standards, sensitive as they are to human
cconcerns and interests, and ultimately grounded in the dispositional constitution
of our species, not in extra-human principles.

3.1.5. The targets of Dennett’s criticism

Dennett mostly criticizes theories and doctrines which, in his view, commit
the two related errors of essentialist thinking and skyhook postulation, notably in
regard to the mind, and of values (morality in particular). This conjures up (at
least one) illusory metaphysical gap, which leaves ill-explained how the mind and
mental phenomena, and values and morality, could fit within the world described
and explained by the physical sciences. In other words, they entrap themselves
within two-rung versions of the Cosmic Pyramid (today’s version of Cartesian
dualism), following which phenomena supposed to be “essentially human” are
isolated from the rest of nature by a mysterious metaphysical gap.

Interestingly, both those philosophers who are sympathetic to reductionism
(and naturalism) and those who are hostile to it are likely to commit that mistake.

He agrees with the critics of reductionism on the following point:
recognizing only the physical and the functional levels of description and
explanation as rationally acceptable can only deliver gravely impoverished (and
inadequate) accounts and worldviews. He condemns this overly restricted form of
reductionism as “greedy reductionism”.

256
Dennett suggests that the advocates of greedy reductionism typically pronounce the exclusion of non-referential families of terms (areas of discourse), but they often let skyhooks in their own “rationally acceptable language”. Thus, they deprive themselves of areas of discourse which, for being non-referential, are nevertheless indispensable; but (partly in compensation for this deprivation), they tend to favor realistic interpretation of terms (and items) that should not be so interpreted. Given their acceptance of essentialist thinking, their propensity to realism opens the door to skyhooks. For being overly restrictive on rationally acceptable levels of discourse they end up with impoverished worldviews; yet, for not being discriminative enough on the attribution of the “reality” status, the views of these philosophers are contaminated by skyhooks. As a result, their accounts of mind, meaning and morality are unsatisfactory.

Among the philosophers that belong to this category, Dennett notably targets those who defend what he calls “Cartesian materialism”, a theory which smuggles in (as we will see in section 3.4.2.2.) a certain conception of the self which Dennett regards as a major skyhook, and which I call the “Conscious Active Self”. This tends to entail in turn other skyhooks: “original intentionality”, “original understanding”, etc. On the other hand, it must be said that reductionists who smuggle in such skyhooks are usually critical of libertarian conceptions of free will (which often imply more metaphysically fantastic skyhooks).

Dennett also targets philosophers who assume that there is a fact of the matter about the propositional content entertained by people, and who propose to account for that propositional content with functionalist models of the mind’s
interaction with the environment (some of them also avail themselves to more or less Fregean construals of propositions, as abstract entities outside of time and space).

As for antireductionist philosophers, their accounts might not be impoverished in the way those of the reductionists are, but they are prone to use skyhooks. Moreover, as the antireductionists focus on the inadequacy of “greedy reductionism”, they miss out on the deeper aspirations that motivate reductionism, and do not tackle head on the requirement of providing credible answers to the placement question.

Thus, antireductionist philosophers tend to underestimate the philosophical relevance of empirical findings, and the importance of the principle that a theory which goes against well-confirmed empirical findings bears a burden of proof proportional to the centrality and degree of confirmation of the facts in question. As a result, their views of the mind and of morality typically feature skyhooks that reductionist philosophers keep at bay, e.g. libertarian conceptions of free will.

Among the opponents to reductionism, Dennett reserves his harshest criticism for those philosophers (usually hostile to naturalism) who claim to account for what they present as essentially human phenomena by invoking essential, metaphysically special, sui generis properties. The thought is that the metaphysical specialness of these properties is supposed to suffice to account for those phenomena which seem “above” simple physical phenomena; their
essential character suggests that they correspond to a distinct metaphysical realm, isolated from the rest of nature by a metaphysical gap; and their sui generis character implies that no explanation is required as to how they could fit in the world described by the physical sciences (attempts to do so are typically condemned as instances of the “naturalistic fallacy”). Dennett rejects this as a form of “mysterianism” (invoking essential properties and characterizing them as sui generis looks like a way of giving up on explaining) and indicts Colin McGinn for committing that sort of fault.

Thus, Dennett equally dismisses two traditional opponents: those who adhere to reductionism, and those who oppose it. For Dennett, they all fail to acknowledge the gradualist implications of evolutionary theory, and consequently to eliminate skyhooks.

I take Dennett’s notion of rational acceptability to capture the deepest aspirations and concerns of reductionism, while at the same time it avoids its traditional shortcomings. What matters for Dennett is to eliminate skyhooks. Science and its standards are powerful tools for that. But Dennett does not accord a “privileged status” to science. Scientific credibility is a mark of rational acceptability, but the latter is not identical to the former.
3.1.6. The intentional level of discourse

The intentional level of discourse, both in its ordinary and in it “rationally acceptable” form, is centrally concerned with two forms of intentional ascription. One consists in the immediate and mostly transient attribution of restricted packages of propositional attitudes to agents (not only persons, but also animals and even some machines), in order to make them readily predictable and intelligible. The other concerns mostly human beings, and consists in “recorded” attributions of propositional attitudes, and their assembling into bodies of propositional attitudes — what Dennett calls an individual’s “notional world”. An observer attributes notional worlds to familiar individuals, gradually, over an indefinite period of time. Dennett remarks that communal notional worlds can be attributed to communities.

Propositional attitudes are attributed along three modes: from oneself to others; from others to oneself; from oneself to oneself. Self-understanding must be construed, not as introspection (observation of the content one’s own consciousness), but as self-attribution of propositional attitudes.\(^6^2\)

\(^6^2\) When useful for purposes of simplifying the presentation and avoiding confusions, I will use the term “alter” as a substitute for the term “others”, and the term “ego” as a substitute for the term “oneself”. Thus I may use expressions such as “alter-to-ego attribution”, or “ego-to-ego attribution”, etc.
Paragraph 3.1.6.1. presents an overall portrait of ordinary intentional ascription, which I take to be congenial to Dennett’s thought, without being specific to him—with the exception of a few specifically Dennettian points that I flag as such. Most, but not all, of the other observations presented here are to be found in his writings, but they are not exclusive to Dennett.

Paragraph 3.1.6.2. presents the rationally acceptable version of intentional discourse, i.e. the theory of intentional system.

3.1.6.1. Ordinary intentional ascription

A large part of ordinary psychology (or “folk psychology”) concerns our ordinary ways of making minimally predictable and intelligible the behavior of other agents. This involves intentional ascription, that is, the attribution to agents of propositional attitudes, principally desires and beliefs.

According to Dennett, intentional ascription directly serves our capacity to anticipate (or “predict”) the behavior of agents, upon which its intelligibility depends. The same principle works in the reverse temporal direction: making retrospectively intelligible the behavior of agents is based on attributing to them desires and beliefs such that their behavior becomes retrodictable (i.e. it would have been predictable for an observer suitably located in time):
We use folk psychology — interpretation of each other as believers, wanters, intenders, and the like — to predict what people will do next. Predictions isn’t the only thing we care about, of course. Folk psychology helps us understand and empathize with others, organize our memories, interpret our emotions, and flavor our vision in a thousand ways, but at the heart of all these is the enormous predictive leverage of folk psychology. Without its predictive power, we could have no interpersonal projects or relations at all; human activity would be just so much Brownian motion; we would be baffling ciphers to each other and to ourselves — we could not even conceptualize our own flailings [sic]. In what follows [more broadly, in Dennett’s writings in general — P.C.] I will concentrate always on folk-psychological prediction, not because I make the mistake of ignoring all the other interests we have in people aside from making bets on what they will do next, but because I claim that our power to interpret the actions of others depends on our power — seldom explicitly exercised — to predict them. (1998 [1991], p.97-98)

Restricted packages of propositional attitudes are attributed in a transient, unrecorded way to agents whose conduct is currently being interpreted and anticipated. The episode of intentional ascription is typically of short duration, and aims at predicting immediately what the agent is going to do in the near future (“what the agent is going to do next”). The interpreter may be completely unknown to the observer, or may be an individual that the observer reidentifies, i.e. already “knows”, to some extent. In any case, the observer does not “hold for the record” the propositional attitudes that she attributes.

Considering the quasi-instantaneity of the performance, and the short-term range of the purpose, the interpreter might rapidly forget having attributed
particular propositional attitudes to a particular agent. In fact, considering the automatic, routinized character of the practice (the “mindless” naturalness of it), the interpreter might not even be aware of being engaged in it. Attributed propositional attitudes are not necessarily apprehended, and then forgotten: the interpreter may never have “noticed” them in the first place.

Again, Dennett insists on the dependence of intelligibility on predictability, but as time lapses predictions soon turn into ordinary explanations of past behavior. These are, to use philosophical characterizations, “belief-desire explanations”, or “reasons explanations”, or “rational-purposive explanations”.

The other form of intentional ascription corresponds to the case where the attributed propositional attitudes are held for the record, i.e. they are integrated to a general reconstruction of the interpreter’s body of beliefs and desires. We progressively attribute notional worlds to individuals we are familiar with, as we routinely keep open-ended “records” of the propositional attitudes that we attribute to them. Dennett calls “notional world psychology” the ordinary practice consisting in the progressive attribution of notional worlds.

Of course, an observer may attribute transient propositional attitudes to an agent of whose notional world she keeps a record, in which case the transient attribution is bound to be influenced by the content of the reconstructed notional world.
Attributing propositional attitudes partakes in the spontaneous, unthinking interpretation of the behavior of agents. In the case where we explicitly articulate such attributive judgments, we cannot avoid using an intentional vocabulary. Ordinary psychology can only be intentional.

The practice of intentional ascription involves certain rules, which we follow unthinkingly. We will see later how Dennett explicitly articulates three trans-cultural “rules of attribution” that we (unthinkingly) follow when we ascribe propositional attitudes to others. A debate has arisen about the organization of those rules and the cognitive form of intentional ascription. Some have argued that it has a theoretical form (a view known as “the theory theory”), while others have argued that it comes down to a form of reasoning by analogy with one’s own case (a view known as “simulation theory”). Those two options are not mutually exclusive, however: ordinary psychology might involve the combination of “theoretical” elements with the “simulation” form of reasoning. Although Dennett has not written much about this debate, his view of ordinary intentional ascription seems to be that it has a roughly theoretical form, which does not exclude simulation reasoning.63

---

63 As we will see in 3.4.1., there are well-supported hypotheses to the effect that we are endowed with an ordinary “theory” of interpretation and intentional ascription (somewhat misleadingly referred to as a “theory of mind”), determined by a species-wide, gene-based functional structure.
Arguably, ordinary intentional ascriptions meet with significant predictive success. This suggests that the rules that we unthinkingly follow as we attribute propositional attitudes are adequate; and it is consonant with the hypothesis that those rules are determined by a gene-based functional structure that passed through the sieve of natural selection.

Particular attributions of propositional attitudes that meet with predictive success, in turn, have a good chance of securing intersubjective agreement as to their correctness. This suggests that ordinary attributions of propositional attitudes are answerable to intersubjective standards of correctness, and that predictive success probably has a great importance in securing such agreement in the practice of giving and asking for reasons.

The scope of the transient, immediate form of intentional attribution is not restricted to humans: we also attribute beliefs and desires to animals, in order to anticipate their behavior and make it intelligible. Arguably, the same rules constrain and guide the attribution of propositional attitudes to humans and to animals. This requires that we ordinarily take animals to display minimal practical rationality (it is hardly disputable that they do if we acknowledge that we take them to behave according to what they “desire” and what they “believe”). However, the range of “desires” and “beliefs” that may be attributed to animals is highly restricted, and it is constrained by the characteristics of the species to which the observed animal belongs.
As we will see in the next paragraph (3.1.6.2.), Dennett insists on the fact that the scope of ordinary psychology goes beyond the human species.

Only those of our intentional ascriptions which are directed at human beings, however, get intertwined with moral judgments and moral sentiments. More exactly, as Strawson argues, it is when they are directed at responsible persons that our intentional ascriptions intermingle with moral judgments and moral sentiments. According to Dennett (as we will see in Chapter Four), this manifests the fact that, whereas intentional ascription is morally neutral and reaches beyond human beings, morality by contrast concerns human beings exclusively; and the notion of moral responsibility plays a fundamental role in morality generally. For Dennett, this difference is at the basis of the irreducibility of the personal/moral level of discourse (concerned with the attribution of personhood and of moral properties) to the intentional level (concerned with the attribution of propositional attitudes, which does not commit us morally).

According to Dennett, ordinary intentional discourse is non-referential, i.e. it does not refer to material mental items. Ascribing beliefs and desires to animals and machines does not commit one to any ontological statement. The meaningfulness of belief-desire attribution comes from the predictive power it gives in regard to the behavior of agents, which it consequently makes intelligible. Putting together these two features of intentional discourse (non-referentiality and predictive power), Dennett describes ordinary intentional discourse as a heuristic
overlay, indispensable for anticipating and making sense of the behavior of agents.

3.1.6.2. A rationally acceptable intentional theory: the theory of intentional systems

According to Dennett, the predictive success of ordinary intentional psychology reflects its overall adequacy. Moreover, it does not directly involve essentialist thinking, and does not directly imply skyhooks. By and large, it is acceptable as it is. Consequently, it serves as a “source theory” for the version of it that constitutes the intentional level of the rationally acceptable language. To be precise, there are two rationally acceptable theories, corresponding to the two forms of intentional ascription. Dennett accords more importance to the “theory of intentional systems”, which derives from the immediate, transient form of intentional ascription, than to the rationally acceptable version of notional world psychology (the “recorded” form of intentional ascription).

The “theory of intentional systems” mostly consists in an explicitation of the principles of ordinary intentional ascription.

According to Dennett, there is simply no fact of the matter about the propositional content associated to intentional states. The latter may really be about an identifiable feature of the environment, but any propositional content
that we may associate to them can only be an indeterminate interpretation. This, he claims, is a case of indeterminacy in the sense of Quine’s thesis of the indeterminacy of radical translation. In the context of Dennett’s doctrine, we might say that interpretation is radically indeterminate.

Although there is no fact of the matter about the propositional content “borne” by a given intentional state, our attributions of propositional attitudes are answerable to the standard of correctness that Dennett designates as the most important: predictive success. In this case, Dennett claims that predictive success indicates that we correctly capture certain patterns in the agent’s behavior — not any real, determinate propositional content supposedly borne by any intentional state inside the agent.

The scope of the theory of intentional systems is the same as that of ordinary intentional ascription. It includes humans, animals and some machines, notably computers. The principle which determines the scope of the theory, is that we attribute propositional attitudes to all agents whose behavior it makes easier to predict.

Dennett calls “intentional systems” all those agents (persons, animals, machines…) whose behavior can only be predicted by attributing propositional attitudes to them, or becomes at least easier to predict when we engage in such attribution. The category of “intentional systems”, then, is not defined by firm
criteria. It is not something that agents essentially are (or are not), but something that an observer takes them as, for practical reasons.

The degree of predictive power afforded by intentional attribution varies according to the complexity of the agent's behavior. As the complexity of behavior increases, the predictive leverage provided by propositional attitude attribution becomes more significant. At the lower end of the complexity scale, propositional attitudes can in principle be attributed to sub-organic, self-duplicating macro-molecules. They can also be attributed to contraptions as simple as a thermostat. However, the predictive power thus gained is of little significance. However, significant predictive leverage appears surprisingly early along the complexity scale. As we get to human beings, the attribution of propositional attitudes is required to make their conduct predictable and intelligible.

The theory that Dennett proposes to build from ordinary psychology is a theory of intentional systems in general. Dennett calls “intentional stance” the level of discourse corresponding to the use of intentional vocabulary. Thus, the theory of intentional systems is articulated at the level of the intentional stance. (Somewhat confusingly, Dennett sometimes uses the expression “intentional stance” to designate the ordinary practice, at other times he uses it to designate his theory of intentional systems.

The theory of intentional systems, then, formulates three rules, which ordinary agents follow in their ordinary practice of intentional ascription. One concerns the ascription of beliefs, another concerns the ascription of desires, the
third concerns the rationality of the agents to whom beliefs and desires are being ascribed. How Dennett formulates them need not concern us, however. What matters is that we have established that we can offer, in the rationally acceptable language, rational-purposive explanations of the behavior of intentional systems, which make it predictable and intelligible. These rational-purposive explanations are skyhook-free, and partake in a gradualist picture of the world.

3.1.7. The functional level of discourse

I mentioned in 3.1.4. that the philosophical doctrine of functionalism, as used in the philosophy of mind and in the cognitive sciences, partook in the rationally acceptable use of the functional vocabulary. What rationally acceptable use Dennett makes of the functional vocabulary — what theory of mind he proposes — will appear progressively throughout the following sections of this chapter (3.2., 3.3. and, last but not least, 3.4.).

We only need to clarify a few points in order to understand the background against which Dennett deploys his theory of the mind.

From the mid 1960’s on, the idea has been propounded, against behaviorism, that mental phenomena, and behavior, could be explained by functional mental systems, posited “inside” the human mind. This idea, which is the central insight of functionalism, has developed between two doctrinal and
disciplinary poles: that of research in artificial intelligence (particularly, so-called “Good Old-Fashioned Artificial Intelligence”, or GOFAI), and that of the neurosciences. In a very schematic way, the difference between these two poles could be summarized as follows.

GOFAI is characterized by the predominance of a radically top-down approach, which abstracts almost completely from issues concerning the realizability in nervous systems of posited functional units. In other words, researchers build models which involve the positing of functional structures (or “systems”), but they are not too concerned about how those functional structures could be physically realized in organisms.

In the neurosciences, researchers stick to an empirical, bottom-up approach, and question the legitimacy of positing functionalist units. That is, they engage in close empirical study of actual neurophysiological structures in the central nervous systems of various species (but mostly in the human brain), and prefer to avoid the postulation of “functional structures”.

Again, this polarization is very schematic. Most researchers in cognitive sciences, and most philosophers of the mind, have steered a middle course between those two poles.

These are thinkers who admit that the top-down and bottom-up approaches are complementary. On the one hand, the former is heuristically indispensable in order to account for a broad range of mental phenomena and for behavior. On the other hand, the latter imposes, as an adequacy condition for
functionalist models, that the posited functional structures be physically realizable.

The “pandemonium model of mental activity” was an early model, close to the neurosciences, but open to the possibilities of functionalism. Dennett describes it as a “pioneer model”: «The pioneer model […] is Oliver Selfridge’s (1959) early Pandemonium architecture in Artificial Intelligence, in which many “demons” vied in parallel for hegemony» (1991, p.189). The cardinal idea of the model is that cognitive activity involves parallel distributed processing (“PDP”) of information, through neuronal networks that “care about” various task-related domains (color, location, motion, etc.) (1991, p.262) As Dennett puts it, «all varieties of thought or mental activity are accomplished in the brain by parallel, multitrack processes of interpretation and elaboration of sensory inputs.» (1991, p.111) As we will see in 3.4.2., this model has exerted a strong influence on Dennett.

Now, the middle-course approach, attentive as it is to the empirical investigation of neuronal networks, does not shy away from a form of theorizing which involves the positing of nested hierarchies of functional structures. Because each structure enjoys a sufficient degree of autonomy to be treated like a homunculus, this sort of theorizing has been called “homuncular functionalism”. Again, it acknowledges the significance of the realizability constraints imposed by the configuration of the actual studied by neuroscientists.
Dennett has championed homuncular functionalism, because it reveals the *gradualist* character of cognitive activity, both as regards its functionality, and as regards its physical realization in neurophysiological structures. That is, homuncular functionalism presents us with «a finite regress that peters out without marked foundations or thresholds or essences.» (1998 [1994], p.362). Thus, homuncular functionalism articulates «the idea that the way to explain the miraculous-seeming powers of an intelligent intentional system is to decompose it into hierarchically structured teams of ever more stupid intentional systems, ultimately discharging all intelligence-debts in a fabric of stupid mechanisms.» (ibid)

As we will see in 3.4.2., Dennett’s theory of the mind, then, combines the empirically rich pandemonium model with heuristically rich, and gradualist, homuncular functionalism, to deliver a centerless picture of the mind, in which mental activity is radically distributed across the nested hierarchies of functional systems, realized in multiple neuronal networks.

Now, Dennett places this theory of the mind in an evolutionary perspective. That is, increasingly impressive “kinds of minds” have emerged throughout evolutionary history, endowed with increasingly complex functional designs and increasingly powerful capacities, accomplishing ever more complex tasks, producing ever more complex mental phenomena. We have to see how this went in order to obtain a skyhook-free picture of the human mind.
3.2. EVOLUTION: FROM SELF-REPLICATORS TO PRIMATES

Dennett proposes (3.2.1.) an evolutionary, and therefore gradualist, anti-essentialist, skyhook-free conception of a series of fundamental phenomena, including agency, interests, reasons, normativity, intentionality. These are evolutionary phenomena which gradually come into being. Although the use of language entails a quantum leap in their evolution, these phenomena are not “essentially human” — only their specifically linguistic version is.

Dennett proposes (3.2.2.) an evolutionary classification of the primitive ancestors of the human species into three progressively more intelligent categories, endowed with increasingly powerful cognitive functional designs, implemented in increasingly complex nervous systems. Tracing this evolution is important for the reason that, in virtue of the “principle of accumulation of design”, the more evolved species typically inherit of the primitive functional structures of their (less evolved) evolutionary ancestors, over which new functional structures are added. As a result, the design of the human species — our evolutionary heritage — is constituted of distinct layers that have their origins in all three pre-human categories, a minimal examination of which is thus necessary.

A significant part of Dennett’s work is dedicated to drawing some philosophical implications from recent work in cognitive ethology (3.2.3.). It appears that many animal species manifest remarkably advanced cognitive abilities, such as: the possession of “proto-concepts”; the ability to entertain
intentional states of the second order, and to use them in tactical behavior; and finally, the ability to hold “track records” of reidentified individuals”, such that in certain higher species of social animals (living in groups), each group member has a “reputation” in the eyes of the other members of the same group. These findings help us in appraising what range of cognitive abilities we can reasonably attribute to our immediate ancestor, prelinguistic Homo sapiens. As the natural extension of the history of intelligent agency, this brief excursion out in the field of cognitive ethology contributes in establishing the gradualist genealogy of our own human agency, and of human nature in general.

3.2.1. The evolution of some phenomena

Agency, according to Dennett, is not something for which we can provide a criterial definition, with sufficient and necessary conditions. Roughly speaking, agency is “autonomous goal-oriented activity”, the most elementary form of which is the self-replication performed by those large molecules which, like DNA and the more primitive RNA, have the ability «to mindlessly construct and then shed exact — or nearly exact — copies of themselves.» (1996, p.20) The most primitive agents, then, are self-replicating macromolecules, «the first macromolecules that have enough complexity to perform actions, instead of just lying there having effects.» (ibid; emphasis added)
Arguably, self-replicating macromolecules are a borderline case of agency, but Dennett’s point is that they constitute the starting point of a series of closely related phenomena, namely the “biological self”, “interests”, “goals”, “reasons”, “practical rationality”, “normativity”, “function”, “intentionality”…

Agency implies a biological individual, distinguished from everything on the “outside” (everything that is not itself) by boundaries of some sort. The expression *biological self*, in a concrete sense, designates the biological individual. In an abstract sense, it designates the principle of biological organization by which the individual is constituted.

As soon as something gets into the business of self-preservation, boundaries become important. […] this distinction between everything on the inside of a closed boundary and everything in the external world is at the heart of all biological processes, not just ingestion and excretion, respiration and transpiration. (1991, p.174)

In the following quote, Dennett ties “the birth of the biological self” to “the birth of reasons”, as in “reasons for performing certain actions”:

[…] the birth of reasons was also the birth of boundaries, the boundary between “me” and “the rest of the world”, a distinction that even the lowliest amoeba must make, in its blind, unknowing way. This minimal proclivity to distinguish self from other in order to protect oneself is the biological self, and even such a simple self is not a concrete thing but just an abstraction, a principle of organization. (1991, p.414)

64 Like propositional attitudes, the biological self is attributable (in principle) to vegetals.
In order to understand Dennett’s point, it must be clear that all biological individuals, all members of all species, have certain goals and interests, the ultimate goals and interests of all biological agents being self-preservation and reproduction. The crucial point that Dennett is making is that the individual need not be aware that it has certain goals and certain interests in order to have them. The same point applies to reasons. From the moment that an individual’s action is oriented towards certain goals, this individual (this “biological self”) has reasons for performing the actions it performs, even though it is not aware of them. The birth of agency, then, is by the same token the birth of goals, and so of interests and of reasons (for acting):

In the beginning, there were no reasons; there were only causes. Nothing had a purpose, nothing had so much of a function; there was no teleology in the world at all. The explanation for this is simple: there was nothing that had interests. But after millennia there happened to emerge simple replicators, and while they had no inkling of their interests, and perhaps properly speaking had no interests, we, peering back from our Godlike vantage point at their early days, can nonarbitrarily assign them certain interests — generated by their defining “interest” in self-replication. That is, [...] we can assign them interests conditionally. If these simple replicators are to survive and replicate, thus persisting in the face of increasing entropy, their environment must meet certain conditions: conducive to replication must be present or at least frequent. (1984, p.21)
The birth of goals, interests and reasons is, by the same token, the birth of practical (instrumental) rationality: agents that behave according to their interests have good reasons for doing what they do. Therefore, they exhibit a natural, primitive form of practical rationality.

The birth of interests (implied by the birth of agency, and of goals) implies the birth of normativity. It is also linked to the birth of function and of intentionality. Let us take these points in turn.

The birth of goals and of interests implies the difference “right” and “wrong”, “better” and “worse”. That is, it implies that there is a difference between “getting things right” and “getting things wrong”; and between things that are objectively “good” in regard to an individual’s self-preservation and reproduction (and therefore to be sought for), and things that are objectively “bad” in regard to the same things (and therefore to be avoided):

When an entity arrives on the scene capable of behavior that staves off, however primitively, its own dissolution and decomposition, it brings with it into the world its “good”. That is to say, it creates a point of view from which the world’s events can be roughly partitioned into the favorable, the unfavorable, and the neutral. And its own innate proclivities to seek the first, shun the second, and ignore the third contribute essentially to the definition of the three classes. (1991, p.174)

Simply put, the interests of an agent are interests in doing what is good for it and avoiding what is bad for it. Now, goal-oriented activity implies the possibility of error: «Right from the beginning, the cost of doing something is running the risk of doing it wrong, of making a mistake. Our slogan could be: No taking
This, Dennett claims, applies to the most elementary, primitive form of activity, i.e. the self-replication of macromolecules. As mentioned in an earlier quote, self-replicators «construct and then shed exact — or nearly exact — copies of themselves.» This is (was) the original error:

The first error that ever was made was a typographical error, a copying mistake that then became the opportunity for creating a new task environment [...] with a new criterion of right and wrong, better and worse. A copying error “counts” as an error here only because there is a cost to getting it wrong: termination of the reproductive line at worst, or a diminution in the capacity to replicate. These are all objective matters, differences that are there whether or not we look at them, or care about them, but they bring in their train a new perspective. (1995, p.203)

This, then, is natural, non-linguistic normativity. The standard of correctness to which the responses (cognitive, emotional and behavioral) of higher animals will be answerable is strictly pragmatic: it comes down to the distinction between success or failure in the satisfaction of the interests (needs, goals) that they serve (as we will see, this changes with the development of language and rational criticism).

Goal-oriented activity, then, is answerable to pragmatic standards of correctness: correct action is that which succeeds in regard to the goals it is directed at; incorrect action is that which fails in the same regard.

The birth of interests brought about the birth of function and of intentionality:
Even the simplest organisms, if they are to favor what is good for them, need some sense organs or discriminative powers — some simple switches that turn ON in the presence of good and OFF in its absence — and these switches, or transducers, must be united to the right bodily responses. This requirement is the birth of function. (1996, p.32)

Note that switches that turn on or off exhibit a primitive form of intentionality, or aboutness. The following quote expresses the fact that Dennett considers intentionality as a much more primitive phenomenon than consciousness. It also shows how Dennett’s gradualist construal of intentionality opposes the essentialist construal, according to which only human beings display genuine, original intentionality:

There was a time, before life on earth, when there was neither intentionality nor consciousness, but eventually replication got under way and simple organisms emerged. Suppose we ask of them: Were they conscious? Did their states exhibit intentionality? It all depends on what these key terms are taken to mean, of course, but underneath the strategic decisions one might make about pre-emptive definition of terms lies a fundamental difference of outlook. One family of intuitions is comfortable declaring that while these earliest ancestors were unconscious automata, not metaphysically different from thermostats or simple robotic toys, some of their states were nevertheless semantically evaluable. […] According to this vision, then, the intentionality of our unconscious ancestors was as real as intentionality ever gets; it was just rudimentary. (1998 [1994], p.358)
The most rudimentary form of intentionality is that which is exhibited by a lock and the key that fits in it. Molecular receptors, and the molecules which they are designed to accept, exhibit that primitive form of aboutness. For example, opioid receptors in brain cells are «receptors designed to accept the endorphin molecules that nature has been providing in brain for millions of years.» (1996, p.35). The former can be said to be “about” the latter in the same sense that a lock can be said to be “about” the key that fits in it. This is the most primitive form of aboutness, and the most primitive form of function.

This lock-and-key variety of crude aboutness is the basic design element out of which nature has fashioned the fancier sorts of subsystems that may more deservedly be called representation systems, so we will have to analyze the aboutness of these representations in terms of the quasi (?) aboutness of locks-and-keys in any case. (ibid)

Thus, Dennett’s gradualist construal of the basic phenomena reviewed here is, like the Darwinian theory of which it purports to be an extension, strongly opposed to conceptions of rationality, intentionality, reasons (for action) and normativity according to which these phenomena have an essence (or something like it) that is specifically human.

3.2.2. The evolution of intelligent agency
Dennett classifies the primitive ancestors of the human species into three categories. From the most primitive to the more evolved, they are: Darwinian creatures, Skinnerian creatures, and Popperian creatures. I present here brief summaries of Dennett’s descriptions of them. Each is a step that takes us towards Dennett’s account of human practical reasoning, decision, agency, and free will. His portrait of the evolution of intelligent agency is part of his gradualist, anti-essentialist, skyhook-free picture of human nature.

3.2.2.1. Darwinian creatures

“Darwinian creatures” are the most primitive sort of animals. They cannot learn anything, so their phenotype is rigid, i.e. entirely programmed by their genotype. More precisely, the dispositions they are endowed with are determined by certain hard-wired mechanisms, themselves determined their species’ genotype. As a result, their behavior is tropistic, unchanging, and uniform across all conspecifics. That is, all members of a species of the “Darwinian” type exhibit the same characteristic “fixed action patterns”, aimed at the same goals and subgoals.

Dennett christened them “Darwinian creatures”, because they exemplify the crudest expression of natural selection: genotypical variations between individuals were generated by random mutation, and the best-adapted genotypes
reproduced at accelerated pace from generation to generation, pushing out to extinction the genotypical variations that showed suboptimal adaptability.

Dennett’s prime example of a Darwinian creature is the digger wasp, *Sphex ichneumoneus*. He mentions that the characteristic fixed action patterns of the sphex have been vividly described by D. Wooldridge, and that Douglas Hofstadter used it an illustration of certain fears that many persons entertain about the human condition, namely the fear that human life might be nothing but a particularly complex version of the sphex’s antics; the fear, in other words, that the human condition may display a property, as Hofstadter called it, of “sphexishness”. If human cognitive activity and human conduct are ultimately determined by the laws of physics; if our condition is not distinguished from that of sphex by something metaphysical (like some essences), does it really add up to anything more than a sophisticated version of the condition of the digger wasp? It is of genuine philosophical relevance that the fear of determinism (which Strawson evokes, but rejects, in “Freedom and Resentment”), as well as the absurdist interpretations of Nietzsche’s idea of the “eternal return”, find a graphic illustration in the condition of sphex.

Dennett’s point is precisely to show that morally, we are indeed extraordinarily different from the sphex, even though what separates us from it is nothing metaphysical, but simply the gradual increase in complexity of our capacities and of our possibilities.
3.2.2.2. Skinnerian creatures

“Skinnerian creatures” mark the first step up. They are endowed with phenotypic plasticity, although of the most primitive sort. That is, members of “Skinnerian” species are not wholly designed at birth: they have the capacity to learn by trial-and-error. As a result, their behavior exhibits *conditionable* plasticity. When confronted with an unfamiliar situation, a Skinnerian creature will produce any behavioral response, more or less at random. “Good” results are positively reinforced, while “bad” results are negatively reinforced. This is how it learns.

The fact that Skinnerian creatures enjoy a minimal range of behavioral elbow room implies that they are confronted (as Darwinian creatures are not) with “the problem of what to do next” i.e. the problem of expressing, within a narrow range of possibility, the behavioral response that is best adapted to perceived circumstances. Their solution to this problem relies on memory of past experience: a Skinnerian agent has to keep track not only of what is going on, but also of the outcomes of its own past performances in certain circumstances, so as to be able to reproduce the right performances in the right circumstances. However, it manifests no sense of anticipation. As Dennett puts it, «Skinnerian conditioning is a fine capacity to have, so long as you are not killed by one of your early errors.» (1995, p.374)
Dennett calls them “Skinnerian creatures”, because the only form of learning that they exhibit is based on reinforcement, which is the form of learning championed by the head figure of behaviorism, B.F. Skinner.

Few species belong to that category, as most species seem to fall either in the lower category of Darwinian creatures, or in the higher category of “Popperian creatures”. Dennett suggests that the sea slug *Aplysia californica* could be an example of the Skinnerian category.

3.2.2.3. Popperian creatures

The second step up is marked by “Popperian creatures”, which are endowed with a much richer form of phenotypic plasticity than that of Skinnerian creatures. Dennett models their competence as follows.

Like Skinnerian creatures, Popperian creatures learn from experience, by trial-and-error. But over the structures that make possible this empirical learning (based on the memory of past experience) is added a cognitive structure that Skinnerian creatures lack, and which performs two complementary tasks: the *generation* of multiple options of response (say, behavioral responses); and the *selection* of one option, upon which the agent actually responds.

According to Dennett, the selection is effected out of the virtual “testing” of all candidate options within an “inner”, virtual environment. A multitude of
candidate options that have never been actually tried (some of which may be “tempting”) are rejected, to the profit of one candidate option that may never have been actually tried, but which is nevertheless better-than-chance (in fact, it stands a good chance of being the most appropriate in the circumstances). That is because, in spite of never having been actually tried, the candidate options have been virtually tested in the creature’s “inner environment”, which Dennett describes as (note the functionalist description) «an inner something-or-other that is structured in such a way that the surrogate actions it favors are more often than not the very actions the real world would also bless, if they were actually performed.» (1995, p.375).

Somewhat more precisely, Dennett compares the testing carried in the inner environment of the creature to a falsification procedure, i.e. the rejected options are like falsified hypotheses, and the selected option is the one that resists falsification. Karl Popper argued that, thanks to science, we humans can guide our action on knowledge, whose acquisition and growth is based on the falsification of hypotheses. This, he claimed, “permits our hypotheses do die in our stead”. According to Dennett, an unconscious and automatic version of the falsification process goes on inside the brain of those creatures which he characterizes, consequently, as “Popperian”.

The virtual testing in the inner environment confers to Popperian creatures a capacity for anticipation. We saw that Skinnerian creatures, for having minimal elbow room as regards their behavioral options, were confronted to “the problem of what to do next”. Popperian creatures, whose capacity for pre-selection of
behavioral options give them much more elbow room, for their part, are confronted to “the problem of what to think about next”: 

[…] nervous systems solved the “Now what do I do?” problem by a relatively simple balancing act between a strictly limited repertoire of actions — if not the famous four F’s (fight, flee, feed, or mate), then a modest elaboration of them. But now, with increased functional plasticity, and increased availability of “centralized” information from all the various specialists, the problem of what to do next spawn a meta-problem: what to think about next. (1991, p.188)

According to Dennett, the development of brains is an answer to the problem of anticipation: «The key to control is the ability to track or even anticipate the important features of the environment, so all brains are, in essence, anticipation machines.» (1991, p.177)

The difference between the respective cognitive functional designs of Skinnerian and Popperian creatures corresponds to the distinction proposed by cognitive scientist Gary Drescher’s between “situation-action machines” and “choice machines”. Dennett describes the distinction as follows:

Situation-action machines are a collection of relatively simple switches, each one embodying an environmental rule of sorts: If you encounter condition C, do A. They are cost-effective for relatively simple organisms whose behavior is innately specified. Choice machines have a different set of mechanisms, which embody predictions: If you encounter condition C, doing A would result in outcome Z (with probability p). They generate several or many such
predictions, and then evaluate them (using whatever values they have, or have developed), and this arrangement is cost-effective for organisms that are designed to learn during their own lifetimes. An organism can have both sorts of machines in its kit, relying on the former for quick-and-dirty life-saving choices and relying on the latter for serious thinking about the future — a rudimentary faculty of practical reasoning. (2003, p.163)

The design of Popperian creatures does include both sorts of machines, in addition to primitive Darwinian mechanisms. In virtue of the principle of accumulation of design, the “choice machines” that characterize Popperian creatures are added to the primitive situation-action machines that they inherited from their Skinnerian ancestors, which are themselves added more primitive mechanisms that characterize Darwinian creatures. Popperian creatures are still describable as being endowed with dispositions: some are underlain by choice machines, other by situation-action machines, others by Darwinian mechanisms.

3.2.2.4. The evolution of practical rationality

According to Dennett, Darwinian creatures, Skinnerian creatures and Popperian creatures manifest a gradual deployment of practical rationality.

What practical rationality Darwinian creatures manifest, however, stems entirely and directly from the free-floating rationale of their species. All members of a particular Darwinian species uniformly manifest the fixed action patterns of
their species, as they pursue a fixed hierarchy of goals and subgoals which is
that of the species, down to the most immediate and partial subgoal. The reasons
why they do what they do are uniformly identical to the overall, free-floating
rationale of their species, which connects their fixed behavioral routines to their
fixed hierarchies of goals and subgoals in a way that is practically rational, thanks
to natural selection. The practical rationality of a Darwinian individual is the rigid
rationality of its species’ free-floating rationale.

Skinnerian creatures have a narrow margin of “elbow room”. A Skinnerian
individual pursues the higher goals that are fixed by its species, but as it learns
“good tricks”, by blind essay-and error, it develops mildly original (individualistic)
routines, oriented towards specific subgoals of its own. For example, it may have
learned effective tricks for hunting, involving very specific subgoals, which are
unknown of other members of the same species. There is very little individuality
in the behavior of a Skinnerian creature. The reasons why they do what they do,
and the rationality they manifest in doing it, are largely the reasons, and the
rationality, of their species. Nevertheless, a small fraction of these reasons, and
of this rationality, is partly that of the individual, as it has learned to do certain
things that other members of the same species do not necessarily learn.

Popperian creatures have learning capacities that connect with their
capacity to consider different candidate courses of action in advance, and to
select one candidate in a non-random, better-than-chance way, thus manifesting
a primitive form of “decision-making”. Popperian creatures develop behavioral
patterns that are more largely individual that what Skinnerian creatures can
afford, aimed at a more largely individual range of subgoals. The reasons why they do what they do, and the rationality they manifest in doing it, is more largely their own than is the case with Skinnerian creatures.

In sum, the transition from Darwinian to Skinnerian to Popperian creatures involves an increase in the “individualization” of the reasons agents have for doing what they do, and of the rationality they manifest. Darwinian creatures, with their rigid, tropistic behavior, do what they do for reasons that compose a fixed free-floating rationale, which is that of their particular species. By contrast, individual members of species of the Popperian sort, equipped as they are with choice machines, do what they do for reasons that are, at least partly, their own: «[…] the individual creation [i.e. the individual agent — P.C.] generates its own reasons for doing x or y, by anticipating probable outcomes of various candidate actions and evaluating them in terms of the goals it also represents (since these goals can change over time, in response to new information gathered).» (2003, p.46) Dennett, as always, insists on the gradual character of the transition from <reasons attributed to the species>, to <reasons attributed to the individual agent>: «we may find that there is a seamless blend of intermediate steps, with more and more of the design work off-loaded from [virtual] designer to designed agent.» (ibid)

As we will see, the advent of language and of rational criticism marks a quantum leap in the development of rationality. Human rationality is not only
practical, it is also theoretical. But there is no difference of essence between the rationality displayed by humans and that displayed by animals. Only a gradual development, which follows something like an exponential curve of development (where the advent of language and rational criticism marks the point where the curve “takes off”).

3.2.3. Cognitive abilities in animals

Recent work in cognitive ethology suggests that animals can exhibit surprisingly advanced cognitive capacities. Such findings support Dennett’s gradualist, anti-essentialist view, and support (as we will see in the next section) the attribution to prelinguistic Homo sapiens of a host of cognitive capacities which started the “arms race” that gave rise to proto-morality. In other words, cognitive ethology helps us trace the genealogy, and hence the gradual deployment, of human cognitive capacities.

Dennett extracts from cognitive ethology three particularly significant results. One consists in pre-linguistic concepts, which I call proto-concepts (3.2.3.1). Another is tactical behavior, and the more specific ability it seems to imply, viz. the capacity to entertain intentional states of the second order (3.2.3.2.); moreover, certain species seem to have a certain ability to make a distinction between “getting it right” and “getting it wrong”. Finally, there is the ability to hold “track records” of reidentified individuals, such that in certain higher
species of social animals (living in groups), each group member has a
"reputation" in the eyes of the other members of the same group (3.2.3.3.).

3.2.3.1. Proto-concepts

"Proto-concepts" are non-linguistic, unconscious, and can be attributed to animals as primitive as a bee. Dennett approvingly refers to psychologist Julian Jaynes’ claim that “The bee has a concept of the flower”, and adds the following (in characteristic casual but insightful style):

We have a very salient theoretical role for something which we might as well call concepts, but if you don’t like it we can call them schmoncepts, concept-like things that you don’t have to be conscious to have. For instance, computers have them. They [the computers] are not conscious — yet — but they have lots of concepts […] (1998 [1986], p.128-129)

Somewhat more elaborately, Dennett specifies in the following quote that the difference between non-linguistic, unconscious concepts (what I am calling proto-concepts) and linguistic concepts is not extensional; rather, it concerns this crucial fact that linguistic concepts (occurring in judgments) partake in a reflective, critical articulation of one’s cognitive activity — what we call thinking:

Does a dog have a concept of a cat? Yes and no. No matter how close a dog’s “concept” of cat is to yours extensionally (you and the dog discriminate the same sets of entities as cats and noncats), it
differs radically in one way: the dog cannot consider its concept. It cannot ask itself whether it knows what cats are; it cannot wonder whether cats are animals; it cannot attempt to distinguish the essence of cat (by its lights) from the mere accidents. Concepts are not things in the dog’s world in the way that cats are. Concepts are things in our world because we have language. A polar bear is competent vis-à-vis snow in many ways that a lion is not, so in one sense a polar bear has a concept that the lion lacks — a concept of snow. But no languageless mammal can have the concept of snow in the way we can, because a languageless mammal has no way of considering snow “in general” or “in itself”. (1996, p.159)

3.2.3.2. Tactical behavior: second-order intentionality and the distinction between “getting it right” and “getting it wrong”

Certain animal species, in specific circumstances involving the presence of potential predators, engage in what appears as a tactical sort of behavior. This means that the description and explanation of their behavior is bound to involve intentionality of the second order. That is, we describe certain animals as entertaining intentional states, which bear on the intentional states of other agents. Simply put: certain animals seem to have the desire to induce certain beliefs in other animals that they see as potential predators. As we describe the behavior of these animals, we invest these desires and belief with propositional contents (not to be realistically interpreted).
Dennett gives a few examples of such tactical behavior, involving intentionality of the second order. Hares dissuade potential predators (e.g. foxes) from chasing them by displaying their capacity to escape them: the hare conspicuously stands up on its hind legs, as if it entertained the desire to induce in the fox the belief that it (the hare), conspicuously confident as it is, is too far and runs too fast to be caught by the fox (cf 1996, p.123). Gazelles use a similar tactic with their own predators:

Gazelles being chased by lions of hyenas often do something similar, called stotting. They make ridiculously high leaps, obviously of no benefit to their flight but designed to advertise their superior speed to the predators. “Don’t bother chasing me. Chase my cousin. I’m so fast I can waste time and energy doing these silly leaps and still outrun you”. And it apparently works; predators typically turn their attention to other animals. (1996, p.123)

The case of low-nesting bird is particularly interesting, as the sort of tactical behavior they engage in (which, again, involves second-order intentionality) is clearly deceptive.

In order to protect their eggs from predators, low-nesting birds desire to induce a wrong belief in them. The following quote begins with a typical Dennettian reminder: cognitive capacities are attributed to animals on the basis of observable behavior, and are to be construed in a gradualist perspective:

But if higher-order intentionality is, as I and others have argued, an important advance in kinds of minds, it is not as clearly the watershed we are looking for between thinking and unthinking cleverness. Some of the best-studied examples of (apparent) higher-order intentionality among nonhuman creatures still seem to fall on the side of unreflective adroitness. Consider “distraction display”, the well-known behavior of low-nesting birds who, when a predator approaches the nest, move surreptitiously away from their vulnerable eggs or nestlings and begin in the most ostentious way to feign a broken wing, fluttering and collapsing and calling out most piteously. This typically leads the predator away from the nest on a wild goose chase, in which it never quite catches the “easy” dinner it is offered. The free-floating rationale of this behavior is clear [...] (1996, p.121-122)

The main point here is that it is practically impossible to describe and explain the behavior of these animals, in a way that enables us to capture their patterns and to anticipate them, without attributing to them intentional states of the second order, which we invest with propositional content.

Accounting for tactical behavior, then, apparently requires that we attribute second-order intentionality to certain animals. However, coming down to first-order intentionality is not a priori excluded. It is not excluded either that ethologists might eventually observe behavioral patterns whose description and explanation will require third-order intentionality. But for the time being, it is

---

66 Consider the case of bees removing dead conspecifics from the hive. That behavior is most rigorously described in physicalist terms, as a tropistic pattern: dead bees secreted oleic acid, whose detection by the other bees triggers a “remove it” mechanistic routine in them.
second-order intentionality that has imposed itself as a practically indispensable component of the description of the behavioral patterns of certain animal species.

In any case, Dennett insists that animal intentionality, of any order, originates in free-floating rationales honed by natural selection, and does not require its conscious appreciation by animals who display it. In the following quote, Dennett refers to the deceptive behavior of low-nesting birds, but he is making a very general point about the fact that in accounts of animal behavior, free-floating rationales play a more important theoretical role than whatever “consciuosness” animals might or might not have:

[…] the raison d’être of this instinctual behavior is its deceptive power. That’s why it evolved. If we want to know why this strange dance came to be provokable on just these occasions, its power to deceive predators will have to be distilled from all the myriad of other facts, known and unknown, in the long ancestry of the species. But who appreciated this power, who recognized this rationale, if not the bird or its individual ancestors? Who else but Mother Nature herself? That is to say: nobody. Evolution by natural selection “chose” this design for this “reason”. (1987 [1983] p.259).

However, although he downplays the role of what an individual animal may “appreciate”, Dennett does not dismiss the very idea of “appreciation” as a skyhook. This appears in relation with a cognitive capacity that the case of deceptive behavior brings us to: the capacity to distinguish between “getting it right” and “getting it wrong”. What appreciation do animals have of this distinction? Although the previous quote suggests that Dennett’s answer would
be “none”, he has recently qualified this position. The following quote insists on the contrast between the appreciation that we, reflectively rational, language-using human beings have of the reality/appearance distinction, and what dim appreciation of it certain animal species might have. But as we will see in the next section, early Homo sapiens, poised in between, had just enough appreciation of this contrast for a cognitive arms race between deceptive tactics and anti-deception measures to be initiated.

At this very moment, trillions of organisms on this planet are engaged in a game of hide-and-seek. But for them it’s not just a game; it’s a matter of life and death. Getting it right, not making mistakes, matters to them — indeed nothing matters more — but they don’t as a rule, appreciate this. They are the beneficiaries of equipment exquisitely designed to get what matters right, but when their equipment malfunctions and gets matters wrong, they have no resources, as a rule, for noticing this, let alone deploring it. They soldier on, unwittingly. The difference between how things seem and how things really are is just as fatal a gap for them as it can be for us, but they are largely oblivious to it. The recognition of the difference between appearance and reality is a human discovery. A few other species — some primates, some cetaceans, maybe even some birds — show signs of appreciating the phenomenon of “false belief” — getting it wrong. They exhibit sensitivity to the errors of others, and perhaps even some sensitivity to their own errors as errors, but they lack the capacity for the reflection required to dwell on this possibility [...] (2003, p.165)

It is worth mentioning that the sort of functional structure that underlies such capacities as that of engaging in higher-order intentionality need not be a
complex “choice machine”. Dennett mentions that the capacity to engage in deceptive behavior, and so to engage in higher-order intentionality, may be realized, at the neurophysiological level, by hard-wired circuits of the sort that equips Darwinian creatures, «abetted here and there by Skinnerian mechanism» (1996, p.130) The choice machine characteristic of Popperian creatures may hone the capacity to engage in such primitive intentional attribution, but is not a requirement for it.

3.2.3.3. Reidentification of individuals, “track records”, and reputations

Tracking is a capacity that is widespread in the animal world, as it is involved in most forms of sexual reproduction and hunting (two sorts of tracking may be distinguished, cooperative and competitive). Tracking often involves the reidentification of particular individuals: «The practices and projects of many creatures require them to track and reidentify individuals — their mothers, their mates, their prey, their superior and subordinates in their band [...]» (1996, p.117; emphasis added)

In certain species of higher social animals, the members of a particular group mutually recognize each other, i.e. each individual is routinely reidentified by all the others (by contrast, individuals in colonies of social insects are anonymous). Social organization is usually hierarchized, so that each band member has a hierarchical status.
In certain social species (notably among apes), evidence suggests that band members hold “track records” of each other past behavior. Some of those species also seem to have the capacity to distinguish cases of “getting it right” from cases of “getting it wrong”. Individuals in those species are able to hold track records of their fellow band members, and these track records are both topic-sensitive and false-belief sensitive — a sort of primitive scorekeeping. Within those species, band members may not only have a hierarchical status, but also a topic-relative “credibility”, or “reputation”. In this connection, Dennett refers to the work of primatologists Cheney and Seyfarth, who have studied how, in bands of vervet monkeys, individuals earn or lose their reputation as reliable callers of various species of predators.

3.2.3.4. Towards prelinguistic Homo sapiens

We have seen that certain species engage in tactical, even deceptive forms of behavior, the best explanation of which calls for higher-order intentionality — most plausibly intentionality of the second order, although there are no a priori bounds on attainable degrees of intentionality.

We have also seen that it is plausible that certain species of primates hold topic-sensitive and “false-belief sensitive” records of the past behavior of other members of their group; and that such records determine the credibility that those
individuals have, in the eyes of the other members of their group, in regard to those topics.

These results are relevant in regard to the theoretical reconstruction of the abilities of early Homo sapiens. It gives credibility to the hypothesis that our immediate, prelinguistic ancestor had their capacity to suspect deceptive behavior on the part of others, and to hold track records of his fellow band members in regard to cooperation and defection, so that individuals could have a reputation as cooperators or as defectors. That these developments were part of a “cognitive arms race”, from which proto-morality arose, is the hypothesis that Dennett defends, and that we will now examine.

### 3.3. PROTO-MORALITY IN PRELINGUISTIC HOMO SAPIENS

The next step in the Dennettian Just So Story tells how prelinguistic Homo sapiens came to acquire proto-moral dispositions, to follow proto-moral norms, and (by the same token) to possess proto-moral proto-concepts. This section presents a consequentialist rationale of how prudential, proto-moral dispositions, norms and sentiments were selected. As we will see in section 3.4., language and language-produced phenomena (notably rational criticism) will transform proto-morality into genuine morality, so the consequentialist account of the selection of the former should not be mistaken for a theory of the latter.
Dennett endorses this broad movement of evolutionary accounts of the emergence of proto-morality which basically corresponds to successive waves of increasingly complex game-theoretic models (3.3.1.) Dennett emphasizes the importance of the cognitive “arms race” generated by the tension between the necessity of cooperation (Homo sapiens is a social animal) and the ever-present possibility of defection. This line of inquiry suggests that cooperative attitudes and behavior were adaptive enough to be selected, although defection is never quite eradicated.

Subsection 3.3.2. focuses on the question as to whether the capacities with which the later game-theoretic models endow their virtual agents can be reasonably attributed to prelinguistic Homo sapiens. Generally, we may ask: what capacities can we reasonably attribute to our immediate, prelinguistic ancestor?

Subsection 3.3.3. presents an account of the natural selection of proto-moral sentiments, fostered by economist Robert Frank, which Dennett designates as a good example of a sort of evolutionary account that complements game-theoretical analyses, although he does not necessarily endorse Frank’s theory as a whole. Frank’s account proposes that proto-moral sentiments are an emotional phenomenon which had the double adaptive effect of facilitating the adhesion of human agents to cooperative attitudes and behavior (against temptations to defect), and of displaying one’s identity as a cooperator.

It will be remembered from Chapter 2 that Wiggins endorsed (by and large) Sober and Wilson’s evolutionary account of the emergence of “altruistic” —
proto-moral, might we say — attitudes and behavior, and remarked that this theory indirectly supports many features of Hume’s genealogy of moral sentiments and artificial virtues. Dennett also endorses Sober and Wilson’s account, but he accords more importance to game-theoretic models and to Frank’s theory of the selection of proto-moral sentiments (2003, p.194-195).

In any case, this sort of investigation manifests the point evoked by Wiggins, that evolutionary theory by and large supports — and makes more precise and explanatorily powerful — much of Hume’s theory of human nature, or rather of our “human-cum-acquired” nature. More broadly, it supports and makes much more explanatory powerful the sort of historical reconstruction about the emergence of societies, initiated by Hobbes, and to which Hume adhered.

When we get to Chapter Four, we will examine another, complementary point of origin of morality: the enhancement of self-control provided by the attitude of taking responsibility for one’s decisions and performances.

To situate prelinguistic Homo sapiens historically, it might be worth mentioning that (as Dennett does (1991, p.189-190)) the common primate ancestor of humans and chimpanzees probably flourished some six million years ago. Three and a half million years ago, our hominid ancestors still had ape-sized brains. The period known as “the Great Encephalization” began about two and a half million years ago, and was essentially completed roughly 150 000 years ago. Early Homo sapiens appeared at the end of the Great Encephalization, before the development of language, of cooking, and of agriculture. Language emerged
approximately 100 000 years ago, together with specifically linguistic functional structures. Apart from this “language organ” (as Chomsky famously dubbed it), the genome, and thus the gene-based functional design of early Homo sapiens, and ours, are supposed to be identical.

3.3.1. The emergence of proto-moral cognitive dispositions and norms: game-theoretical accounts

Evolutionary hypotheses about the emergence of proto-moral dispositions in early Homo sapiens have become increasingly popular during the 1970’s. Dennett’s remarks on this topic trace the story of a theoretical development pioneered by sociobiologist Robert Trivers (among others), and driven by increasingly elaborate and credible game-theoretic models.

There seems to have been a consensus early on about the basic predicament of prelinguistic Homo sapiens, in connection with the emergence of proto-morality. It could be summarized as follows.

Early Homo sapiens is a social animal, living in groups within which individuals mutually reidentify one another. Individuals in these groups have the two following options: cooperation, or “defection” (i.e. whatever combination of might and deception that might allow one to appropriate another’s resources, or to monopolize resources which would otherwise would be split more evenly). It is
uncontroversial to postulate that early Homo sapiens has the capacity to engage in deception with conspecifics, notably members of the same group.

The hypothesis that Dennett defends is that the tension between the necessity of cooperation and the ever-present possibility of facing defectors (also called “freeloaders”) initiates a cognitive (but, at first, non-linguistic) “arms race”: would-be cooperators and potential defectors try to anticipate (to “outguess”) the moves of others, and to secure their own ends (be they cooperative or deceptive). Thus is generated a series of ever more complex anti-defection measures, always matched by new tactics of defection.

This “arms race” (the hypothesis continues) stimulates the emergence, and the selection, of new cognitive, behavioral and emotional dispositions. Plausibly, gene-based functional structures underlie some of these dispositions, while others are cultural. Moreover, a crucial social phenomenon appears within groups of early Homo sapiens: certain rules get enforced across the whole group, and their transgression is systematically punished. In other words, social norms emerge. Finally, certain proto-concepts accompany these dispositions and norms. The idea is that these dispositions, norms, and proto-concepts, borne out of the attempt to protect cooperation against defection, can be described as proto-moral.

Dennett mentions that Robert Trivers reflected upon the role, in the development of human intelligence, of what he called “reciprocal altruism”: an interactive relation between agents, who can «form mutually beneficial
arrangements of *quid pro quo*, the first step towards human promise-keeping.» (1995, p.479) Trivers suggested that early Homo sapiens had for starters the cognitive capacities required for engaging in reciprocal altruism (notably «a rather specific memory capable of reidentifying one’s debtors and creditors, and the capacity to spot a cheat» (ibid)); and that reciprocal altruism further stimulated the development of human intelligence. Dennett describes Trivers’ theory as “largely speculative”, but it eventually obtained support from game-theoretic models (as we are about the see, early models delivered pessimistic conclusions, but later models tend to support Trivers’ optimism).67. Dennett informally distinguishes four generations in the history of evolutionary game-theoretic models.

3.3.1.1. The Prisoner’s Dilemma and its early applications

The first generation corresponds to the work of John Maynard Smith. The question that initiated his research was whether cooperation was an “evolutionary stable strategy” (ESS).68. Maynard Smith considered that game theoretic analyses,  

67 Dennett mentions in passing that animal psychologist Nicholas Humphrey arrived at conclusions similar to those of Trivers, «by a different and in some regards less speculative route.» (1987 [1983], p.244)

68 Dennett describes an evolutionary stable strategy as «a strategy that may not be the best imaginable but is un-subvertible by any alternative strategy in the circumstances» (2003, p.149).
like the “Prisoner’s Dilemma”, could adequately model the evolutionary dynamics between two categories of agents (cooperators and defectors), and from there could suggest an answer to his question.

The Prisoner’s Dilemma models the possible choices faced by two agents, A and B. Each of them must individually choose one of two options: to cooperate with the other agent, or to defect. The context of decision-making is the following: imagine that A and B are two imprisoned associates, separately incarcerated, interrogated by the authorities. Each prisoner knows that a defection on the part of the other is necessary for her to be sentenced. If both of them remain silent, both will be released. Now, the authorities propose to each prisoner separately, on the condition that she denounces the other, liberation plus a reward. The denounced prisoner, in addition to being sentenced, would have to pay a fine to cover the other’s reward. But if each prisoner denounces the other, they both get sentenced, without any fine to pay (nor any reward).

Understood that “to cooperate” here means for one prisoner to cooperate with her associate (and so to refuse to denounce her), and that “to defect” means to denounce her associate, there are four possible outcomes: A cooperates/B cooperates; A defects/B cooperates; A cooperates/B defects; A defects/B defects. To those four possible outcomes correspond four different payoffs, to which numeral values are assigned: “0” corresponds to the sentence plus the fine; “1” to the sentence without the fine; “2” to liberation without the rewards; “3” to the liberation with the reward. We then obtain a payoff matrix according to
which defection on the part of both agents has a payoff of “1” for them both; mutual cooperation has a payoff of “2” for both agents; and when one agent cooperates while the other defects, the defector obtains a payoff of “3” while the cooperator gets a payoff of “0”.

Of course, this fictional setting is just a way of making more concrete an abstract model of the respective benefits and risks of cooperation and defection in general. One could think up a virtual population (not necessarily human) whose members would have to face the Prisoner’s Dilemma, according to varying parameters, and this could be taken as a model of the evolutionary dynamics of cooperation and defection.

In the early 1970’s, then, Maynard Smith proposes an evolutionary version of the model, with a mixed population of cooperators and defectors (also called “freeloaders”), and where payoffs represent reproductive success. Agents are ascribed a fixed behavior-type, as if it was determined by a single gene (some agents are programmed as “cooperators”, others are “defectors”). Agents are randomly matched, and interact. In such conditions, if a cooperator is lucky enough to be randomly matched with another cooperator, both will thrive (they get the 2-2 payoff). If a defector is matched with another defector, both have mediocre success (they get the 1-1 payoff). Finally, in the case where a cooperator is randomly matched to a defector, the defector has the greatest reproductive success, and the cooperator the lowest (the 3-0 payoff). Maynard Smith concludes that cooperation is not an ESS.
Indeed, the early models of the 1970’s in general suggest that defectors tend to take over, although this tendency creates various oscillations which impede its straightforward course: as defectors proliferate, the frequency of the defector/defector type of match increases. This eventually favors a comeback by cooperators, who then get preyed on again. However, the overall tendency favors defectors:

These models show that defectors tend to do very well indeed, though they can pollute their own nest: as the proportion of freeloaders increases, they tend to meet each other more often, in costly bouts of mutual defection [...] So the cooperators start making a comeback, but only until there are enough of them around to be worth preying on, at which point the freeloaders begin thriving again. (2003, p.198-199)

Many evolutionary theorists have acknowledged that game-theoretic analyses like the Prisoners’ Dilemma may indeed model certain forms of interactive dynamics that play an important role in evolution. As Dennett simply puts it: «Many species, in many regards, face various sorts of Prisoner’s Dilemma». (1995, p.255) However, the results delivered by the early models were not entirely credible:

[T]he models also exhibited some strange effects, settling into equilibria that don’t match our expectations, and thus raising the prospect that at least some of the behavior of the models was artifactual, an unintended by-product of the oversimplifications rather than a reflection of something in the real world. (2003, p.199)
3.3.1.2. Later models

The problem of the early game-theoretic models was that they were extremely schematic. Subsequent generations of models have added what are taken to be more realistic specifications to the agents’ profiles, and to the conditions of their interaction.

One result of this increase in the complexity of the models is that they rapidly became theoretically intractable, i.e. it became impossible to draw mathematical demonstrations from them. On the other hand, as increasingly powerful computers became available, it became possible to run simulation programs, and to observe (instead of calculating) the evolution of virtual populations. Computer simulations are not theoretical demonstrations, nor are they equivalent to empirical testing. But they are now generally considered as a credible way of probing certain hypotheses, and of generating others. Behavioral patterns (e.g. patterns of cooperation and defection), leave no fossil records to seek out in the field, but it is possible to observe simulated versions of them, and so to obtain credible, quantitative data.

After the original generation, Dennett informally distinguishes three subsequent “waves” of increasingly complex game-theoretic models (2003, p.199-200). Regarding the models of the second wave, he simply mentions that instead of pairing agents at random, they modelize certain factors of spatiality
(like a “viscosity factor”), following which certain “neighborhoods” emerge, including concentration of cooperators that increase the probabilities for cooperators of interacting with their own kind.

The third wave of models introduces “selective altruism”. These models endow agents with the capacity to hold track records of other agents, and to set conditions on cooperative interaction. These conditions can be developed into varying “policies of forgiveness and forgetfulness” — and the more widely a particular policy is shared across a population, the more it resembles a social norm. Agents thus have the capacity to select who they interact with, and on what condition. This allows cooperators a certain protection against freeloaders. Dennett commends the contribution that philosopher Philip Kitcher has brought to this sort of model:

Kitcher shows, in careful mathematical detail, how “discriminating altruists” (who keep a tally on who has defected in the past) can flourish under certain — not all — conditions, and also begins to sort out the conditions under which varying policies of forgiveness and forgetfulness can hold their own against the ever-present prospect of a resurgence of antisocial types. (1995, p.480)

The fourth wave of models is of course the most interesting. Agents are endowed with greater discriminativeness, as well as with a finer capacity to react to the behavior of others. They also have the capacity to acknowledge certain norms, to recognize what their transgression consists in, as well as to identify norm-violators. These emerging norms, then, are standards of correctness.
shared by the members of a group (compare Wiggins, who evoked “shared standards for the evaluation of characters”).

Moreover, agents have the capacity to create coalitions of norm-enforcers, and to punish violators, as well as those who, although not violators themselves, refuse to punish them. In such conditions, agents learn, norms evolve, the situation becomes complex, and optimistic prospects arise:

[In preceding models] there would be no capacity of the various individuals to notice what was happening, to raise the alarm, to deplore, to propose sanctions, to form vigilante groups, to brand or punish the freeloaders among them. Once we add simple versions of this reactivity, it ushers in a wave of new complexities. Dire conditions that had seemed inevitable now turn out to be preventable after all, thanks to the timely and well-aimed use of information by group members. [...] A group’s evolution of the capacity for policing its members, by adopting the disposition among its members to punish violators (of whatever its other policies are) opens the floodgates to the social or cultural evolution of all manner of local norms. (2003, p.200-201)

Here we have, Dennett tells us, a «self-maintaining societal structure», where cooperative behavior has become proto-moral behavior, regulated by proto-moral norms. Defection is not eradicated, but it is prevented from taking over.

In order for the models to be credible, the virtual agents that they “set up” must only be endowed with capacities that early Homo sapiens could reasonably have been endowed with. Did our immediate prelinguistic ancestors really had
the capacities required to be selective in his interactions with others, to punish norm-violators, etc? What capacities our immediate ancestors actually had is a difficult issue, but we can circumscribe a range of capacities that it is reasonable to attribute to them. According to Dennett, the fourth-wave models are credible, and attribute nothing but reasonably basic, prelinguistic capacities to early Homo sapiens:

[...] since this sort of self-maintaining societal structure can now be seen as a necessary precondition for the long-term flourishing of genuinely altruistic agents, it is reassuring to see how little must be presupposed to get it to evolve and to sustain itself: The very simplicity and relative rigidity of the abilities to discriminate the freeloaders from the good citizens, and the dispositions to “punish”, show that as far as this feature of culture is concerned, it could predate language and convention and ceremony. We’re not talking about trial by jury and public denunciation here; we’re talking about an unreflective, “brute” inclination to channel some risky aggression against those of one’s group one has discriminated as norm-violators. (2003, p.201-202)

The capacities required for proto-moral dispositions and norms to emerge, then, are basic enough for their attribution to our prelinguistic ancestor to be reasonable. However, Dennett does not provide anything like a list, however incomplete, of those capacities that early Homo sapiens had to have in order for the scenario suggested by recent game-theoretic models to be possible. In the next subsection (3.3.2.), I propose a minimal, and no doubt incomplete, list of my own. I take it that Dennett endorses the idea that prelinguistic Homo sapiens was indeed endowed with the capacities featured in the list.
3.3.2. Capacities and faculties of prelinguistic Homo sapiens

First, let us review those capacities that it seems a creature must have, in order for the arms race that stimulates the rise of proto-morality (as described above) to be possible. Then, we will see if it is reasonable to attribute such capacities to prelinguistic Homo sapiens.

I have mentioned in the introduction of the present section that it is more or less taken for granted that our immediate prelinguistic ancestor had the capacity to engage in deception with conspecifics (notably members of the same group). We can now take a further step: in order for the “arms race” which drove the developments sketched by the game-theoretic models to be possible, early Homo sapiens had to have the capacity, not only to engage in deceptive behavior, but to suspect deceptive tactics on the part of any other agent. This implies the two following, more specific capacities.

First, the capacity to entertain intentional states of the third order (intentional states about intentional states bearing on yet other intentional states). This can be described as the capacity to entertain (for example) the belief that X desires that Y believes something.
Second, the capacity to discriminate between “getting things right” and “getting things wrong”, and so an appreciation (however dim) of the phenomenon of “false belief”.

Together, these two capacities add up to the capacity to entertain the belief that X desires that Y believes something wrong. This enables prelinguistic Homo sapiens to suspect deception (including attempts to induce false beliefs) on the part of other agents.

For the recent models sketched in the previous subsection to be plausible hypotheses about what might have happened, early Homo sapiens must also be able to engage in nonlinguistic scorekeeping: the capacity to hold track records of how reidentified individuals behaved in the past, in regard to cooperation and defection (including attempts to induce false beliefs). Thus, each group member has a “reputation” in the eyes of the others, and can be seen as a reliable, trustworthy cooperator, or as likely to defect.

An implication of what precedes is that early Homo sapiens acquired a certain repertoire of proto-moral proto-concepts, which I propose to divide in three rough-and-ready categories.

First, if the last wave of game-theoretic models are any guide, early Homo sapiens must have had proto-concepts of “norms”, of “right” (corresponding to the respect of norms) and of “wrong” (corresponding to the violation of norms).
Second, if our evolutionary ancestors recognized human agents as the authors of their performances and attitudes, including "right things" and "wrong things", then they must have had proto-concepts of “accountability”, “desert”, “merit”, and “fault”.

Third, if they had the aforementioned proto-concepts, then they must have had proto-concepts of deontic statuses, namely commitment (members of a community are committed to obeying its norms) and entitlement (one is entitled or not to certain things given certain factors: one’s hierarchical status in the group, one’s “track record” and therefore “credibility” and “reputation”, etc.) Dennett suggests as much when he writes, echoing Trivers, that «mutual recognition and the capacity to communicate a promise […] are necessary conditions for the evolution of morality.» (1995, p.481) The idea of “commitment” will be evoked in the next subsection (3.3.3.), in connection with the emergence of proto-moral sentiments.

Did prelinguistic Homo sapiens indeed possess such proto-concepts? Did he possess all the cognitive abilities required for the emergence of the sort of proto-morality sketched by the last wave of game-theoretic models? I propose to make two points here (which I take to be “Dennettian” points). One takes us back to cognitive ethology, the other is about the evolutionary interpretation of certain empirical results of recent psychological research.
As we have seen in 3.2.3., there are arguments and evidence to the effect that many animal species possess proto-concepts, the ability to entertain intentional states of the *second* order which may be used in deceptive behavior, and the capacity to keep track records of reidentified individuals, so that, within a group, each member has a certain “reputation” (or a certain “credibility”) in the eyes of the other members. From this, we cannot directly infer anything about what capacities early Homo sapiens actually had, but it does suggest that it is reasonable to attribute to him the capacities listed above.

Empirical evidence in psychology shows that human reasoning displays certain patterns of strengths and weaknesses, of which it is tempting to make certain evolutionary interpretations. For example, some experiments have been carried off concerning differential success in the performance of certain inferential tasks, presented according to two different modes, but otherwise identical. One mode of presentation is abstract, the other is applied to situations in which the inferential tasks take the form of “cheater detection” (or “patrolling of social contract”). Subjects who perform those inferential tasks in their applied, “cheater detection” form are significantly less prone to error than those who perform the same tasks presented in their abstract form. As Dennett himself points out (1995, p.488-492), some evolutionary psychologists (notably Leda Cosmides and John Tooby) suggest that this well-confirmed cognitive distorsion is a manifestation of our evolutionary heritage. The hypothesis is that early Homo sapiens was equipped with functional structures (“cognitive modules”) *dedicated to reasoning*
applied to the patrolling of social contract; and that we have inherited of these modules, which are still operative in us.

Certain scientists and philosophers (including notably linguist Ray Jackendoff and psychologist Simon Baron-Cohen) designate a certain set of cognitive abilities manifested by humans, hypothetically underlain by gene-based dedicated modules, as constituting a “faculty of social cognition”. The hypothesis is that the modules composing this faculty were selected throughout evolutionary history, because they endowed those who possessed them with abilities that proved to be adaptive in regard to problems raised by sociality (first animal, and then human). According to this hypothesis, we are endowed with the same gene-based faculty of social cognition as prelinguistic Homo sapiens was (we have the same genome as our immediate ancestor, the only significant difference being the recent addition of the “language organ”).

Dennett’s attitude toward those hypotheses is sympathetic but careful. That is, he thinks that the hypothesis is correct, but he reiterates the importance of this methodological principle that hypotheses of gene-based functional structures are always rivaled by hypotheses of cultural transmission, and ruling the latter is extremely delicate. On Cosmides and Tooby, and the position they defend, he makes the following comment:

Even if they are right — and I am confident that they are — that such rationality as we human beings have is the product of the activities of a host of special-purpose gadgets designed by natural selection, it does not follow that this “Swiss-army knife” of ours cannot have been used,
time and time again, to reinvent the wheel. It still has to be shown, in other words, that any particular adaptation is not a cultural product responding quite directly (and rationally) to quite recent conditions. They know this [...] but in the heat of battle they sometimes forget. (1995, p.490)

In any case, evidence is gradually accumulating in favor of the sort of view that Dennett and others defend: prelinguistic Homo sapiens was endowed with proto-moral cognitive dispositions, possessed proto-moral proto-concepts, and lived in groups within which proto-moral norms were enforced. (Spelled out more elaborately: a cognitive arms race exerted selective pressure on cognitive and behavioral policies, with the following results: although defection can bring immediate benefits, in the long run, cooperative attitudes and behavior prove to be more adaptive than defection. Consequently, they are selected. Non-linguistic proto-moral norms, and non-linguistic proto-moral behavior, were naturally selected, in reason of their adaptivity.)

Dennett emphasizes that proto-morality must not be confused with genuine morality. The conflict between cooperative attitudes and behavior on the one hand, and deception and defection on the other, comes down to a tension between the far-sighted pursuit and protection of one’s long-term self-interests, and the myopic pursuit of one’s narrow, immediate self-interest. That is, if what the later wave of game-theoretic models suggest is right, then it is in one’s long-term interest to go beyond selfish motivation, to cooperate, and to stand in solidarity with fellow cooperators. Defection may be profitable in the short term,
but the individual who is recognized as a defector is likely to face severe consequences, at least in the long run.\footnote{Hence it is something of an abuse of language to talk, as Trivers and Kitcher did, of “reciprocal altruism” and “selective altruism”, respectively.}

We will in section 3.4. see how the advent of language and of rational criticism transformed proto-morality into genuine morality. But before we get there, we may add another element to the evolutionary account of the emergence of proto-morality: an account of the natural selection of proto-moral sentiments, as contributing to one’s long-term interests by making one stick to cooperative commitments.

### 3.3.3. Proto-moral sentiments

Let us admit, as the last wave of game-theoretic models suggest, that while defection can be profitable in the short term, defectors are likely to get eventually caught and punished, so one’s long-term interests lie on the side of cooperation. Then, it may appear that defecting \textit{without} getting caught, while passing for a cooperator, could be the most profitable strategy. Is not the mere perspective of such a possibility enough to seriously destabilize the cooperators
and norm-enforcers within a group, and to ruin the dynamics optimistically suggested by the last wave of game-theoretic models?

According to economist Robert Frank, defecting while attempting not to get caught, and feigning cooperation, is not the most beneficial policy in the long run, but it is an immediate temptation. The most effective policy in the long run is a conjunction of two things: to cooperate unconditionally, even when cooperation disserves one's short-term interests; and to have others recognize that one indeed cooperates unconditionally. Frank has recently proposed a theory according to which moral sentiments were naturally selected because they facilitated both the attitude of unconditional cooperation, and the social result of recognition and reputation, while at the same time barring the temptation of secretly defecting while feigning cooperation.70

Dennett is sympathetic to Frank's theory (I will shortly present a summary of it, based on Dennett's own reading). But what matters most, he claims, is not the details of Frank's particular combination of speculation and argumentation. Rather, it is the fact that it illustrates the fruitfulness of the evolutionary approach. The following quote is a typical programmatic claim of Dennett's:

70 In this subsection, the expression “moral sentiment” should be understood as referring to what are proto-moral sentiments, according to the terminology I have been using to distinguish the prelinguistic forerunners of morality, and genuine language-dependent morality.
Frank’s arguments and conclusions have not yet won anything like general acceptance among his fellow economists or evolutionary theorists (or philosophers), and there remains serious problems — and alternatives — that need to be carefully addressed. What is mainly important to me here is that is that Frank’s project […] is an instance of a type of approach to these issues, a Darwinian approach, that, I claim, is both obligatory and promising. It is obligatory because any theory of ethics that just helps itself to a handy set of human virtues without trying to explain how they might have arisen is in danger of positing a skyhook, a miracle that “explains” nothing because it can “explain” anything. It is promising because, contrary to what the enemies of Darwinian approaches declare, novel insights tumble out of the exercises of these theorists with quite gratifying frequency. (2003, p.217)

Frank’s theory, then, proposes that the most effective policy for any agent is to internalize cooperative attitudes, and to have the other members of one’s group acknowledge that one is, essentially, an unconditional cooperator, who sticks to one’s commitments come what may71.

There might seem to be a certain tension between these two elements: if one internalizes certain attitudes, then they — not the advantages they bring — are their own rewards. How then could one consistently want to have everybody acknowledge that she has internalized those attitudes, as part of an “effective policy”?

The answer is that this is not a strategy that an agent might deliberately adopt. It is, rather, the free-floating rationale that accounts for the emergence of

---

71 This chimes with Wiggins’ amendment of Hume’s view of moral education.
moral sentiments. Frank’s proposal is that the emergence of moral sentiments is explained by their being part of the most effective policy. Frank is not proposing a substantive moral theory (i.e. a theory that defines “the good”, “the right”, and their mutual relation). What he proposes, rather, is a genealogical account of the origins of moral sentiments. His point is that these (proto-moral) origins reside in a consequentialist rationale. This does not imply that actual morality as we language-using humans understand it today, in all its elements (the good, the right, moral sentiments), is definable and explainable in consequentialist terms. It is not moral problems that Frank poses in terms of “effective”, or “cost-effective” policies, but rather the proto-moral, prudential problems which are at the origins of morality — in particular, of moral sentiments. This squares with Dennett’s theory of morality, according to which the prelinguistic origins of morality have a consequentialist rationale, which allows the takeoff of proto-moral attitudes, behavior, sentiments, and standards; this becomes fully moral with the emergence of language, which brings about the practice of giving and asking for reasons.

The policy that Frank describes is called upon with particular sharpness when one faces what he calls the “commitment problem”, which arises when one realizes that a previously taken commitment turns out to be contrary to one’s self-interest. But again, it is advantageous in the long run to be locally and temporally irrational (in the sense of forsaking one’s short-term interest) in order to be globally rational (favoring now another’s interests may favor one’s own long-term
interests). So even in cases where it appears that one has undertaken a commitment that goes against one’s self-interest, sticking to the commitments that one has undertaken is the best policy.

Then there is the issue of being recognized by others as a committed cooperator, a reliable partner, a loyal ally, a promise-keeper… As Dennett puts it: «the main problem [is] not just how can you make yourself into an agent that can be trusted in commitment problem cases, but how can you credibly advertise the fact that you are to be so trusted?» (2003, p.205)

Frank speculates, then, that moral sentiments are a “cost-effective way” of achieving the two following results.

First, Frank proposes that moral sentiments partake in our adherence to our commitments and in the internalization of cooperative attitudes (and so in self-control): «Moral sentiments may be viewed as a crude attempt to fine-tune the reward mechanism, to make it more sensitive to distant rewards and penalties in selected instances.» (Frank 1988, p.90; quoted in Dennett 2003, p.214)

Second, moral sentiments partake in the visibility of the committed cooperator’s character: «Others get to see that you are one of those emotional folks who can be counted on to care passionately about your commitments; it is not that you are crazy or irrational but that you put an irrationally high price (from the myopic perspective of the critic) on your integrity.» (2003, p.213-214)
Frank speculates that this makes of moral sentiments an adaptive trait, and so accounts for their selection. In his model, moral sentiments appear independently of language (and so the internalization of cooperative attitudes gets a start independently of language). However, although moral sentiments have a natural base, *moral education* makes them much more “effective”: it molds them into the right sort of disposition, and it makes it possible to solicit them, and to co-opt them in favor of the self-control required to stick to one’s commitments. Moral education, however, requires language.

### 3.4. MIND, LANGUAGE, AND LANGUAGE-PRODUCED PHENOMENA

We now come to Dennett’s theory of the human mind, and to what I have called Dennett’s “linguistic extension” to evolutionary theory. The topics presented in this section are important components in Dennett’s gradualist, skyhook-free picture of human nature. A minimal grasp of them is important for understanding his conception of responsibility and free will, and of the sphere of the unforsakeable.

The human mind is built upon the kinds of minds we have encountered in section 3.2. From our more or less distant ancestors, we have inherited multiple layers of Darwinian mechanisms, situation-action machines, and choice-
machines. Those functional items have constituted primitive cognitive areas within the mental functional design of our species, which language has invested. We have also inherited from the proto-moral dispositions (cognitive, behavioral, emotional) surveyed in section 3.3.

The development of language in early Homo sapiens was accompanied by the parallel development of functional systems dedicated to language. This late addition brought about major changes in the cognition of Homo Sapiens, but it did not change the mental functional design that was already there, and over which it was superimposed. The mental functional design of language-using Homo sapiens, then, is pretty much the same as that of pre-linguistic Homo sapiens.

How did language change the cognition of Homo sapiens? For one thing, language invested the primitive cognitive areas of the mind of pre-linguistic Homo sapiens, and thus entailed the linguistic articulation of already-shaped primitive

---

72 Chomsky’s hypothesis that language is produced by a dedicated cognitive system (a “language organ”) that significantly constrains syntax is widely accepted today. Language probably emerged somewhere around 100 000 years ago, so the “language organ” is a very recent addition in the history of our functional design. Chomsky and Stephen J. Gould have notoriously disputed the hypothesis that language evolved out of natural selection as an adaptation. But the adaptive hypothesis has been defended by many scholars, including Pinker and Bloom, whose position, expressed notably in (Pinker and Bloom 1990), Dennett endorses. He adds that the adaptivity of language probably has to do with the lever it provided in the cognitive arms race between cooperation and defection.
cognitive areas. For another thing, it introduced families of concepts that had no
pre-linguistic forerunners, and thus opened new cognitive areas, initiated new
practices and brought about new phenomena and items. Among these, two are of
particular importance for the purposes of the present dissertation: the self, as a
“center of narrative gravity”; and rational criticism, as the practice of giving and
asking for reasons (typically in the form of “why” questions) in support of
judgments, decisions and actions.

Subsection 3.4.1. is dedicated to the linguistic articulation of primitive
cognitive areas. “What happens”, Dennett asks, “to a human or hominid brain
when it becomes equipped with words?” The answer appears like an empirical
confirmation of Strawson’s hypothesis regarding our “fundamental cognitive
structure”.

Subsection 3.4.2. is dedicated to Dennett’ theory of the human mind. The
irreducibility of the intentional level of discourse is accounted for by the following
insight: the brain is a syntactic engine (part of our overall functional design) that
approximates a semantic engine.

The three last subsections present the linguistic extension as such.
Subsection 3.4.3. presents the very idea of “language-produced phenomena”.
Subsection 3.4.4. is dedicated to the self. Subsection 3.4.5. is dedicated to
rational criticism. The irreducibility of the personal/moral level of discourse is
accounted for by the insight that human beings, from the moment that they can
represent the reasons for their actions and beliefs, can examine those reasons,
endorse them, or reject them, and look for other avenues. In any case they make their judgments, decisions and actions their own, as well as the reasons that support them. Consequently, no moral standards can be derived from the free-floating rationale that accounts for how our prelinguistic ancestors behaved, nor from the functional design of our species, nor from any part of it.

3.4.1. Primitive cognitive areas and language

We have inherited from the evolution of species a broad set of behavioral, cognitive and emotional dispositions, some of which are underlain by gene-based functional structures (or “systems”, or “units”), while others are “good tricks”, transmitted from generation to generation by learning. To certain gene-based functional structures are associated “tacit theories”, constituted by information about Homo sapiens’ environment. This subsection presents evidence concerning some gene-based functional structures and associated tacit theories, which constrain how we conceive of certain subject matters. The set of all relevant functional structures seems to correspond to what Strawson evoked as a “fundamental cognitive structure”, and thus provides empirical support for his speculative hypothesis.

Dennett points out that, following empirical research in developmental psychology and in the neurosciences, operative, domain-relative functional
structures have been identified (hypothetically but plausibly) in prelinguistic children. These constrain not only the language acquisition process, but also the very linguistic articulation of concepts in the domains in question. In other words, these functional units constitute primitive cognitive areas, which shape how we conceive the corresponding subject matters:

What happens to a human or hominid brain when it becomes equipped with words? In particular, what is the shape of this environment when words first enter it? It is definitely not an even playing field or a tabula rasa. Our newfound words must anchor themselves on the hills and valleys of a landscape of considerable complexity. [...] A number of investigators are currently exploring portions of this terrain. The psychologist Frank Keil (1992) and his colleagues at Cornell have evidence that certain highly abstract concepts — such as the concepts of being alive and ownership, for instance — have a genetically imposed head start in the young child’s kit of mind-tools; when the specific words for owning, giving and taking, keeping and hiding, and their kin enter a child’s brain, they find homes already built for them. Ray Jackendoff (1993) and other linguists have identified fundamental structures of spatial representation — notably designed to enhance the control of locomotion and the placement of movable things — that underlie our intuitions about concepts like beside, on, behind, and their kin.

Nicholas Humphrey (1976, 1983, 1986) has argued that there must be a genetic predisposition for adopting the intentional stance, and Alan Leslie (1992) and others have developed evidence for this, in the form of what he calls a “theory of mind module” designed to generate second-order beliefs (beliefs about the beliefs and other mental states of others). [...] So the words [...] that take up residence in a brain [...]
enhance and shape pre-existing structures, rather than generating \textit{entirely} new architectures. [...] (1995, p.378-379)

A few remarks are in order about higher-order intentionality, and the "theory of mind module" hypothesized by Alan Leslie.

First, a reminder: empirical evidence in cognitive ethology suggests that certain animal species access second-order intentionality; and, according to the hypothesis examined in section 3.3., the third level of intentionality must have been accessible to early Homo sapiens.

Second, if the capacity to attribute second-order beliefs is underlain by a gene-based "theory of mind module", it does not follow that this module imposes any limit on the level of intentionality accessible to humans. There are, obviously, practical limits to the level of intentionality we may reach, but probably no principled limit. Our capacity on this count might not depend on language. Dennett points to an empirical fact that echoes the empirical research results evoked in 3.3.2., to the effect that reasoning abstractly is more difficult (and for most people less reliable) than reasoning in applied contexts of cheater detection: in a similar way, engaging in higher-order intentionality is easier when we actually interpret the conduct and attitude of other persons than when we engage in more abstract forms of iteration:

Sometimes higher orders are so easy as to be involuntary. Why is the fellow in the movie trying so hard to avoid smiling? In the context it's deliciously obvious: his effort show us he knows she doesn't realize he already knows she wants him to ask her to the dance, and he wants to keep it that way! Other times, simpler iterations can stump us. Are you
sure that I want you to believe that I want you to believe what I’m saying here? (1996, p.121)

Some areas of our ordinary outlook, which bear upon “primitive” domains, and which seem to be organized in “theories” shared by all humans, might be conceptual articulations of tacit theories associated to gene-based mental functional structures. For example, if all humans share a common “folk physics” (bearing on the physics of middle-size visualizable objects), then it is highly plausible that this folk theory is a linguistic articulation of a tacit theory, associated to cognitive functional structures which care about the behavior of middle-size objects. Similarly, if all humans share a common “ordinary psychology” (bearing on the attribution of propositional attitudes to intentional systems), then it is highly plausible that it is a conceptual articulation of a tacit “theory of mind”, associated to mental functional structures that care about the behavior of intentional systems.

Now, consider this possibility: if all humans share the same demand for minimal regard, and for a minimal demonstration of good will, on the part of those persons whom they see as “responsible”, and with whom they interact; if all humans have similar emotional and attitudinal responses to what they interpret as a lack of minimal regard, then it is plausible that this is a conceptual articulation of certain requirements associated to a primitive faculty of social cognition.
3.4.2. Dennett’s theory of the mind

Dennett’s functionalist theory of the mind proposes to eliminate what he regards as a major, dual skyhook, corresponding to two concepts (or clusters of concepts). One corresponds to our ordinary idea of the self (which I call the “Conscious Active Self”), the other corresponds to the philosophical version of the latter (I call it the “transcendental ego”). The ordinary notion of the Conscious Active Self has been rather supported, and elaborated, by various philosophical traditions and doctrines — a case of unjustified vindication of an erroneous feature of our ordinary outlook. Our concept of the Conscious Active Self, together with its philosophical version, implies the existence of (something like) a central authority within the mind, acting like a Central Observer, Central Feeler, Central Representer, Central Intender, Central Meaner, Central Judge, Central Executive, etc. Thus, the skyhook of the Conscious Active Self (and transcendental ego) entails other more specific skyhooks, e.g. ideas of “original intentionality”, “original understanding”, etc.

But, Dennett claims, empirical findings show that no such central authority exists, and suggests that the true nature of the appearance of central authority is a sort of “user illusion” (somewhat like the illusion of a unified interlocutor created by interface software). Conceptually speaking, postulating a Conscious Active Self — and specifically a central authority within the mind — is a mistake not unlike postulating an entity called a “pair of gloves”, over and above the two individual gloves.
Paragraph 3.4.2.1. presents the cardinal ideas and principles which constitute the core of Dennett’s theory of the mind. Then we switch to the critical side of Dennett’s theory (3.4.2.2.), which is directed to our ordinary conception of the self, as well as to its philosophical versions, especially its most recent variety, which he calls “Cartesian materialism”.

Paragraph 3.4.2.3. presents another insight of Dennett’s theory of the human mind, which throws light on Quine’s thesis of the indeterminacy of radical translation, and explains why the intentional level of discourse is irreducible to the functional level. The human mind is responsive to meaning, and as such it seems justified to describe it as a “semantic engine”. But the human brain is in fact a syntactic engine, functionally organized and mechanistically driven, designed by the evolution of species in a way such that it approximates a semantic engine. Therefore, the human mind is not a real semantic engine, but an approximation realized by an exquisitely designed syntactic engine.

Paragraphs 3.4.2.4. and 3.4.2.5. present the models of speech production and of understanding that Dennett draws from his overall theory. These contribute in clarifying and grounding the semantic engine/syntactical engine insight.

3.4.2.1. The human mind as centerless and fragmented
The human mind can be described as a broad set of dispositions, underlain by nested hierarchies of functional structures, which are physically realized in parallel neuronal networks, and which perform nested hierarchies of goal-oriented tasks to which they are dedicated.

Examining the nested hierarchies of functional systems, downwards, we reach a ground floor: simple neurons, acting like ON/OFF switches. Examining the same hierarchies upwards, we shall not reach one ultimate, unitary functional structure, including all others. The human mind was built through evolutionary history by the accumulation of distinct functional systems, realized in layers of neural networks. It has remained a more or less fragmented coalition of systems which work together, but which never merged into one all-inclusive unit.

Cognitive tasks are performed according to a principle of generation-and-selection which we have already encountered, in our paragraph on Popperian creatures (3.2.2.3.). That is, cognitive tasks can be analyzed into two main complementary processes: on the one hand, the generation of options (of data interpretation, of action) by functional systems that may be generically designated as "generators"; on the other hand, the selection of one option, which is acted upon. These are "mechanistic" process, the result of many mechanisms functioning together. The selection of options involves no other inner agency or "authority" than the mechanism of natural selection that drove the evolution of species involves a "higher" agency or authority. In fact, Dennett suggests that there is no difference in kind between the selection process that occurs, as part
of various cognitive performances, within the internal environment of the mind (across all the nested hierarchies of functional structures, and the neuronal networks that realize them), and the process of natural selection that has driven the evolution of species.

Dennett suggests that this principle is involved in the performance of many human cognitive tasks and phenomena, including decision-making and consciousness.

Dennett remarks that as we go about our daily routines, we constantly and automatically take decisions, about which we barely think. Some of them are plausibly performed by some “choice machines”, that we inherited from our Popperian ancestors. But sometimes, we consciously deliberate. The Dennettian hypothesis is that in the case of deliberation, the selection of options is particularly complex, and governed by a great number of rules which we do not merely follow, but also consult. We will return to deliberation in Chapter Four.

The same principle is also responsible, according to Dennett, for the serial character of what we experience as “consciousness”. However, even though some mental events are clearly conscious while others are clearly not, Dennett’s “multiple drafts” model of consciousness denies that a clear-cut border separates all that is conscious from all that is not. Similarly, it denies that a clear-cut border separates “passive” experience from agency (or “volitional activity”).
Overall, Dennett proposes a picture of a centerless, fragmented mind, within which tasks are exhaustively distributed across a multitude of functional structures and neuronal networks, working in parallel.

3.4.2.2. Cartesian materialism and the ordinary notion of the Conscious Active Self

It is hardly disputable that the theory of the mind that Dennett defends goes against our ordinary idea of the self. We spontaneously form an idea of our conscious selves. I am the subject my own experience, and the agent of my own volition. This suggests a dichotomy of passivity and agency, and the fact is that there indeed seems to be clear cases of passivity (the content of my current perceptual experience), as well as clear cases of agency (willful action). But at the same time there is a shady zone in between, where my experience and my agency merge into the flow of my cognitive activity, my memories etc.

According to Dennett, our ordinary, unreflective distinction between “what we do” (acts of volition) and “what happens to us” (non-voluntary mental events) largely contributes in the generation of our ordinary conception of (what I have called) the Conscious Active Self. It seems phenomenally obvious that whatever causes passive experience comes from the input periphery of our senses and
travels from there towards some experiential “center”; and that agency comes from some volitional “center” and travels towards some output periphery:

Thus *feeling pain in my foot* and *seeing the desk* are clearly not acts “in my control”, but things that happen to me as a result of impingements from the world. And *moving my finger* or *saying these words* are obviously things that I do — voluntary actions par excellence. (1984, p.79)

Whatever lies in between, however, is not clearly passive nor active. It is this void, Dennett suggests, that we fill with the idea of the Conscious Active Self:

Faced with our inability to “see” (by “introspection”) where the center or source of our free actions is, and loath to abandon our conviction that we really do things (for which we are responsible), we exploit the cognitive vacuum, the gaps in our self-knowledge, by filling it with a rather magical and mysterious entity, the unmoved mover, the active self. (ibid)

A long series of otherwise different philosophical doctrines suggests that this ordinary idea of the Conscious Active Self corresponds to something real, which stands at the center of the mind (and so at the center of philosophy and of psychology). I group under the heading of “transcendental ego” — that expression understood in a generic sense — the philosophical notions of the Conscious Active Self that were developed by Descartes and the rationalist tradition, by Kant, and by Husserl and the phenomenological tradition.
Some empiricist philosophers have challenged the ordinary idea of the Conscious Active Self and of its philosophical version (the “transcendental ego”), but they had no credible alternative. Notably, Hume, and Mach after him, argued against the transcendental ego, and left the idea of the self in a state of complete dissolution, which is hardly convincing. Ayer, in Language, Truth and Logic, proposed a phenomenalist conception: «the self is reducible to sense-experiences, in the sense that to say anything about the self is always to say something about sense-experiences» (Ayer 1946 [1936], p.128). But this phenomenalist position proved to be problematic.

However, as Dennett points out, the “transcendental ego” tradition had to accommodate growing evidence that all information undergoes parallel distributed processing in the brain. The result is what he calls (somewhat ironically) “Cartesian materialism”. According to Dennett, this theory purports to integrate both functional insights and the pandemonium model of mental activity (and so empirical results of the neurosciences), but in fact it still implies the transcendental ego (and so indirectly vindicates our ordinary notion of the Conscious Active Self). That is because, instead of acknowledging the principles summarized in the preceding paragraph (3.4.2.1.), it construes of cognitive tasks as performed by one central agency, and of experience as “felt” by one central subject. As a result, according to Dennett, the existence of a transcendental ego is implied, if not necessarily claimed (and so the ordinary idea of the Conscious Active Self is indirectly vindicated — as it should not, Dennett insists). Let us see how this works.
According to Cartesian materialism, input systems send masses of information “inside”, below a supposed “border of consciousness”. This information is subconsciously processed through multiple parallel networks, which compose functional structures.

At some point, some information crosses the line of consciousness, and is centralized and “taken in” by some sort of central mental “entity” who experiences the incoming information (including the feeling of qualia), understands linguistic stimuli, makes decisions, and initiates action (linguistic and nonlinguistic behavior) — in other words, a Conscious Active Self, implicitly designated as the subject of experience, the agent of volitional activity, the locus of “original intentionality” and of “original understanding”, etc. The parallel distributed processing of information is seen as the machinery which, working below the line of consciousness, sustains the life of this homuncular Conscious Active Self.

Again, the crucial mistake corresponds to a cluster of faulty assumptions, according to which there must be a point “where it all comes together”, a unified subject who feels, a unified agent who performs those cognitive tasks carried off above the supposed line of consciousness (thinking, taking decision, initiating action). The existence of the Conscious and Active Self is not stated, but it is implied:

Perhaps no one today explicitly endorses Cartesian materialism. Many theorists would insist that they have explicitly rejected such an obviously bad idea. But [...] the persuasive imagery of the Cartesian Theater keeps coming back to haunt us — laypeople and scientists
alike — even after its ghostly dualism has been denounced and exorcized. (1991, p.107)

The “Cartesian Theater” evoked in the quote above is Dennett’s expression to designate the virtual centralization, above the line of consciousness, of felt experience and of the performance and authorship of cognitive tasks. That is a virtual Theater. Cartesian materialism does not postulate (as Descartes did) a spatial location in the brain where all information comes together. In fact, recent versions of the theory, as Dennett portrays them (1994a), accept the hypothesis that, at the “input periphery” of our cognitive systems, sensorial input is fragmented into a series of “micro-takings of information” distributed across the brain. The problem is that it still assumes a line of consciousness, which some information crosses (somewhere between the input periphery and the output periphery), to be compiled and taken in by a central, homuncular subject/agent.

3.4.2.3. The brain as a syntactic engine approximating a semantic engine

In his very first book, Dennett remarks that brains discriminate stimuli by their environmental significance. But «since environmental significance [...] is not an intrinsic physical characteristic, the brain, as a physical organ, cannot sort by significance by employing any physical tests.» (1969, p.48) This is one of the earliest expression of the insight that the brain is a syntactic engine.
approximating a semantic engine, an insight which, in his “Self-Portrait” (1998 [1994]), he designates as “the first stable conclusion” that he reached:

The first stable conclusion I reached [...] was that the only things brains could do was to approximate the responsivity to meanings that we presuppose in our everyday mentalistic discourse. When mechanical push came to shove, a brain was always going to do what it was caused to do by current, local, mechanical circumstances, whatever it ought to do [...] but over the long haul, brains could be designed — by evolutionary processes — to do the right thing [...] The appreciation of meanings — their discrimination and delectation — is central to our vision of consciousness, but this conviction that I, on the inside, deal directly with meanings turns out to be something rather like a benign “user-illusion”. [...] This vision tied in beautifully with a doctrine of Quine’s that I had actually vehemently resisted [...] the indeterminacy of radical translation. (1998 [1994], p.357)

Humans, then, are responsive to meaning, i.e. we discriminate things by meaning. This suggests that brains are “semantic engines”, but (according to Dennett) it is impossible that they be such. Rather, brains are “syntactic engines”, functionally organized and mechanistically driven, behaving as semantic engines.

Between our semantic activity (language production, understanding, interpretation of behavior) and the syntactic activity that produces it, there is an indeterminacy — precisely what Quine described as the indeterminacy of radical translation. Contrarily to what appears in the Cosmic Pyramid, the gap between meaning and the neurological machinery that supports it is not metaphysical.
From the point of view of the intentional stance, we deal with propositional content. Accounting for this propositional content in functional terms, as we must, demands that we acknowledge the indeterminacy of a syntactic/semantic approximation — which confirms the irreducibility of the intentional level.

The brain, as intentional system theory and evolutionary biology shows us, is a *semantic engine*; its task is to discover what its multifarious inputs mean, to discriminate them by their significance and “act accordingly”. That’s what brains *are for*. But the brain, as physiology or plain common sense shows us, is just a *syntactic engine*; all it can do is discriminate its inputs by their structural, temporal, and physical features and let its entirely mechanical activities be governed by these "syntactic" features of its inputs. That’s all brains *can do*. Now how does the brain manage to get semantics from syntax? It couldn’t. The syntax of a system doesn’t determine its semantics. (Dennett 1987, p.61)

Yet, as Dennett points out, the brain does extract, from syntactically driven operations, «semantically reliable results»:

[the brain] could be designed to *approximate* the impossible task, to *mimic* the behavior of the impossible object (the semantic engine) by capitalizing on close (close enough) fortuitous correspondences between structural regularities — of the environment and of its own states and operations — and semantic types. (ibid)

The “structural regularities” involved result from evolution — they are not simply claimed by a doctrine of pre-established harmony. This phenomenon is nothing more than the application to the human case of a phenomenon we have
already encountered in the discussion of tactical behavior in animals (3.2.3.2.), and of the capacity, that certain species are endowed with, to somehow distinguish between “getting it right” and “getting it wrong”. About those species that are endowed with such a capacity, Dennett says that «[t]hey are the beneficiaries of equipment exquisitely designed to get what matters right [...]» (2003, p.165). Meaning, then, is explainable as the result of the syntactical activity of functional (and ultimately mechanistically driven) systems, exquisitely designed by evolution to “get things right”.

3.4.2.4. A model of language production

Dennett remarks that speech production has been investigated much less than most other features of linguistic competence:

One of the skeletons in the closet of contemporary linguistics is that it has lavished attention on hearing but largely ignored speaking, which one might say was roughly half of language, and the most important half at that. Although there are many detailed theories and models of language perception, and of the comprehension of heard sentences (the paths, from phonology, through syntax, to semantics and pragmatics), no one — not Noam Chomsky, and not any of his rivals or followers — has had anything very substantial (right or wrong) to say about systems of language production. (1992, p.231)
In this context of scarcity, what we have to fall back on are Dennett’s insights about the general principle of generation-and-selection. Thus, language production involves the generation of many options of what to say according to the circumstances. Multiple layers of phonetic, syntactic, semantic and pragmatic systems assemble many options of candidate phrases, i.e. “things to say”. At the same time, other systems constitute multiple layers of filters, selecting the sentences that are actually uttered. Moreover, Dennett refers to various connectionist models of “constraint satisfaction processes” as examples of models that could eventually deliver convincing accounts of language production. To what extent they may be successful or not remains an empirical question, but, Dennett insists, we can be sure that an adequate account will include a basic schema of generation/selection, akin to the evolutionary process: «We know, however, that somewhere in any successful model of language production we must avail ourselves of an evolutionary process of message generation, since otherwise we will be stuck with a miracle (“And then a miracle occurs”) or an infinite regress of Meaners to set the task.» (1991, p.239)

Note that the selection of one “thing to say” is a phenomenon distinct from the entertaining of propositional attitudes. The fact that one phrase candidate gets selected does not imply realism about propositional attitudes.

The elimination of the Conscious Active Self, in its guise as Central Meaner, eliminates by the same token the idea of “original intentionality”. Intentionality, Dennett argues, is distributed across the hierarchies of nested
functional structures. At the ground level of these structures, simple neurons exhibit the simplest form of intentionality. As they constitute the nested functional units, the intentionality gets more complex.

3.4.2.5. A model of understanding

Dennett’s model of understanding is based on the same principles, and it stands in opposition to the view that human understanding must consist in some “original understanding”, performed by the central entity that is the Conscious Active Self (in its role as the Central Understander). Theories that do not eliminate the Conscious Active Self partake, Dennett suggests, in some Cartesian view, either the “Cartesian materialism” that he challenges, or the old Cartesian dualism that few philosophers adhere to, but which still casts its shadow on contemporary philosophy of mind. The following passage is part of Dennett’s critique of John Searle’s Chinese Room experiment, which defends the idea of “original understanding”. According to Dennett, Searle, as much as Cartesian dualists, commits the fallacy of the sorites (which, as we have seen in 3.1.1., is linked to essentialist thinking and to the postulation of skyhooks):

Cartesian dualists […] think that even human brains are unable to accomplish understanding all by themselves; according to the Cartesian view, it takes an immaterial soul to pull off the miracle of understanding. If, on the other hand, we are materialists who are convinced that one way or another our brains are responsible on their
own, without miraculous assistance, for our understanding, we must admit that genuine understanding is somehow achieved by a process composed of interactions between a host of subsystems none of which understand a thing by themselves. The argument that begins as “this little bit of brain activity doesn't understand Chinese, and neither this bigger bit of which it is a part…” is headed for the unwanted conclusion that even the activity of the whole brain is insufficient to account for understanding Chinese. It is hard to imagine how “just more of the same” could add up to understanding, but we have very good reason to believe that it does, so in this case, we should try harder, not give up. (1991, p.438-439)

What we must “try harder” to do is of course to adopt a gradualist, skyhook-free perspective, from which we can see «how understanding could be a property that emerges from lots of distributed quasi-understanding in a large system.» (1991, p.439) Talk of understanding as an “emerging property” has no connotation of metaphysical specialness (or of mystery) within the sort of empirically based, gradualist view that we have surveyed in this section. Now, «Is it “cheating”», Dennett rhetorically asks, «to think of the software as composed of homunculi who quasi-understand, or is that just the right crutch to help the imagination make sense of astronomical complexity?» (ibid) Dennett is referring to the heuristic indispensability of homuncular functionalism: without functionalist thinking (notably the positing of nested hierarchies of homunculi), empirical neurophysiological research cannot, by itself, analyze the distributed cognitive micro-tasks involved in understanding.
3.4.3. Language-produced phenomena

We now come to the linguistic extension proper, which examines how the development of language introduced families of concepts that had no prelinguistic forerunners, and thus opened new cognitive areas, initiated new practices, and brought about new phenomena and items. Before we get to the two linguistic phenomena that are of particular importance for the purposes of the present dissertation (the self and rational criticism), let us clarify the very idea of “phenomena brought about by language”, or “language-produced phenomena”\(^73\).

In *Empiricism and the Philosophy of Mind*, Sellars argues that the origins of concepts is first and foremost in the needs and interests of human beings, not in experience, either empirical or introspective. Sellars’ “myth of Jones” is a long argument to the effect that a large part of our vocabulary consists in concepts originally introduced as theoretical posits, in order to satisfy pragmatic needs.

\(^73\) Culture is, of course, another important creation of language. However, the present study is not concerned with the “nature” of culture, nor with the processes by which it is reproduced and transformed. Dennett endorses, and develops, a theory of culture initially proposed by Richard Dawkins, according to which culture is constituted by those supposedly self-replicating units that he famously called “memes”. I find meme theory problematic in many regards, but, for the purposes of the present thesis, it can be set aside.
Their status as theoretical posits implies three things. First, although Sellars himself favored the realist interpretation of theoretical terms, one may remain ontologically neutral in regard to posits. Second, the rules of their use constitute a sort of theory, or partake in such a theory. Third, it follows from the preceding point that originally, the use of these concepts must have been inferential, as their application implied the recourse to a theory. However, as the use of these concepts became familiar, it also became noninferential. From then on, the theory that includes the rules for the use of concepts originally introduced as theoretical terms is not consulted in their routine, noninferential use. It still underlies it, however, and it is consulted when, in the practice of giving and asking for questions, we assess a particular instance of the use of these concepts, and so the cognitive status of the judgments in which it occurs.

Sellars argues that two concepts of “inner episodes” (“thoughts” and “sense impressions”) were introduced as such theoretical posits, because they allowed speakers to get a better grip on what other agents were up to. Now, according to a “good reasons” approach to ontology, it is possible to argue that from the moment that the basic, ordinary use of a concept is noninferential, and that this noninferential use typically elicits wide agreement, we have a good reason to consider that it refers to something “real”. But the gist of Sellars’ insight is that consciousness and mental states of “inner episodes” are not mental natural kinds which humans introspectively discovered, or which simply popped in their mind, independently of concepts. Rather, consciousness and conscious
mental states were attributed by using concepts, and the origins of this attributive practice are in certain human needs.

Dennett basically endorses the general idea that Sellars introduced, as well as similar insights proposed by psychologist Julian Jaynes (Jaynes 1976), but he brings some precisions and qualifications of his own. First of all, Dennett clearly states that many of our concepts did not create the phenomena and items they designate, although changes in the use of concepts often modifies our relationship to the items they designate:

Our concepts of diseases and earthquakes have also undergone substantial revision over the last few hundred years, but *diseases and earthquakes are phenomena that are largely (but not entirely) independent of our concepts of them*. Changing our minds about diseases did not in itself make diseases disappear or become less frequent, although it did result in changes in medicine and public health that radically altered the occurrence patterns of diseases. Earthquakes may someday similarly come under some measure of human control, or at least prediction, but by and large the existence of earthquakes is unaffected by our attitudes toward them or concepts of them. (1991, p.24; emphasis added)

As examples of phenomena and items that cannot pre-date their concept, Dennett mentions romantic love (1991, p.23-24), as well as history and money (1998 [1986], p.128). Those phenomena that are of greatest philosophical importance, however, are the self, as the “center of narrative gravity” of an individual’s biography; and rational criticism, as the practice of giving and asking for reasons in support of judgments, decisions and conduct. In regard to
Dennett’s moral anthropology, these two concepts, and the phenomena that they brought about (and designate), provide the connection between “human nature” on the one hand, and responsibility and free will on the other.

3.4.4. The self

Dennett’s overall conception of the self comprises three elements: a self is a biological individual (separated from what is not itself by a physical border), who enjoys an expandable measure of auto-control (or “self-control”), and who has a biography.

We have already encountered the biological self in 3.2.1. The notion has an abstract sense, which designates a principle of biological organization; and a concrete sense, which designates the biological individual, distinguished from everything on the “outside” — everything that is not itself — by boundaries of some sort. Let us see how the two other elements are related to the biological self.

3.4.4.1. The self as locus of auto-control and center of narrative gravity
The biological self, enclosed within its boundaries, has an interest in, and a capacity for, self-control. It is, Dennett tells us, a locus of self-control, the human version of which he evokes as follows:

A self is, above all, a locus of self-control. How do you know, when you look in a mirror, that the person whose reflection you see is you? Not just by noting the resemblance, for how do you know what you look like right now? You determine that you are looking at yourself by using the same tests you use to determine that you are controlling the model airplane. You see if you control the motions of the person you see. Can you wave that hand? If you can, it’s yours. The epistemological problem of such self-identification is normally trivial, but in contrived cases it can be difficult. In a suitably disguised and complicated tangle of arms and legs — say, a group wrestling match in identical costumes and gloves — the question “Which of these hands are mine?” can become an occasion for careful experimentation and reflection. Control is the ultimate criterion: I am the sum total of the parts I control directly. (1984, p.81-82; emphasis added)

The self, then, designates agents as biologically organized entities with an ability to directly control themselves.

Now, consider what a biography is. It is constituted by the narration of all the events that have made up the existence of an individual, from its beginning to the current point. There is no unique, canonical narrative, but rather an indeterminate number of descriptions of series of events in which the individual is involved, as their author or as one of their protagonists. There is no necessary unity to this indeterminate number of descriptions: an individual’s life is an inexhaustible source of information, and can lend itself to infinitely many
descriptions. But, Dennett claims, we spontaneously make such a unity, and we give it tractable compactness, by finding, in the intractable mass of biographical information, the equivalent of the center of gravity that we find in a physical system.

We have already encountered centers of gravity in section 3.1. They are one of Dennett’s favorite examples of what Reichenbach called abstracta. A center of gravity is an immaterial, abstract point, but its position within a system is objective, and can be calculated, for the purpose of making the system predictable. Its ontological status is undetermined: although it can be described as “a theorist’s fiction”, it nevertheless has a certain objectivity, since its location is a relatively precise slice of time-space (so it is not a mere projection). “Attributing” a center of gravity to a system is answerable to standards of correctness set in the practice of giving and asking for reasons, and predictive success is the most significant property of well-groundedness that a particular center of gravity “attribution” can exhibit.

Dennett, then, proposes the following analogy: as we find a center of gravity in a physical system in order to make it predictable, so we find a center of “narrative gravity” in the biography of an individual, in order to make that individual informationally tractable, (minimally) predictable, and so intelligible (or, maybe more exactly, it enhances informational tractability, predictability, and intelligibility). That is, circumscribing a center of narrative gravity in an individual’s biography fosters psychological and moral identities to individuals (including oneself). It constitutes a coherent entity out of what would otherwise remain
multiple streams of information. This center of narrative gravity is the biographical (or “narrative”) self.

These strings or streams of narrative issue as if from a single source — not just in the obvious physical sense of flowing from just one mouth, or one pencil or pen, but in a more subtle sense: their effect on any audience is to encourage them to (try to) posit a unified agent whose words they are, about whom they are: in short, to posit a center of narrative gravity. Physicists appreciate the enormous simplification you get when you posit a center of gravity for an object, a single point relative to which all gravitational forces may be calculated. We heterophenomenologists appreciate the enormous simplification you get when you posit a center of narrative gravity for a narrative-spinning human body. Like the biological self, this psychological or narrative self is yet another abstraction, not a thing in the brain, but still a remarkably robust and almost tangible attractor of properties, the “owner of record” of whatever items and features are lying about unclaimed. (1991, p.418)\(^7\)

Note that enhancing informational tractability, predictive power and intelligibility of other agents partakes in the interest for control of one’s environment. The same applies to oneself. We will see in Chapter Four how self-

\(^7\) There are also obvious differences between the two concepts (which Dennett does not mention). A physical center of gravity is a point whose location requires calculation. As a concept, it is relatively sophisticated. By contrast, finding the center of narrative gravity, as the core of a person’s biography, is something that we do naturally and, to some extent, unthinkingly.
control can develop into free will when one takes responsibility for one’s decisions and conduct.

3.4.4.2. Selves, notional worlds, animals

We have seen in section 3.1. that propositional attitudes are attributed to intentional systems to make them minimally predictable, and so intelligible. In the case of human beings, propositional attitudes can be attributed in a transient way; but they can also be attributed in a “cumulative way”, so that the propositional attitudes attributed to an individual partake in a “notional world”, attributed to that individual.

Dennett does not give any indication as to the relationship between notional worlds and biographical selves. As I read him, we can, for all practical and theoretical purposes, assimilate the notional world of a person to the biography of that person. So the biographical self can also be thought of as the center of narrative gravity of a person’s notional world.

The notion of biographical self is not a priori restricted to humans. An animal can also be seen as having a biography, and so as having a biographical self. Racehorses, or oxen used as breeders, might have quite elaborate biographies. There is no more point in attributing a biographical self to a rat, than there is in attributing beliefs and desires to a thermostat. But the pointlessness of
such exercises comes from practical concerns, not from metaphysical matters of fact. Moreover, those animals who (as we have seen in 3.4.3.) hold “track records” of other band members are themselves performing the primitive, animal version of the attribution of biographical selves.

The biographical self, then, is not a metaphysically special item: «Since selves and minds and even consciousness itself are biological products (not elements to be found in the periodic table of chemistry), we should expect that the transitions between them and the phenomena that are not them should be gradual, contentious, gerrymandered.» (1991, p.421)

Again, we will see in Chapter Four what distinguishes humans from animals.

3.4.5. Rational criticism: the practice of giving and asking for reasons

At this point, we can see what a Dennettian, naturalist version of what I called the “underdetermination thesis” looks like.

As we have seen, there are reasons why animals do what they do, of which they are of course unaware. Similarly, our hominid ancestors adopted certain attitudes and patterns of behavior for practical, self-interested reasons they were not aware of. By contrast, language-using human beings do not merely experience sensorial impressions and sentiments, and act upon decisions produced by the “choice machines” that they are equipped with. Humans
linguistically articulate the content of their experience, their dispositions and the responses that manifest them, the considerations they generate, and the processes by which they select some of those and reject others. In other words, they make judgments, deliberate, make self-conscious decisions, and act accordingly. From the moment that their experience, dispositions, decision-making, and action are linguistically articulated, they can give and ask for reasons in support of their judgments, decisions and actions.

The practice of giving and asking for reasons in support of decision, action and judgment — typically in the form of “why questions” — has opened what Sellars famously called “the logical space of reasons”. Within the space of reasons, language-using human beings can endeavor to appreciate some reasons in support of their judgments, experience and action, and endorse and act upon those reasons which they take to be good. In the practice of giving and asking for reasons, then, human beings constitute self-conscious standards of correctness, to which judgments, decisions and actions are answerable. In sum, language brings about a theoretical form of rationality: it gives to humans the capacity to examine and to question the reasons that purportedly support their judgment, decision and action.

The most salient feature of the question, for the purposes of the present dissertation, concerns how the practice of giving and asking for reasons induces in all forms of human judgment and action a dimension of endorsement which is properly moral. Let us examine Dennett’s treatment of this insight.
Kant famously claimed: «Everything in nature works according to laws. Only a rational being has the capacity of acting according to the conception of laws (i.e., according to principles).» (Groundwork 412)

Now, let us compare Dennett:

Once we have captured the free-floating rationales of the natural motives and represented them, […], we are no longer bound by the inefficient, wasteful, mindless trial and error of natural selection. We can hope to replace an equilibrium of sheer replicative power with a reflective equilibrium of rational agents who have engaged in the communal activity of mutual persuasion. This shift from undirected trial and error to intelligent (re-)design is […] a major transition in evolutionary history […] (2003, p.267)

Although Dennett’s overall framework is not Kantian, the naturalistic examination of rational criticism and its effects can only resonate with Kant’s famous thought about how rational beings act according to the conception of laws.

The capacity to examine the reasons that purportedly support judgments and actions is a capacity to reject or endorse those reasons. As humans assess reasons for acting and reasons for believing as more or less good or bad, they can modify their beliefs, decisions and conduct consequently. It follows that the judgments we hold, the decisions we make and the actions we perform, and the reasons why we hold, make and perform them, are endorsed by us, at least implicitly.
Thus, even in those cases where language-using humans act like their evolutionary ancestors, and adopt attitudes similar to those that the ancestors adopted; and even when they act in these ways and adopt these attitudes for the same (practical, self-interested) reasons that their ancestors had, these reasons are endorsed, given the fact that they have the capacity to examine them critically. Moreover, language-using humans can act in similar ways, adopt similar attitudes, but for reasons different from those of our prelinguistic ancestors. Notably, language-using humans may engage in the protection of cooperators and the repression of defectors, not in reason of the beneficial consequences that this strategy brings about, but for the reason that cooperating is valued in itself, and defecting condemned in itself. Finally, language-using humans can act in ways that their ancestors never did, and adopt attitudes that their ancestors never adopted. Most humans institutions are not directly related to the mores of early Homo sapiens.

In any case, whether they endorse the reasons that correspond to evolutionary rationales, or whether they give themselves entirely new reasons for doing certain things and believing certain things, humans give themselves their own reasons to believe what they believe, and to do what they do; and they have the power to choose their own finalities.

As Dennett puts it, we have the capacity to capture the reasons why we do what we do, to bring them under critical scrutiny, and so to make them our own: endorsing them, amending them, or rejecting them. As language makes the
workings of the primitive dispositions describable, but also disputable, we can consider our dispositions critically. Humans can tamper with the expression of their natural dispositions. To a certain extent, human beings can re-design themselves by the practice of rational criticism, and by the traces it leaves in culture, and in language itself, over the generations.

In the end, the standards of correctness to which we take our judgments to be answerable emerge from the practice of giving and asking for reasons. Although those standards plausibly bear the mark of the primitive purposes, or functions, of the various areas of discourse, they are not determined by them. This is, as Dennett puts it, how humans “bootstrapped themselves free” (2003, p.259). That is, the dimension of endorsement brought about by our capacity to examine reasons makes us responsible, in the moral sense, for our judgments, decisions and actions. And as we will see in Chapter Four, free will is nothing but the maximization of self-control allowed by the attitude of taking responsibility for one’s judgments, decisions and actions. Rational criticism, then, has emancipated our natural dispositions and abilities from the consequentialist rationales and the self-interested practical thinking that accounts for the emergence of their proto-moral form.

In virtue of the dimension of endorsement and responsibility it induces, rational criticism brings about morality (and values in general). Thus, it transforms the proto-moral dispositions (and the proto-moral behavior and attitudes that
express them) that are part of our evolutionary heritage into genuinely moral dispositions. Mere instinctual responses are transformed into attributions of moral properties, genuinely *moral* judgments, sentiments, decisions and actions, answerable to standards of correctness.

Whether we manifest the same cognitive, emotional and behavioral patterns as our ancestors, and so express the same dispositions that they had, or whether we act differently, the fact that we have the capacity to offer reasons in support of our actions and judgments implies that we endorse the reasons why we do what we do, and hold the judgments we hold; and that has introduced a moral dimension in our relationship to ourselves, to others, and to the world.

The transformation operated by language and rational criticism, then, is not only cognitive; it is also moral. That is, language and rational criticism have transformed our proto-moral dispositions, both natural and artificial, into genuinely moral dispositions, whose manifestations are answerable to moral standards. The space of reasons is not merely a space of epistemic grounding, but also a space of responsibility, freedom and morality.
CHAPTER 4. THE PERSONAL STANCE: DENNETT’S PHILOSOPHY OF MORALITY

I divide Dennett’s philosophy of morality in two parts. What I call the “genealogical” part (4.1.) presents Dennett’s take on how Hobbes, Hume and Nietzsche conceived of the emergence of morality. Dennett emphasizes what he thinks those philosophers saw correctly. His comments on those three philosophers serves as a transition between his evolutionary theory of our given cum acquired nature, and his anthropology of moral life per se (4.2.)

4.1. THE GENEALOGY OF MORAL LIFE: THE “BIRTH OF VALUES” INSIGHT

We have seen in 2.2. that Hume adhered to the Hobbesian sort of theoretical reconstruction, which proposes to identify various implications of human sociality, given the dispositional constitution of the human species; and we have seen that Hume’s own version of this reconstruction is centrally involved in his theory of artificial virtues, as well as (according to Wiggins) the intersubjectivist strand of Hume’s analysis of morality (according to which
morality reaches beyond the sentiment of benevolence which originally gave rise to it.)

We have also seen in 3.5.3. that, according to what I have called “Dennett’s linguistic extension” of evolutionary theory, certain phenomena cannot pre-date their concepts, but have been brought about by language. According to Dennett, morality is one such phenomenon (or set of phenomena). It is the coining of moral concepts, their use in judgments, and their discussion in the practice of giving and asking for reasons that brought about values and moral phenomena\textsuperscript{75}. There were no values before there was a possibility of giving and asking for reasons. Values, then, are not ahistorical abstract entities.

Dennett synthesizes those insights as he remarks that Hobbes, Hume and Nietzsche argued, each in his own way and with his own emphases, that morality (and values in general) is a historical phenomenon, brought about by society, language and culture. More precisely, Dennett finds in the respective works of those three philosophers the following insights.

Hobbes suggests, as part of his theoretical reconstruction of the origins of sociality, that “the covenant” (i.e. the social contract), in combination with other phenomena — notably rational criticism — brought about values (and morality in particular).

\textsuperscript{75} The term “values” is understood broadly here (I take it to be equivalent to the expression “moral properties”, and to include moral responsibility).
Hume might not have thought that values as such were brought about by language and sociality, but he stressed that the practice of giving and asking for reasons brought about the development of morality way beyond anything prelinguistic that we could recognize as “morality”.

Nietzsche insisted on the fact that morality is an essentially human phenomenon, undetermined by natural factors related to the evolution of species.

The main point that Dennett wants to establish, and that the three aforementioned philosophers (each in his own way and with emphases of his own) foresaw, concerns the historical setting of the underdetermination thesis. Whereas the cognitive and emotional responses of animals, together with their behavior, are determined by their dispositional constitution, directed as it is at the satisfaction of a means-end hierarchy, the decisions and the conduct of human beings are not so determined. Arguably, they are conditioned and constrained by the dispositional constitution of our species, but not determined. From the moment that language-using humans could assess the reasons why they acted as they did, and the reasons why they held the judgments they did, they appropriated those reasons.

Moreover, humans started to offer, in support of their judgments and actions, reasons which did not partake in the sort of means-end rationality that governed the development of our ancestors — which can be translated, in evolutionary terms, into the insight that humans could therefore offer reasons that were not drawn from the free-floating rationale which had governed the
development of our evolutionary ancestors, up to the point where Homo sapiens started to use language. Thus, humans started to hold judgments, to make decisions and to conduct themselves for new reasons, which were entirely their own.

Dennett does not engage in a close reading of any of the three aforementioned philosophers, he merely brings out what he takes to be the main thrust of their respective formulations of the “birth of values” insight. In spite of the brevity of Dennett’s historical excursions, their content is worth summarizing.

4.1.1. Hobbes: the covenant and rational criticism

The idea that values were brought about by rational criticism and by the covenant originates in Hobbes, who emphasizes that human sociality entails the self-conscious adoption of certain norms, since some attitudes and behavior have consequences that are socially desirable, others have consequences that are socially undesirable. Escaping the “state of nature” requires that we promote the former (and reward meritorious members of the community), and prohibit the latter (and punish the culprits), and this in turn requires a covenant. Thus, for Hobbes, values do not pre-date human sociality:

Once upon a time, [Hobbes] said, there was no right and wrong, and people lived in “the state of nature” and engaged willy-nilly in the war
of all against all […] Then it happened that some of them came together and formed a covenant, and thus began society, and with it, *right and wrong came into existence*. Not just the ideas or concepts of right and wrong in the people’s minds, but right and wrong themselves. Before the covenant nothing was morally right or wrong. It is not that when people did evil or good things they had no inkling of how evil or good they were. Nothing was evil or good — in a moral sense. For people lived then the way lions and tigers live now in the state of nature, and surely our intuitions do support the idea that there simply are no moral facts about the relationships between wild animals. The covenant changed all that, by bringing into existence a different way of life, with *new* reasons for acting, just as the emergence of replicators brought into existence the first way of life, and with it the first sort of reasons for acting. (1984, p.44)

Dennett acknowledges, of course, that Hobbes’ point does not pertain to an actual historiography of humankind, in the sense that it does not conjure up an actual historical event (consisting of people drawing some sort of collective agreement). Still, Hobbes’ point is historical: it does refer to an actual period in prehistory when a certain practice emerged (namely the practice of giving and asking for reasons), which brought about values.

For Hobbes, the covenant and rational criticism (notably the critical examination of what is desirable and what is undesirable) mutually imply one another. No doubt, many distinct but related phenomena and practices emerged together, and their mutual interactions are practically intractable. What Dennett emphasizes, from among the cluster of Hobbes’ genealogical insights, is the importance of *rational criticism* for the birth of values:
The emergence of social order, language, consciousness, and morality was surely a tangle of interpenetrating and interacting developments for which Hobbes’ account, considered as a “model”, would be simplistic in the extreme. But Hobbes’ story distills and isolates for thought one important strand: the social practices of verbal advocacy, mutual criticism and debate not only permitted the creation of a new relationship between creatures and their reasons (the relationship of considering those reasons); it also permitted the creation of new reasons to consider. (1984, p.45; emphasis added)

In the last line of the quote above, Dennett attributes to Hobbes an insight that latches onto the linguistic extension of evolutionary theory. Dennett credits Hume with a similar insight, as we will now see.

4.1.2. Hume: natural inclinations and rational criticism

Dennett considers that, by and large, Hume’s account of natural and artificial virtues is sound. He suggests that Hume’s overall theory of morality, provided a few modifications, can be readily integrated to evolutionary theory cum its linguistic extension. The theory of morality in The Treatise of Human Nature is characterized as follows by the author of Freedom Evolves:

According to David Hume, we begin with what he calls the natural motives: sexual appetite, affection for children, limited benevolence, interest, and resentment — a list that twenty-first century evolutionary
psychologists would look on with favor. These dispositions have rationales that are not our rationales, though they have set the stage for our practice of demanding and giving reasons. [...] From the outset we find ourselves approving of some attitudes and practices — as “intrinsically” good somehow — and these attitudes and practices were shaped over millennia without foresighted design but with their own raison d’être. [...] The natural motives, Hume maintained, have “offspring”, which he dubbed the “artificial” virtues of morality — such as justice. Hume [...] saw reflection as the tool we get from nature that permits us to revise our natural instincts, enhancing them with prosthetic elaborations whose rationales (free-floating until Hume and others pinned them down and represented them) are, in fact, aimed at still more freedom, consistent with security from harm. (2003, p.260-261; emphasis added)

The main point on which Dennett disagrees with Hume is whether it is better to reserve the term of “morality” for the developments brought about by language and rational criticism (in which case those dispositions and patterns of response of pre-linguistic Homo sapiens which gave rise to morality will be characterized as proto-moral); or whether we may characterize those pre-linguistic dispositions and patterns of response as “moral”. Hume may be interpreted as suggesting the latter, as he claims that natural virtues and vices were recognized as such by humans in their “uncultivated state”. I have interpreted Dennett as adopting the other option and, speaking in his name, I have characterized as “proto-moral” those dispositions and patterns of response of pre-linguistic Homo sapiens which gave rise, with the development of language and rational criticism, to genuine morality. The proto-moral dispositions that gave
rise to natural virtues and vices may be more primitive than those that gave rise to artificial virtues and vices, but all became genuinely moral with the advent of language and rational criticism, not before.

This, however, is obviously a matter of terminological precision, and not a substantive disagreement. Dennett emphasizes that Hume rightly saw (like Hobbes before him) that rational criticism allows human beings to endorse, in support of their conduct and judgment, reasons that are their own, and not those of nature. We may attribute to Hume, then, a pre-evolutionary version of the view according to which humans have the capacity to examine and to reject the reasons on the basis of which they act and hold certain judgments, even those that stem directly from the free-floating rationales which tightly constrained the behavior of our evolutionary ancestors. Dennett suggests that Hume would have agreed that we *endorse* those reasons upon which we hold certain judgments, and act. These are *our* reasons, not those of nature. Neither natural dispositions, nor our essential sociality, *determine* the standards of correctness by reference to which we assess our moral responses and judgments. Our natural dispositions certainly create natural attractors in the space of reasons; but even in a deterministic universe, there is a crucially significant sense in which it is *up to us* to follow the natural attraction, or to resist it. In some instances, we might even give ourselves standards of correctness that are at odds with our natural dispositions, which we then decide to “revise”, in the sense that we set conditions of acceptability upon them.
We may say that our responses and judgments are genuinely moral to the extent that they are answerable to rationally endorsed standards of correctness. From the moment that we are engaged in the practice of rational criticism, it is that very practice, not our natural dispositions, which ultimately set the standards of correctness to which moral responses and judgments are answerable, and those standards are adequate.

4.1.3. Nietzsche: the autonomy of values

Dennett commends Nietzsche’s insistence on the autonomy of our moral standards from our evolutionary origins and our natural condition; that is, the autonomy that human concerns and practices, culture and history, conferred upon the domain of values: «the upshot of Nietzsche’s genealogy of morals is that we must be extremely careful not to read into the history we extrapolate from nature any simplistic conclusions about value.» (1995, p.466)

Even though we may endorse our natural dispositions, and the reasons our evolutionary ancestors had for acting in certain ways; even though evolutionary rationales may be offered as reasons in support of certain judgments, and as counting against others, it remains that there is no direct derivation from some evolutionary rationale to some standards of correctness. Nietzsche might not accord due importance to rational criticism, to which (according to Dennett) values owe their autonomy, and the moral level of
discourse its irreducibility. But he is right to claim that human beings make their own values. His criticism hits at once sociobiology and utilitarianism (and consequentialist theories in general): no order of values is derivable from what we take to be the desirable or undesirable consequences of certain actions, any more than from the natural dispositions that we have inherited from evolution (on the count that they provide adaptive advantages). Human standards of correctness enjoy an autonomy such that they are neither derivable from evolutionary rationales, nor from the calculation and anticipation of desirable or undesirable consequences. We human beings can consider the consequences of certain actions, and the adaptive advantages provided by certain dispositions, and take them as reasons counting for or against certain judgments (evaluative or practical). We can acknowledge the influence of our natural dispositions and of our sociality on our thinking. But natural and social facts do not directly determine the standards of correctness to which our responses and judgments are answerable. Rather, we constitute them in the practice of giving and asking for reasons.

Dennett, then, approves of Nietzsche’s emphasis on the historical character of values, presented notably in the Genealogy of Morals (which Dennett describes as a Just So Story). As Dennett remarks, Nietzsche’s position goes against the simplistic sociobiology and “evolutionary ethics” of the “Social Darwinians” (notably Herbert Spencer, and Paul Rée), who thought it was
possible to directly infer an order of values from certain traits of the evolution of species (as they understood it).

However, Nietzsche was wrong to think that his rebuttal of the Social Darwinians refuted Darwin’s theory itself (which, Dennett remarks, he probably never read). Once we have seen that it is rational criticism which gives birth to values, that this capacity (like all our rational capacities) results from the evolutionary process, and that it was added to an evolutionary heritage of natural dispositions that it can call into doubt but that it does not obliterate; once we have seen that, we understand that the Nietzschean, historicist insight is fully compatible with Darwin’s theory of evolution.

Like Nietzsche, Dennett stresses the autonomy of values, and consequently opposes the crudest forms of sociobiology (which, he insists, are a theoretical development distinct from Darwin’s theory). Sociobiologists are right to emphasize that evolutionary rationales tend to influence human decisions, and that we must study the origins and the workings of that influence. But they do not acknowledge, as they should, that rational criticism is a capacity that can «overthrow innate preferences»:

So although we arrive on this planet with a built-in, biologically endorsed set of biases, although we innately prefer certain states of affairs to others, we can nevertheless build lives from this base that overthrow those innate preferences. We can tame and rescind and (if need be) repress those preferences in favor of “higher” preferences, which are no less real for not being directly biologically (that is, genetically) endorsed.[…]
It is our communal activity of mutual persuasion, reflection, and evaluation that creates the values that then take precedence over the cruder interests of our ancestors. (1984, p.45-46)

It is from this perspective that we should understand the Nietzschean idea that moral meaningfulness, if there can be such a thing, can only be created by human beings themselves, out of a natural world which, by itself, is devoid of any such thing. It is worth remarking (as Dennett does not) that the very practice which gave birth to values and constituted humans as moral beings (rational criticism) is by the same token a capacity to undermine the same values.

4.2. DENNETT’S ANTHROPOLOGY OF MORAL LIFE

Dennett’s anthropology of moral life articulates, at the specifically moral level of the rationally acceptable language, an account of some essential features of ordinary moral life. The account is partly revisionary, and partly vindicatory.

First, I will present his account of moral responsibility as a deontic status and of free will as the maximization of self-control by the self-attribution of that deontic status (responsibility).76

76 In presenting my reconstruction of Dennett's anthropology of moral life, I will sometimes talk about the “attribution” of responsibility (and of moral personhood), along the three modes of attribution that I distinguished: ego-to-ego, alter-to-ego-, ego-to-alter. It should be clear that it is a deontic status that is being attributed, not a metaphysical property.
Second, I present his heuristics of moral decision-making, which I contend is shaped by the underdetermination thesis. Its pivot is the idea that certain universal maxims are bound to put a stop to our deliberations, and to motivate decision and action. I contend that Dennett’s notion of *conversation-stoppers* is a variant of Wiggins’ “unforsakeable concerns and motives” — a variant which focuses on the role of these concerns and motives in the deliberations of temporally and cognitively limited, but nevertheless committed, agents — which is the case for all human beings. In comparison with Wiggins’ fourfold classification of concerns, motives, etc, and his description of our “first-order ethic of solidarity and reciprocity” (including arguments which purport to justify this double characterization), Dennett’s evocation of "conversation-stoppers" is extremely sketchy. Yet, considering the richness of what Dennett provides as a theory of human nature, and as an overall doctrinal framework, we should make the most of what he offers us in this connection. I also clarify what (minimalistic) metaethical position could be attributed to Dennett, given his pragmatic epistemology and his endorsement of the anthropocentricity thesis.

Third, I present what Dennett has to say about moral meaningfulness. Dennett tackles the issue indirectly, as he comments on a passage of (Taylor 1976) which bears on what Charles Taylor calls “radical re-evaluations”. The Dennettian answer to the challenge of moral meaningfulness is that the issue is necessarily constrained by the deontological, as well as by the practical, material constraints of decision-making.
Dennett suggests that his anthropology of moral life is largely descriptive of our actual common moral lives, and constitutes a vindication of it. The one point about which his anthropology is resolutely revisionary concerns our natural tendency to invoke skyhooks, which he means to bar completely.

4.2.1. Responsibility, free will and decision-making: a compatibilist account

Dennett’s evolutionary and compatibilist conception of moral responsibility and free will can be summarized as follows. Human beings have cognitive capacities which endow them with considerable behavioral and cognitive self-control, and autonomy. “Responsibility” is a deontic status — not a metaphysical property. It is, more precisely, a status of commitment that one attributes along the three modes: ego-to-ego, alter-to-ego, ego-to-alter. We have already encountered, in our discussion of Strawson, the commitment to manifest the minimal regard due to other human beings. Responsibility is an even more fundamental, “primitive” commitment: a commitment to take oneself — and to take others, and to be taken by others — as the author of one’s decisions and performances, and as the owner of the desires and beliefs upon which one decides and acts.
The result of attributing this status to oneself is the maximization of one’s autonomy (that is the rationale that explains the emergence of this practice of self-attribution in the first place). “Free will” is maximized autonomy — not a “special” metaphysical capacity.

Whether determinism is true or not is a metaphysical consideration that has no direct import on that state of affairs. Dennett adopts Van Inwagen’s formulation of determinism: «there is at any instant exactly one physically possible future.» (Van Inwagen, 1983, p.3; quoted in Dennett 2003, p.25). He considers that it is probably true, that it applies to human agency, but that it is not morally significant. Dennett’s position, then, corresponds to the traditional compatibilist conception of responsibility and free will: «one acts freely and responsibly just so long as one does what one decides, based on what one believes and desires.» (1984, p.83; emphasis added).

Dennett’s compatibilist model of decision-making (4.2.1.1.) is built along the lines of his generate-and-test model, which delivers an account of the personal authorship of decision, which he completes (4.2.1.2.) with a view of the ownership of propositional attitudes. Dennett explains (4.2.1.3.) how taking responsibility for one’s propositional attitudes, decision and performances maximizes one’s autonomy; how free will is the maximization of autonomy; and how it is compatible with determinism. He also examines (4.2.1.4.) the concept of personhood, and the conditions of attribution of responsibility, and thus of full, moral personhood. Dennett’s account of responsibility and free will has both
4.2.1.1. Decision-making and the personal authorship of decision

Dennett’s model of human decision-making provides an account of the personal authorship of decisions, including moral decisions. As such, it is an essential joint between his conception of the self and his conception of moral responsibility and free will.

The account of personal authorship of decisions is presented most notably in “On Giving Libertarians What They Say They Want” (Dennett 1978). The title suggests that for Dennett: a) personal authorship is a significant feature of our selves and of our lives that libertarians justifiably want; and b) we have it whether determinism is true or not. Dennett’s text is a direct, critical answer to Wiggins’ “Towards a Reasonable Libertarianism” (Wiggins 1973). I will return to the specifically critical side of Dennett’s model in paragraph 4.2.1.6.
Although the traditional issue of whether reasons for acting are intrinsically motivating (or whether they require a appetitive element), and the related problem of akrasia, are briefly discussed in (Dennett 1984), these issues are not an essential part of his model of decision-making.

Dennett’s model divides up the overall process of human decision-making into three parts: at the outset, considerations are generated; at the end, one consideration is selected; in between deliberation takes place. Following the “reverse engineering” heuristic which characterizes functionalist thinking, we may (Dennett tells us) assign the task of consideration generation to a functional structure that we will call the “consideration generator”, and the task of consideration selection to a functional structure that we will call the “consideration selector”. As for deliberation, we may assign it to a set of functional structures, but what Dennett insists on is the importance of the heuristic rules involved in the process: rules about how to organize the considerations generated, what weight to assign to them, and when to terminate deliberation and select one consideration. Here, the heuristics of deliberation is tackled in a purely abstract way. In 4.2.2., we will see how Dennett considers substantively the termination of the deliberation process in moral decision-making by specific conversation-stopping maxims.

For Dennett, whether the whole decision-making process (consideration generation, deliberation, selection) is determined, or whether non-determined events occur within it, is not a morally significant issue. By contrast, the two
following things are indeed morally significant: personal authorship; and the
intelligibility of human conduct, thanks to the rational patterns it manifests.

We can deal rapidly with the issue of the intelligibility of conduct. For
Dennett, whether the decision-making process is determined, or whether non-
determined events occur within it, *decision-making and human conduct are
intelligible, and to a certain extent predictable, in virtue of the rational patterns
that they manifest*, and so of the rationales that underlie them. i.e. the chains of
reasons that support human decisions, conduct and attitudes. The occurrence of
non-determined events in the decision-making process does not threaten those
rationales, and so the intelligibility and predictability of the decisions and of the
conduct of agents. On this count, Dennett refutes “hard determinism”, according
to which the intelligibility of conduct requires that it be determined.

As for the *personal authorship of decision* (including moral decision), it is
distributed across all the functional structures (or “systems”) and the processes
involved, in particular those that are related to the (sub)task of *selection*. Taking
his cue from writer Paul Valéry, Dennett (1978, p.293) suggests that intelligence
(cognitive fitness) has to do not with the power of one’s “consideration generator”,
so much as with the adequacy of the heuristic rules that one consults in
deliberation, and with the capacity of one’s “selector” to grasp the most adequate
consideration, and to take action consequently.
In the light of his overall theory of the mind (3.4.2.), it should be clear that Dennett’s model of decision-making does not imply the dissolution of the self. Authorship is distributed across many structures and processes which are responsive to education, and so constitute a record of the individual’s personal history (including education). This is the stuff the self is made of, and the “substrate”, so to speak, of moral responsibility and free will (whether the whole process is determined or not).

Dennett points out that the deliberation process and the selection of one consideration largely respond to education (including moral education). Thus, the history of one’s education does much to set the heuristic rules in use in one’s deliberation and decision-making (including moral decision-making). He suggests, then, that a significant part of intelligence lies in our competence in deliberation and in choice selection; and that a large part of the personal authorship of decisions (including moral decisions) lies in the deliberation process and the “consideration selector” — not in an illusory Conscious Active Self (or any of its avatars, e.g. “Central Intender”, “Central Executive”, etc). Finally, a significant part of one’s identity lies in the way one deliberates and takes decisions.

Dennett insists on the genuineness of the personal authorship of moral decisions. First, he claims that «the sense in which we wish to claim authorship of our moral decisions, and hence claim responsibility for them, requires that we view them as products of intelligent invention, and not merely the results of an
assiduous application of formulae.» (1978, p.297) He then asserts that deliberation and consideration selection involve, precisely, the sort of intelligent invention that makes us want to claim authorship for them (they do not come down to an “assiduous application of formulae”). These processes being the stuff the self is made of (and not impersonal algorithms), personal authorship claims are fully justified. Moreover, the processes involved partake in what makes responsibility and free will possible:

Finally, the model I propose points to the multiplicity of decisions that encircle our moral decisions and suggests that in many cases our ultimate decision as to which way to act is less important phenomenologically as a contributor to our sense of free will than the prior decisions affecting our deliberation process itself: the decision, for instance, not to consider any further, to terminate deliberation; or the decision to ignore certain lines of inquiry.

These prior and subsidiary decisions contribute, I think, to our sense of ourselves as responsible free agents, roughly in the following way: I am faced with an important decision to make, and after a certain amount of deliberation, I say to myself: “That’s enough. I’ve considered this matter enough and now I’m going to act”, in the full knowledge that I could have considered further, in the full knowledge that the eventualities may prove that I decided in error, but with the acceptance of responsibility in any case. (1978, p.297; emphasis added)

We have, then, the capacity to make decisions of which we truly are the authors, independently of whether the process is determined or whether undetermined events occur within it. In order to enhance this capacity, however,
we have to adopt a practical attitude: that referred in the last sentence of the quote above, i.e. we have to hold ourselves responsible for our attitudes and performances. In other words, we have to attribute responsibility to ourselves. And, as Strawson would put it, «experiencing it ourselves, we attribute it also to others.» (Strawson 1992 [1983], p.135) We attribute, then, responsibility to persons, along the three modes: ego-to-ego, ego-to-alter, alter-to-ego.

4.2.1.2. Personal ownership of desires and of beliefs

We have already seen, although briefly, what it means to be the owner of one’s own desires, in connection with Strawson’s investigation of our ordinary sense of free, responsible agency. In (Strawson [1983]), Strawson, availing himself to Frankfurt’s notion of higher-order desires, remarks that we experience our higher-order desires as our own, as part of our selves; and that we notably experience in deliberation and decision-making that those desires which have superior status are more closely involved in the definition our self-identity: «That experience [i.e. deliberation — P.C.] heightens our sense of self; in the higher-order desire which determines what we call our choice we identify ourselves more completely; and this is why we call it our choice.» (Strawson 1992 [1983], p.135)

Dennett makes the same point. Deliberating and acting upon second-order desires shapes one’s first-order volitions and desires:
As Frankfurt points out, second-order desires are an empty notion unless one can \textit{act} on them [...] Acting on a second-order desire, doing something to bring it about that one acquires a first-order desire, is acting upon oneself just as one would act on upon another person: one \textit{schools} oneself, one offers oneself persuasions, arguments, threats, bribes, in the hope of inducing oneself to acquire the first-order desire. (1978 [1976], p.284)

As Strawson remarked, this implies that second-order desires are more directly involved in the definition of one’s identity. Consequently, we can take ourselves to be the personal owners of our second-order desires, and of the first-order desires that they shape, just as much as we can take ourselves as the personal authors of our decisions and performances.

Dennett thus establishes the following equivalence: taking oneself as the personal author of one’s decisions and performances, and as the personal owner of one’s desires (in the sense just specified) means \textit{holding oneself responsible} for what one decides, does, and desires; conversely, holding oneself responsible for one’s decisions, performances and desires, means taking oneself as the personal author of the former and the personal owner of the latter.

Dennett makes the same point about beliefs, from the “holding oneself responsible” angle, along the following lines. We have the capacity to notice incoherences in our bodies of belief. Now imagine a case of “implanted belief”: one then acquires a belief independently of normal belief acquisition processes, but does not know that the belief has been thus acquired. However, for having
been thus acquired, an implanted belief is expected to have logical consequences that will contradict other beliefs in one’s overall body of beliefs. Even in the case where we would have beliefs implanted by some foreign agency, we would still have the capacity to notice the resulting incoherences, and to inquire about them. Since one has the capacity to “patrol” one’s body of belief and to investigate incoherences, one has the responsibility of the cognitive maintenance of one’s whole body of beliefs.

It follows that a belief is (in more deontic terms) something that we endorse: «a belief is essentially something that has been endorsed (by commission or omission) by the agent on the basis of its conformity with the rest of his beliefs.» (1973, p.252) We have, then, very good reasons to call our own the beliefs upon which we act, and no good reason not to. We are the owners of our action-guiding beliefs, just as we are the authors of our decisions and actions. This implies moreover the acknowledgement that the consequences of our decisions and actions follow from decisions and performances of which we are the authors. Now, we will see what results from taking responsibility in this sense.

4.2.1.3. Free will: maximizing autonomy by taking responsibility
Moral responsibility is a deontic status that one attributes along the three modes: ego-to-ego, alter-to-ego, and ego-to-alter. Self-attribution of responsibility means, basically, claiming personal authorship for one’s decisions and performances, and claiming ownership for the desires and beliefs upon which one makes decisions and takes action. This enables one to engage in self-criticism (involving the practice of giving and asking for reasons), to actively seek self-improvement, and to enhance one’s self-control.

A crucial trait of self-control enhancement is that it allows one to acquire new capacities, including the capacity to act upon undetermined features of our dispositional constitution which are involved at basic levels of decision and action. One can acquire the capacity to shape one’s own second-order desires (and projects and goals), to hone and improve the heuristics of one’s deliberation, and even to gear “primitive” dispositions (as well as one’s cognitive control systems) to the realization of one’s superior goals and projects (Dennett provides an example that will be presented shortly). As Dennett puts it, the practice of giving and asking for reasons and self-criticism give human beings the possibility to emancipate themselves from natural determinants, «not in any mysterious, “contra-causal”, vitalistic way but by permitting our cognitive control systems to acquire goals that can not be genetically controlled.» (1984, p.45)

The capacity for self-control can thus exert a retroactive influence on its sources: the more we gain self-control, the more we can exercise this capacity on our dispositions, and shape them so as to gain even more self-control. As it acts on basic features of decision and action, self-control enhancement can feed on
itself, following a virtuous circle that has no clear limits. Free will corresponds to the maximization of human self-control and autonomy.

At basis, as long as one acts upon decisions that are one’s own, and on propositional attitudes, goals and projects that are one’s own, one acts freely. This is independent from metaphysical facts about persons, in particular from the question as to whether or not human agency partakes in a deterministic world.77

Although free will, as the maximization of self-control, does not depend upon whether the world is “determined” or “undetermined”, it is, however, directly tied to a distinct contrast (more pragmatic and less metaphysical, although not altogether devoid of metaphysical implications), which Dennett describes as that between “evitability” and “inevitability”. He explicates the latter concept as a version of fatalism, according to which the life of human beings is determined at birth like a ballistic trajectory, so that anyone’s attempts to enjoy genuine self-control, and to improve through self-criticism, are illusory, and morally insignificant. According to Dennett, what matters to us as moral beings is not the issue of determinism so much as the contrast between evitability and inevitability.

Dennett argues that the dreadful predicament of “inevitability” does not obtain, thanks to our free will, i.e. our capacity to enhance self-control and autonomy indefinitely. I mentioned that self-control can act upon undetermined

77 As Dennett reminds us: «The issue is not about determinism […]; the issue is about what we can change whether or not our world is deterministic.» (2003, p.160)
features of our dispositional constitution, including primitive dispositions. Among these are "harm-avoidance dispositions", the most primitive of which are underlain by dedicated functional structures. Dennett remarks that even these are ultimately responsive to self-control, and can be bent to the decisions that we take, and the policies that we adopt. He takes our capacity to act on such primitive harm-avoiding dispositions as an illustration of the fact that our predicament is not that of "inevitability", but on the contrary that of "evitability".

Our open capacity for self-control makes our lives very different indeed from the trajectory of ballistic objects, and very different from the life of animals. Not metaphysically different, but morally different, as morality follows in the stride of the capacities for self-control, rational-criticism, and self-criticism, which characterize human beings. Animals are programmed to avoid harm, and have a capacity for self-control, but it is minimal. Our own capacity for self-control (again, it combines with linguistic and rational capacities) is so great as to allow us to re-design ourselves.

Dennett illustrates this point with a case of a person’s capacity for self-control enrolling very primitive harm-avoidance dispositions so as to make them serve that person’s goals. Imagine, he proposes, that you are playing baseball. You are in the batter position, and the ball just thrown by the pitcher turns out to be directed at your face. It “is bound to” hit you in the face — but you dodge it. This, he claims, is much more than a reflex. It is an instance of what allows us to be very different from the ball itself, i.e. an object following a ballistic strategy:
But in what sense was that baseball “going to” hit you in the face? You dodged it; you were caused to dodge it by the elaborate system evolution has built into you to respond to photons bouncing off incoming missiles on certain trajectories. It was “never really going to” hit you precisely because it caused your avoidance system to go into action. But that avoidance system is more sophisticated than a simple blink reflex, and it can respond to further information, when it is available, and countermand its initial decision. Noticing that you can win the game for your team by being hit by the incoming pitch, you can decide to take the hit. You avoid doing the avoiding that was well within your power — thanks to (caused by) the advance notice you had of the wider context. And you can also avoid avoiding avoiding [sic], when circumstances warrant it. (2003, p.59)78

Moreover, Dennett adds, our genuine capacities to avoid harm and to develop self-control are deployed through unpredictable courses of events. Even if we live in a determinist world, a person’s life remains largely unpredictable, given the complexity of the context in which it unfolds. Although, as the theory of intentional systems makes plain, human agency involves forms of rule-following that make it to some extent predictable (and intelligible at the same time), it remains impossible to predict what course anyone’s life will take, because it is impossible to predict what purchase rational reflection will have on one’s agency, what a person will reap from her capacity to avoid harm, from her capacity for

78 Dennett’s example is weakened (in an interesting way) by the specification that the ball is directed at the batter’s face. The batter can indeed take a hit on the body, so as to earn the first base. But the dodging reflex is practically uncontrollable when objects are directed at the face.
self-criticism and self-improvement, from her encounters with other persons, from the opportunities that come up her way, etc.

Determinism, then, is distinct from inevitability, and does not imply it. Dennett suggests that the truth of determinism would not undermine the genuineness of the potential for self-control and autonomy that the evolution of species has equipped us with, and that we have the power to develop, so as to set us off as far as possible from the predicament of inevitability. By taking responsibility for what we do, we make ourselves avoiders of fatality:

We, unlike the cells that compose us, are not on ballistic trajectories; we are guided missiles, capable of altering course at any point, abandoning goals, switching allegiances, forming cabals and then betraying them, and so forth. For us, it is always decision time, […] no consideration is alien to us, or a foregone conclusion. For this reason, we are constantly faced with social opportunities and dilemmas of the sort for which game theory provides the playing field and the rules of engagement but not the solutions. (1995, p.460)

Again, the practical attitude of holding persons responsible for their deeds and words is genuinely moral. As I mentioned in the subsection on the birth of values, Dennett suggests that the moral dimension of our dispositions and practices comes from our capacity to critically examine the reasons of our actions and beliefs, and the consequent fact that the actions we perform, the judgments we hold and the attitudes we adopt are endorsed, and the standards to which we take all these to be answerable, are endorsed, if only indirectly. Holding
ourselves responsible makes this endorsement explicit, and implies that we acknowledge the moral dimension of our actions, judgments and attitudes.

Dennett suggests that most ordinary agents are not so much concerned with the metaphysical issue of determinism as with the more pragmatic issue of inevitability, and this is reflected in the ordinary use of the concept of “could have done otherwise”, of which he proposes an analysis which chimes with Strawson’s in (Strawson [1983]).

The concept has two orientations in time, past and future, both of which express its nature as a conceptual tool that we use, in association with the concept of “moral responsibility” itself, to install a practical attitude (or “policy”, or “strategy”) — not to refer to a metaphysical principle of “alternative possibilities”.

The past-orientation of the concept comes down to the acknowledgement that one’s action has fallen short of the ideal (independently of any judgment about the metaphysical status — “determined” or “undetermined” — of one’s past action). The future-orientation of the concept is a commitment to do things otherwise, i.e. better, in the future.

According to Dennett, the use of the concept of “could have done otherwise” is acquired in the first instance as one learns to hold oneself responsible for one’s own decisions and performances, and is then extended to the two other modes: acknowledging that others hold one responsible for one’s decisions and performances, and holding others responsible for their own. The following quote evokes how the ordinary use of “could have done otherwise” is
linked to self-criticism. The first paragraph is a short self-critical soliloquy that elucidates the use of the concept. The second paragraph evokes the complications that arise from the temptation to look for exculpations (a topic to which we will return in connection with the topic of moral formation):

This time I made a fool of myself; if the situation had been quite different, I certainly would have done otherwise; if the situation had been virtually the same, I might have done otherwise and I might not. The main thing is to see to it that I will jolly well do otherwise in similar situations in the future.

That, certainly, is the healthy attitude to take towards the regrettable parts of one’s recent past. […] Of course if I would rather find excuses than improve myself, I may dwell on the fact that I don’t have to “take” responsibility for my action, since I can always imagine a more fine-grained standpoint from which my predicament looms larger than I do. (If you make yourself really small, you can externalize virtually everything.) (1984, p.143)

4.2.1.4. Conditions of responsibility and of moral personhood

In Chapter One, I pointed out that Strawson’s analysis of the commitment-responsive standard (to which our moral sentiments and the judgments that accompany them are answerable) and of the responsibility-attributive condition (presupposed by the commitment-responsive standard) implied the following two-level concept of personhood. In a basic, “metaphysical” sense, personhood is universal, unconditional, and entitling: all humans are persons in this minimal
sense, and as such are entitled to regard. However, full personhood requires that one be held responsible for one’s decisions and performances — hence the importance of correct attributions of responsibility. That is the conditional, and committing, level of personhood. It will be remembered that Strawson argued that responsibility, and thus full personhood, do not depend upon metaphysical considerations, and consequently could not be affected by the putative truth of determinism.

Dennett explicitly defends such a conception of personhood, and proposes an extensive analysis of the mental capacities required to be held responsible, and therefore to enjoy full personhood. But first, let us see precisely how Dennett’s conception of personhood is similar to Strawson’s.

Dennett remarks that the ordinary notion of personhood seems to intertwine two distinct notions: the *metaphysical* notion of a person, and the *moral* notion of a person. However, it is somewhat confusing to treat them as two distinct notions, as he eventually makes the following claim: «The moral notion of a person and the metaphysical notion of a person are not separate and distinct concepts, but just two different and unstable resting points on the same continuum.» (1978 [1976], p.285)

Dennett, then, treats these two notions as two components of the notion personhood, one unconditional and entitling, the other conditional and committing.
Strawson described that to which all persons are entitled, and most persons committed, as “a minimum of regard” for human beings. Dennett describes it as *person-solicitude*. He remarks (1978 [1973], p.241) that a gardener may feel solicitude for his flowers, but points out that such “gardener-solicitude” is obviously different from person-solicitude. Only the latter is a *moral commitment*.

Parenthetically, Dennett formulates a critical interrogation regarding utilitarianism: «is the central fault in utilitarianism a confusion of gardener-solicitude with person-solicitude?» (1978 [1973], p.241). This chimes with Wiggins’ criticism that indicts utilitarianism on the ground that «the *locus* or origin of all value has been firmly confined within the familiar area of psychological states conceived in independence of what they are directed to.» (Wiggins, 2002 [1976], p.97)

Dennett acknowledges that some persons are not held responsible, but are still considered as persons in the basic sense. Evoking the case of a man declared “insane”, Dennett asserts: «In one sense of “person”, it seems, we continue to treat and view him as a person.» (1978 [1976], p.269)

---

79 For one thing, gardener-solicitude, not being directed at an intentional system, is not even a candidate for the sort of commitment that person-solicitude is. Second, person-solicitude is directed at a very special kind of intentional system, namely human beings.
Now, on the understanding that «metaphysical personhood is a necessary condition of moral personhood.» (1978 [1976], p.269), we may ask what the conditions are to be attributed “moral personhood” — in other words what the conditions are for being considered morally responsible. In “Conditions of Personhood” ([1976]), Dennett presents seven necessary but not sufficient conditions for the ascription of personhood, six of which are mental capacities. Dennett claims that if there are «clearly formulatable necessary and sufficient conditions for ascription» of the concept of personhood, «we have not yet discovered them.» (1978 [1976], p.267) We will return to the reasons why these seven necessary conditions are not sufficient for the ascription of personhood.

The seven conditions for attribution of responsibility, and so for moral personhood, are implied by our previous discussion (some of them have in fact been explicitly tackled, as we will see shortly). Dennett presents them in the order of their dependence. More exactly, he takes the first three to be mutually interdependent, following which the fourth depends upon these three, the fifth depends upon the fourth, etc. It appears that conditions one, two and four correspond to mental capacities that many animal species possess. Conditions five, six and seven correspond to cognitive capacities that only humans possess. Condition three states that moral personhood is a conditionally attributable deontic status, not a metaphysical property. (It is not clear why Dennett characterizes condition three as being “mutually interdependent” with one and two, and as being presupposed, together with one and two, by condition four, but
nothing much hinges on this unclarity). Let us see how Dennett formulates the seven conditions.

The first condition is rationality: «persons are rational beings» (1978 [1976], p.269).

The second condition is, one might say, mentality: «persons are beings to which states of consciousness are attributed, or to which psychological or mental or intentional predicates, are ascribed.» (ibid). It should be clear that Dennett’s reference to “states of consciousness” must not be read as an invitation to postulate a Conscious Active Self enjoying qualia.

The third condition states that moral personhood is a conditionally attributable deontic status. Again, being a human being is a necessary condition of full personhood. But the attribution of the conditional component partakes in a certain practical attitude, a stance — namely the personal stance. It does not depend upon some metaphysical property:

The third theme is that whether something counts as a person depends in some way on an attitude taken toward it, a stance adopted with respect to it. This theme suggests that it is not the case that once we have established the objective fact that something is a person, we treat him or her or it in a certain way, but that our treating him or her or it in this certain way is somehow and to some extent constitutive of its being a person. (1978 [1976], p.270)

The fourth condition is the capacity to entertain other-regarding higher-order propositional attitudes, i.e. beliefs and desires about the beliefs and desires of other agents. As we have seen in section 3.4., many animal species are
endowed with such a capacity, with the obvious proviso that in the case of animals, “beliefs” and “desires” must be construed as non-propositional intentional states (“perceptual states” and “volitional states”).

The fifth condition is the capacity to engage in verbal communication.

The sixth condition is the capacity to engage in the practice of giving and asking for reasons. In the following quote, Dennett describes how the advent of language and of the practice of giving and asking for reasons introduced the moral dimension of our relationship to the world (and to others):

*If I am to be held responsible for an action* (a bit of behavior of mine under a particular description), I must have been *aware* of that action under that description. Why? Because only if I was aware of the action can I say what I was about, and participate from a privileged position in the question-and-answer game of giving reasons for my actions. (If I am not in a privileged position to answer questions about the reasons for my actions, there is no special reason to ask *me.*) And what is so important about being able to participate in this game is that only those capable of participating in reason-giving can be argued into, or argued out of, courses of action or attitudes, and if one is incapable of “listening to reason” in some matter, one cannot be held responsible for it. The capacities for verbal communication and for awareness of one’s actions are thus essential in one who is going to be amenable to argument or persuasion, and such persuasion, such reciprocal adjustment of interests achieved by mutual exploitation of rationality, is a feature of the optimal mode of personal interaction. (1978 [1976], p.283)
The seventh condition of responsibility and moral personhood is the capacity to entertain second-order desires, which we have already discussed.

Dennett claims that these seven necessary conditions are not sufficient for the ascription of personhood, for the reason that “qualifying as a person” is too context-sensitive for there to be necessary and sufficient conditions for being a person, or for “being a good enough approximation” to an “ideal” person:

Now finally, why are we not in a position to claim that these necessary conditions of moral personhood are also sufficient? Simply because the concept of a person is [...] inescapably normative. Human beings or other entities can only aspire to being approximations of the ideal, and there can be no way to set a “passing grade” that is not arbitrary. (1978 [1976], p.285)

In other words, personhood is attributable on a “good reasons” basis, not on the basis of firm criteria. The conditions may not be sufficient, but on the other hand we obviously would have very good reasons to regard an agent manifestly endowed with those capacities as responsible for his or her performances.

Dennett acknowledges that his conception of personhood is close to Strawson’s, although he faults the latter’s conception (as well as that of many other thinkers) on one count: it does not clearly establish that rationality, while it is required for the attribution of the deontological status of personhood, remains logically distinct from it. Dennett claims that Strawson runs together the rationality that characterizes the behavior of all kinds of intentional agents, and the moral
accountability that is attributable to humans only. As Dennett makes clear, their logical distinctiveness is precisely the reason why an irreducible moral level of discourse is to be established over and above the intentional: «We might, then, distinguish a fourth stance, above the intentional stance, called the personal stance. The personal stance presupposes the intentional stance (note that the intentional stance presupposes neither lower stance)». (1978 [1973], p.240)

Dennett’s point is that the personal stance presupposes that the agents that we regard as persons are intentional systems, to which minimal rationality is attributed (it is inconceivable that entities that are not intentional systems could be candidates for personhood). But some intentional systems, namely animals and some machines, manifest a certain practical rationality without being persons. As we have seen, the unwitting manifestation of practical rationality in animal behavior is accounted for by evolutionary theory, at the functional level. Then, at the intentional level, it is explained how we human observers attribute rationality to all intentional systems, and attribute propositional attitudes to them, so as to interpret and anticipate their behavior. Then, at the moral level, it is explained how we attribute the unconditional element of personhood to all members of the human species, and only to them (which presupposes that we are intentional systems, and that we are therefore endowed with rationality); and how we attribute responsibility, and therefore full personhood (entitlement to person-solicitude and commitment to its manifestation), to those agents who fulfill the seven conditions stated by Dennett.
4.2.1.5. Vindication, revision and moral education

Dennett’s account of our ordinary sense of free, responsible agency is at once revisionary and vindicatory.

It is revisionary in the sense that it eliminates certain skyhooks. That is, it bars what tendencies we might have to think of free will, responsibility and personhood as involving a Conscious Active Self (which, according to Dennett, we tend to postulate spontaneously, as we have seen in 3.5.), as well as sui generis, metaphysically special properties. Dennett acknowledges that we do have a natural tendency to see things in ways that induce skyhooks. Thus, in the exercise of reflection, we may come to think that in order for us to be responsible, there has to be a primary metaphysical fact about a moral property of responsibility:

Why do we hold ourselves and others morally responsible? At first it may appear that this inquiry is parasitic on a metaphysical question about the conditions under which someone truly is responsible. […] Surely, it seems, we can make a distinction between the question of why we hold people responsible, or take responsibility ourselves for various things, and the question of why or whether we actually are responsible. (1984, p.163)

On the other hand, Dennett’s account is vindicatory to the extent that certain ordinary features of ordinary moral life are shown to be well-grounded.
Arguably, we naturally take ourselves as the authors of our own decisions and actions, and as the owners of our desires and beliefs, and in this we are right — as long as we do not appeal to skyhooks. Dennett’s account shows explicitly (in a rather detailed way) something that Strawson’s ethical view implied, viz. that we have a two-level concept of personhood which does not require any skyhook, and that is well-grounded. The conditional component of personhood corresponds to the attribution of responsibility, and on that count, our skyhook-inducing tendencies intermingle with a very pragmatic attitude:

Instead of investigating, endlessly, in an attempt to discover whether or not a particular trait is of someone’s making — instead of trying to assay exactly to what degree a particular self is self-made — we simply hold people responsible for their conduct (within limits we take care not to examine too closely). And we are rewarded for adopting this strategy by the higher proportion of “responsible” behavior we thereby inculcate. (1984, p.164)80

If Dennett is right, most ordinary agents are unconcerned by determinism, and much concerned by issues of evitability, and such basic concerns of moral life are well-grounded, as is the ordinary use of the idea of “could have done otherwise”.

80 It should be clear that the last sentence in the preceding quote does not express a consequentialist construal of responsibility. It simply reminds us that the beneficial consequences of holding persons responsible are still there for us to reap.
In other words, many of our ordinary attitudes are right, and the ordinary standard of correctness by reference to which we assess attributions of moral responsibility is well-grounded — not in extra-human principles, but in features of our dispositional constitution that we are ready to endorse in the practice of giving and asking for reasons. Here, I contend that Dennett would agree with Wiggins, about the relation between responsibility and the sphere of the unforsakeable: the genuinely moral character of responsibility and free will is that human beings have come to value them, not for their beneficial consequences, but because they are indissociably linked to unforsakeable concerns, motives, etc. But Dennett presents a more elaborate argumentation to the effect that we are responsible and that we have free will.

It must be clear that Dennett’s theory rejects, as well as skyhooks, the idea that the moral dimension of human existence follows from consequentialist rationales. In his view, such rationales account for the emergence of proto-moral dispositions, not for the features of genuinely moral life.

Now, holding persons responsible in the right way involves complications and temptations. As suggested by the last sentence in the quote in 4.2.1.3. about the use of “could have done otherwise” («if you make yourself really small, you can externalize virtually everything»), one may hesitate to hold oneself (or other persons) responsible, and deterministic ideas come in handy as a way of unjustifiably extending excusing conditions to cases in which one looks for exculpations. Thus, one may illegitimately invoke excusing conditions, or dispute
the authorship of actions, or dispute their consequences, etc. As natural as it is to hold persons responsible, doing it correctly requires psychological maturity and a certain moral formation.

Moral responsibility, then, is attributed gradually to humans, as children grow up into adults, and as they learn to take responsibility for their decisions and performances, to acknowledge that everybody is to take responsibility in the same way, and to hold others responsible for their own decisions and performances.

As one learns to take responsibility for one’s actions, one acquires a morally responsible self. The generalization of this practice implies, from one’s individual point of view, that others may take one as responsible for one’s actions, and that one may take others as responsible for their actions. As the self grows and matures, it becomes an endorser and attributer of moral responsibility.

Moreover, philosophers can make suggestions as to how we can sharpen our concepts and ideas, our policies and the standards of correctness to which we hold them answerable. Wiggins makes one such suggestion when he proposes to take in a non-transcendental way Kant’s Formula of Humanity as an End in Itself, i.e. to take it as an idealized principle of action guidance. Dennett

81 The narrative self and moral responsibility exemplify two different sorts of gradual attribution. The biographical self is attributed gradually, as events concerning the agent to whom it is attributed accumulate. By contrast, moral responsibility is a moral capacity gradually attributed to an individual according to her degree of maturity.
makes a related proposal, i.e. that the standard of correctness by reference to which we assess attributions of moral responsibility should express an ideal of autonomy, which we can express as a “Default Responsibility Principle”:

If no one else is responsible for your being in state A, you are. This nicely cuts off the infinite regress feared by Kane; it permits us to pass the buck to brainwashers (if such there be in your past) but not to “society” in general or to the agentless environment. (2003, p.281)\(^82\)

This point connects with Dennett’s claim (which we have encountered in 4.2.1.2.) that one is responsible for one’s body of beliefs, for its coherence and its cognitive maintenance: if one is responsible for “being in state A”, then one is responsible for the coherence of one’s body of beliefs, notably those upon which one acts.

4.2.1.6. The critique of Wiggins, of Sartre, and of elementary nihilism

Here, I want to present a brief survey of what appears like a slight change of attitude on Dennett’s part toward libertarian views, which apparently occurred between the publication of “On Giving What Libertarians Say They Want” (1978) and that of Elbow Room (1984). As we will see shortly, in (1978), Dennett does

---

\(^82\) In this quote, Dennett mistakenly attributes to Alfred Mele a principle of which he is the author. Mele himself has corrected this mistake in his comment of Dennett’s book (this was pointed out to me by Daniel Dennett in a personal communication).
not reject as a *skyhook* Wiggins’ idea that moral agency is metaphysically located somewhere between causal determination and randomness. Moreover, he recognizes that, although determinism is (on his view) fully compatible with responsibility and free will, it remains a prospect that may be distressing, and that may induce nihilism. But in (1984), although he does not go back over Wiggins’ position, he bluntly rejects libertarianism à la Sartre, approvingly quotes Strawson’s characterization of libertarian metaphysics as “obscure and panicky”, and dismisses nihilism as a “negligible position”.

This change of emphasis probably manifests the fact that by 1978, Dennett had not fully connected his philosophical framework to evolutionary theory, and it was not clear yet that the critical side of his account of responsibility and free will, as well as of his framework as a whole, implied the rejection of the cosmological view he will come to call, in (Dennett 1995), the Cosmic Pyramid. By 1984, Dennett seems to have come to the view that, even though the idea of a middle ground between causal determination and randomness might not be, as such, a skyhook, it readily lends itself to the faulty cosmology of the Cosmic Pyramid. Moreover, Dennett forcibly asserts that since genuine responsibility and free will do not depend upon metaphysical facts, the vindication of the former should simply not appeal to the latter. It is worth examining how Dennett expresses this, since it is the only juncture at which he explicitly discusses a position of Wiggins’.
“On Giving What Libertarians Say They Want” (1978) is one of Dennett’s earliest expressions of his model of decision-making (it emphasizes the account of personal authorship for moral decision), and a direct answer to (Wiggins [1973]).

Dennett agrees with Wiggins’ point that “non-causally determined” does not imply “random”. Consequently, processes of decision-making and agency within which non-causally determined events would occur need not be random: they might manifest rational patterns, making them intelligible, and to some extent predictable. Dennett’s point is that what matters for ordinary moral agents is that first, we may claim authorship for our decisions and actions; and second, these performances are intelligible, and to some extent predictable, in virtue of the rational patterns they manifest. The possibility that non-causally determined events might occur in the decision-making process (or in the performance of action) is irrelevant. Wiggins’ idea of human agency as being metaphysically located on a middle ground between causal determination and randomness, then, is dismissed as irrelevant (but not rejected as a skyhook).

What about this other point made by Wiggins: the fact that the language of action and responsibility is not something that we need be introduced to does not mean that nothing could count as a reason to opt out of it? I take Dennett’s answer to go as follows: if the compatibilist account of moral responsibility and free will is right, then we have moral responsibility and free will anyway. In which case, why would the truth of determinism count as a reason to opt out of the personal stance (the language of action and responsibility)?
Nevertheless, Dennett concludes this text with something like a psychological, or moral, concession to Wiggins’ libertarian concerns. Even though we have responsibility and free will, the prospect of determinism can be extremely disturbing, and may lead to nihilism (of what I have called the elementary form):

Wiggins speaks of the cosmic unfairness of determinism, and I do not think the considerations raised here do much to allay our worries about that. Even if one embraces the sort of view I have outlined, the deterministic view of the unbranching and inexorable history of the universe can inspire terror or despair, and perhaps the libertarian is right that there is no way to allay these feelings short of a brute denial of determinism. Perhaps such a denial, and only such a denial, would permit us to make sense of the notion that our actual lives are created by us over time out of possibilities that exist in virtue of our earlier decisions; that we trace a path through a branching maze that both defines who we are, and why, to some extent, (if we are fortunate enough to maintain against all vicissitudes the integrity of our deliberational machinery) we are responsible for being who we are. That prospect deserves an investigation of its own. (1978, p.299)

In Elbow Room (1984), and thereafter, Dennett seems to reconsider this moral/psychological concession to libertarian concerns. He acknowledges that his compatibilist account of responsibility and free will may appear disappointingly down-to-earth, but he insists that it is adequate, and emancipating, in a way that “brute denials of determinism” — that of Jean-Paul Sartre, for example — are not:
This may sound trivial, but it flies in the face of a very influential — if ultimately incoherent — vision of self-choosing: Sartre’s idea that one chooses oneself completely by a “radical choice” that brings none of yesterday’s baggage along. In Sartre’s view, it is as if one could create oneself ex nihilo; as if, to borrow another familiar bit of philosopher’s Latin, the self at birth were a tabula rasa or blank slate. We know better.

Sartre knew better too, of course, but unfortunately chose this extravagant metaphor to express his view. It is not a helpful metaphor. It postulates a magical power none of us could have, and encourages people to believe that if science shows that they cannot have it, they cannot be free. [...] It is not hard to see, however, why a philosopher (such as Sartre) would want to hold out for the absolute, for there is an argument that often bubbles up to the surface to the effect that unless one were absolutely responsible for oneself, one could not be responsible at all. (1984, p.83)

Throughout Elbow Room, Dennett asserts the evolutionary character of his framework, and what he will eventually call the “Cosmic Pyramid” is quite clearly (if not explicitly) rejected. This implies the rejection of the factitious alternative between dubiously metaphysical philosophical theories on the one hand and elementary nihilism on the other. Yet, Dennett directs a supplementary rebuttal against elementary nihilism. The reason for this “extra” is that, according to

83 Still, one might want to argue that a young man who asks a sage whether he should join the Résistance or stay home to care for his aging, sick mother makes a better story than a baseball player who lets the ball hit him in the face so that he can get to first base.
Dennett, the fear of nihilism probably provides some psychological support to skyhook-inducing theories, so an explicit dismissal of nihilism, however brief, might be useful.

Dennett does not deny that nihilism is intelligible, nor does he deny that the sort of compatibilism that he advocates might entail it (although he does not think it does). But he dismisses nihilism as a “negligible position”, which we may pragmatically assume is false.

This dismissal expresses Dennett’s pragmatic assumption that in common moral life, moral questions present themselves to ordinary moral agents as, first and foremost, practical problems about what ought to be done “here and now” by the finite beings that we are, i.e. beings with cognitive limitations, as well as time limitations. These problems call primarily for practical, “ought” judgments (rather than evaluative, “is” judgments), and time-pressured decision-making:

Nihilism might, I suppose, be true; it might even follow from determinism, as some suspect (though I see no reason to believe this). Nevertheless, we may assume that nihilism is false. How can I make that complacent (cavalier) claim? Shouldn’t we even stop to consider carefully the prospect that it might be true? Well, if it were true that we ought to take the possibility of nihilism seriously, then nihilism would be false, for if we ought to do anything nihilism is false. But then we may as well assume that it is false, since either it is false or, if it isn’t, nothing matters and we may do whatever we want. Nihilism is, quite literally, a negligible position.

Since there could be nothing better to do, we might just as well affirm our confidence in our starting point. We may assume that something matters, and that we are rational enough so that there can be a point
to our attempts to understand the world. I do not see how those assumptions could be coherently criticized — which is not to say that they must be true. (Dennett, p.156)

On the one hand, Dennett does not dispute the intelligibility of nihilism. More precisely, he does not dispute the intelligibility of the claim that even though we are “avoiders of fatality”, it remains that if determinism is true, then what one believes and desires is ultimately determined, and that this does entail nihilism.

On the other hand, he suggests that nihilism is self-contradictory if we take it to be expressed by the proposition “we ought to consider that we ought not do anything”. On this count, however, I would remark that nihilism may be expressed by propositions that are not self-contradictory. For example, the proposition “it is not the case that anything ought to be done, here and now” is not self-contradictory.

The gist of Dennett’s dismissal of nihilism as a negligible position resides in his pragmatic construal of it, following his pragmatic thesis that moral questions are first and foremost practical problems about what ought to be done by the finite beings that we are. Thus, nihilism is construed as a position (or a proposition) bearing on “what ought to be done”, here and now, by human beings, with finite cognition and finite lifespan. This invites the pragmatic, and commonsensical answer: since we are finite beings and time is lapsing, not deciding anything comes down to a decision by default, and “not doing anything” is a negligible position indeed.
4.2.2. Underdetermination, conversation-stoppers and metaethics

Dennett suggests that our deliberations are put to an end, at some point, by maxims which he calls *conversation-stoppers*. I contend that these maxims can be taken as propositional formulations of unforsakeable concerns, motives, etc., and that therefore they constitute basic standards of correctness. Not only do we refer to those standards to morally assess our judgments, emotions and actions, but these standards also — Dennett adds — put an end to deliberation, and motivate decision and action. I contend that for Dennett as for Wiggins, we are endowed with unforsakeable concerns, motives etc, which underdetermine the propositional content of conversation-stopping maxims, as well as the decisions we make and the actions we take.

Dennett relevantly insists (as Wiggins does not) on the importance of certain features of our "given cum acquired nature", notably the fact that our numerous commitments continually call for decisions and action, in circumstances where we enjoy a limited, but genuine and expandable autonomy; where our cognitive resources are limited, and our lifespan is finite, so that decision-making is always performed under a certain time pressure. Thus,

---

84 Dennett would approve, as much as Wiggins, of this remark of Ross: «[My neighbours] do stand in this relation [of possible beneficiaries of my action] to me […]. But they may stand in the relation of promisee to promiser, of creditor to debtor, of wife to husband, of child to parent, of
common moral life implies a multiplicity of practical problems with a moral
dimension, about what ought to be done here and now by finite beings like us.
We may not have the sort of free will that libertarians dream about, but we have
all the autonomy necessary to benefit from sound principles of decision-making,
and to suffer from inadequate such principles. Hence the importance of a general
heuristic of decision — of moral decision in particular.

The practical necessity of putting a stop to deliberation corresponds to a
commonsensical “metaprinciple” of decision (or, more modestly, a basic
guideline) according to which we must avoid, on the one hand, putting a
premature stop to consideration generation and examination; and on the other
hand, spending too much precious time over decreasingly useful deliberation. In
other words, we must “satisfice”, strike a tradeoff (in a time-pressured context)
between ideal optimality on the one hand, and real-world constraints on the other.
Hence the importance of letting “conversation-stoppers” force the selection of one
consideration upon which to act.

As an illustration, Dennett conjures up a particular instance of decision-
making in a departmental meeting. In the first paragraph of the following quote,
Dennett evokes the fact that sub-optimal, “second-rate conversation stoppers”
are often what we have to settle for, lest we spend all our time in endless

fellow countryman to fellow countryman, and the like […]» (Ross 1930, p.19; quoted in Wiggins
2006, p.218-219)
scrutiny. The second paragraph advocates the commonsensical guideline. The third paragraph suggests that although decisions are usually context-sensitive, and although we often have to settle for second-rate conversation-stoppers, there are certain concerns at the basis of common moral life, which are expressible by sentences that act as universal, or quasi-universal conversation-stoppers, commendable to everyone in practically all situations.

It will not do for all these people to be endlessly philosophizing, endlessly calling us back to first principles and demanding a justification for these apparently (and actually) quite arbitrary principles. What could possibly protect an arbitrary and somewhat second-rate conversation-stopper from such relentless scrutiny? A meta-policy that forbids discussion and reconsideration of the conversation-stoppers? But, our colleagues would want to ask, is that a wise policy? Can it be justified? It will not always yield the best results, surely, and… and so forth.

This is a matter of delicate balance, with pitfalls on both sides. On one side, we must avoid the error of thinking that the solution is more rationality, more rules, more justifications, for there is no end to that demand. Any policy may be questioned, so, unless we provide for some brute and a-rational termination of the issue, we will design a decision process that spirals fruitlessly to infinity. On the other side, no mere brute fact about the way we are built is — or should be — entirely beyond the reach of being undone by further reflection.

We cannot expect there to be a single stable solution to such a design problem, but, rather, a variety of uncertain and temporary equilibria, with the conversation-stoppers tending to accrete pearly layers of supporting dogma which themselves cannot withstand
extended scrutiny but do actually serve on occasion, blessedly, to deflect and terminate consideration. Here are some promising examples:

“But that would do more harm than good.”

“But that would be murder.”

“But that would be to break a promise.”

“But that would be to use someone solely as a means.”

“But that would be to violate a person’s right.” (1995, p.506-507)

Dennett’s formulation in the passage just quoted might give the impression that those maxims are more or less arbitrary, but I contend that we can take them as a Dennettian formulation of the unforsakeable. Most of these conversation-stoppers can be formulated generically, as evaluative or as practical moral judgments. In fact, only the first one is too closely bound to context to be given a generic formulation (other that the empty utilitarian principle to “maximize good”). But if we formulate as moral judgments the other conversation-stoppers in the list above, we could obtain the following results.

The fourth is obviously drawn from one of the formulations of Kant’s Moral Law, viz. the formula of humanity as end in itself: «So act that you use humanity, whether in your own person or that of another, always at the same time as an end, never merely as a means.» (Groundwork of the metaphysics of morals 429) I take it that Dennett construes it in a way similar to Wiggins’: not as a formulation of a transcendentally deduced “moral law”, but as a formulation of an idealized
principle of action guidance, underdetermined by the sphere of the unforsakeable — one maxim among others of comparable importance.

The fifth is drawn from the “human rights” tradition of thought, of which we could give a summary formulation with the judgment that “persons have rights, and these rights are inalienable”.

The third could be formulated as the judgment that “promise-keeping is a virtue, promise-breaking is a vice”, and linked to Hume’s discussion of promise-keeping as an artificial virtue (Treatise, 3.2.5.)

The second could be formulated as “murder is unforsakeably wrong” — a formulation that brings out the kinship between Dennett’s general “conversation-stoppers” and the number (1) category of Wiggins’ unforsakeable concerns, that of “prohibitive aversions”.

I propose that we could add to the list above the standard of correctness to which attributions of responsibility are ordinarily taken to be answerable: “Persons are responsible for their decisions, attitudes and actions, unless specifiable excusing conditions (namely those specified by Strawson) apply”. This proposition seems to be just as robust a conversation-stopper as those above.

I contend that Dennett would agree that those maxims are neither arbitrary, nor transcendentally (or otherwise) deduced, nor determined by the dispositional constitution of the human species, but that they are rather underdetermined (conditioned and constrained, but not determined) by our human predicament, which includes those facts: we have limited time and
cognitive resources, we live in communities that demand that we commit
ourselves, we must take decisions and take action.

Dennett, then, is not just describing concerns and formulating maxims that
put a stop to deliberation and motivate decision and action. He is endorsing them,
and vindicating them theoretically, on the basis of the underdetermination relation
just sketched, in combination with the anthropocentricity thesis, which I take to be
implied by his epistemology.85

There is little textual evidence to bring in support of my contention that
Dennett’s notion of conversation-stoppers is a variant of Wiggins’ “unforsakeable
concerns and motives” (a variant which insists on its role in relation to
deliberation, in the context of our condition of temporally and cognitively limited,
but nevertheless committed agents), but it might bring indirect support to consider
that Dennett opposes his view to consequentialist and to deontological theories in
a way that chimes with Wiggins’ critique of the same theories. Here again,
Dennett insists on moral problem-solving, of which consequentialism and of
Kantian deontological theories propose inadequate treatments.

85 The role that Dennett imparts to those sentences is reminiscent of that which Wittgenstein
imparted to those “hinge propositions which he claimed lie at the basis of certain language
games. Whether these hinge propositions are taken to be “true ” or simply “justified”, they are
bound to elicit much agreement, and to serve as standards for evaluating other, less fundamental
judgments.
Consequentialism demands that courses of events be assigned calculable values, so that idealized agents may engage in a global cost-benefit analysis (featuring those calculable values). Dennett attacks both the idea of assigning calculable values to courses of events, and the idealization of human agents that consequentialism supposes.

First, assigning values to events is a fantasy, Dennett tells us. There is no Archimedean point from which we may assign values to events and their consequences, if only a broad characterization in terms of what is “positive” or “negative”. As an example, Dennett points out that there is no principled way of assigning a “positive” or “negative” value to the accident at the nuclear plant of Three Mile Island in 1980. On some counts, it was a “bad thing”, and on others it was a “good thing” (1995, p. 498-499).

Second, granted that ethical theories must have recourse to certain idealizations, consequentialism involves an inadequate idealization of agents and of the decision-making process:

Ethics, like any theoretical enterprise in science, has always been conducted with the aid of idealizations. Reality, in all its messy particularity, is too complicated to theorize about, taken straight. A favorite idealization in ethics has been the useful myth of the moral

---

86 To those two critical points, we may add the point, mentioned earlier, that utilitarianism mistakenly construes psychological states in isolation from what they are directed to; precisely, the idea that utilitarianism confuses the solicitude that a gardener may feel towards his flowers with the solicitude that we are committed to in relation to persons.
agent with unlimited knowledge and time for ethical decision-making. For instance, consequentialist theories, such as the various brands of utilitarianism, declare that what ought to be done is always whatever course of action will have the best expected consequences \textit{all things considered}. […]

The brute fact that we are all finite and forgetful and have to rush to judgment is standardly regarded, not implausibly, as a real but irrelevant bit of friction in the machinery whose blueprint we are describing. (1998 [1986], p.380-381)

Dennett's point is that setting aside our cognitive and temporal limitations is a legitimate idealizing move, and not a mistake as such. However, as it turns out, it distorts the real situation too much to be a methodologically adequate move. Thus, an adequate model of human decision-making must take into account the fact that humans have finite cognitive capacities and finite lifespan.

Moreover, consequentialism seems to presuppose that agents are endowed with the sort of rationality that Wiggins debunked as “cosmic rationality”:

The mistake that is sometimes made is to suppose \textit{that there is or must be a single (best or highest) perspective from which to assess ideal rationality}. Does the ideally rational agent have the all-too-human problem of not being able to remember certain crucial considerations when they would be most telling, most effective in resolving a quandary? If we stipulate, as a theoretical simplification, that our imagined ideal agent is immune to such disorders, then we don’t get to ask the question of what the ideal way might be to cope with them. (1995, p.505; emphasis added)
Dennett’s point is that this sort of idealization has got it all wrong. Fundamental constraints are involved in decision-making, and our rationality is anthropocentric (I take it that Dennett would agree with the present use of this Wigginsian term) and context-sensitive. Those constraints, not fictitious “utilities”, are precisely what we have to include in models of decision-making, both in their descriptive and prescriptive dimensions.

Kantian deontological theories, according to Dennett, involve a sort of idealization that is different from the consequentialist sort, but it is just as inadequate. The principal problem with Kantian theories is that they leave us clueless as to how the “test” of the Categorical Imperative should be carried on candidate maxims:

Kantian decision-making typically reveals rather different idealizations — departures from reality in other directions — doing all the work. For instance, unless some *deus ex machina* is standing by, a handy master of ceremonies to whisper suggestions in your ear, *it is far from clear just how you are supposed to figure out how to limit the scope of the “maxims” of your contemplated actions before putting them to the litmus test of the Categorical Imperative*. There seems to be an inexhaustible supply of candidate maxims. (1995, p.500; emphasis added)

In the end, Dennett’s position echoes Wiggins’ rejection of traditional moral theories, but on the specific count that they are not computationally tractable: «no
remotely compelling system of ethics has ever been made *computationally tractable*, even indirectly, for real-world moral problems.» (1995, p.500)

Dennett does not directly discuss the traditional semantical and epistemological issues of metaethics: what are the standards of correctness for evaluative judgments and for practical judgments? Are they capable of being true or false? How should we interpret, ontologically, the terms that occur in moral judgments?

We have seen in section 3.1. that Dennett favors a “good reasons” approach to the assessment of judgments, where well-groundedness (warranted assertability) has much greater importance than truth, and where predictive success is the most significant property of well-groundedness for most classes of judgments. Dennett also favors a “good reasons” approach of the ontological interpretation of the terms that occur in sentences. What does that tell us regarding moral judgments?

Although evaluative moral judgments are part of ordinary psychology, and may be linked to the attribution of propositional attitudes, it is difficult to hold that moral judgments, as such, directly entail certain predictions, so that their cognitive status can be assessed on the basis of their “predictive success”. It is difficult to see how, in the practice of giving and asking for reasons, predictive success could constitute a reason in support of moral judgments, either of the evaluative or of the practical sort.
Dennett suggests that propositions expressed by conversation-stoppers correspond to basic ordinary standards of morality. They are not grounded in extra-human principles, and so are not “foundational” in that sense. But, following the anthropocentric thesis to which Dennett adheres, that does not bar us from considering them as adequate. Again, the epistemological role of such propositions seems similar to that of hinge propositions in contextualist theories of knowledge.

Dennett seems to entertain a contextualist view, then, but only in relation to the personal stance. That is because judgments which pertain to the other levels of discourse do entail predictions, and in their case, predictive success is the most significant property of well-groundedness, in virtue of which we may considered them as epistemically justified.

If my interpretation of Dennett’s framework is adequate, contextualism is not accepted as an all-inclusive theory of knowledge. But it is the only theory of knowledge for the personal stance.

4.2.3. Radical re-evaluations and attributions of moral meaningfulness

Dennett does not directly tackle the issue of moral meaningfulness, but he does make a few remarks on a passage from a text by Charles Taylor, which
bears on what Taylor calls “radical re-evaluations”. Since Taylor’s reflection is relevant to the issue of moral meaningfulness, Dennett’s comments bear indirectly on that topic.

Although we are routinely caught in the ordinary pursuits of common life, the practice of giving and asking for reasons may bring us to doubt that anything might have intrinsic worth, and might be apt to constitute a non-derived finality to human action and existence. I have pointed out in the Introduction that in the absence of a positive answer to the challenge of moral meaningfulness, we face the “higher” form of nihilism, that which Camus called “the absurd”.

Taylor, in (Taylor 1976), sketches the form that rational criticism takes when it calls into question, in one stride, the meaning (the correctness conditions) of all of the most fundamental terms of one’s evaluative vocabulary, so that no term remains which could constitute a basis on which the others could be defined. Taylor calls this form of rational criticism “radical re-evaluations”. Taylor suggests that it bars the formulation of moral judgments, and that moreover, it entails a breakdown of intersubjectivity. Such radical re-evaluations make it practically impossible to find anything intrinsically worthwhile, and so carry the threat of what I have characterized as the higher form of nihilism. Taylor evokes radical re-evaluations as follows:

Evaluation is such that there is always room for re-evaluation. But our evaluations are the more open to challenge precisely in virtue of the very character of depth which we see in the self. For it is precisely the deepest evaluations which are least clear, least articulated, most
easily subject to illusion and distortion. It is those which are closest to what I am as a subject, in the sense that shorn of them I would break down as a person, which are among the hardest for me to be clear about. The question can always be posed: ought I to re-evaluate my most basic evaluations? Have I really understood what is essential to my identity? Have I truly determined what I sense to be the highest mode of life? [...] A re-evaluation of this kind, once embarked on, is of a peculiar sort. It is unlike a less than radical evaluation which is carried on within the terms of some fundamental evaluation, when I ask myself whether it would be honest to take advantage of this income-tax loophole, or smuggle something through customs. These can be carried on in a language which is out of dispute. In answering the question just mentioned the term “honest” is taken as beyond challenge. But in radical re-evaluations, the most basic terms, those in which other evaluations are carried on, are precisely what is in question. [...] How then can such re-evaluation be carried on? There is certainly no metalanguage available in which I can assess rival self-interpretations. If there were, this would not be a radical re-evaluation. (Taylor 1976, p.296-297; quoted in Dennett 1984, p.88; emphasis added)

A radical re-evaluation, then, is a bout of rational criticism within which it appears impossible to fix any convincing standard of correctness for moral judgments. As a result, the meaning of our moral terms is suspended. The result is not a principled affirmation of nihilism, but a nihilistic result, as we find ourselves in the impossibility of attributing moral meaningfulness to anything. Clearly, Taylor’s radical re-evaluations bear the threat of the “higher” form of nihilism.
Dennett’s answer focuses on the fact that we human beings have finite lifespan, and face the practical obligation of making decisions, and of acting, under time pressure. I take it that the complete Dennettian rejoinder should stress the role of conversation-stoppers, and that the most radical re-evaluation is eventually halted by them\(^8\).

Consequently, Dennett makes a “decision-making” interpretation of radical re-evaluations, and neutralizes their corrosive potential by an emphasis on the constraints that bear on decision-making in general, in particular time pressure. Spending too much time on radical re-evaluations, and indefinitely deferring action, is so unreasonable that it indirectly sets limit on their radicalness:

[…] this process of reflective self-criticism and meta-level control reasoning must have its limits. If a complacent […] rejection of all opportunities for reflection would be ill-advised, it would clearly also be irrational to embark on a limitless round of self-evaluation. The unexamined life may not be worth living, but the overexamined life is nothing to write home about either. Moreover, there is almost certainly no “book” answer to the question of how much moderation is the right amount of moderation. (1984, p.87)

\(^8\) Extending Dennett’s conception of the problem with a Deweyan thought, we may suggest the hope that action itself might bring elements of answer to theoretical questions about moral meaningfulness.
Indirectly, Dennett asserts the role of “conversation-stoppers”, given our human finiteness. Anyone engaged in radical re-evaluations has a certain lifespan in front of her. In order to go on living, she has to undertake some commitments, and to take care of a number of things — and a number of persons. Her commitments, the care she has to take of persons and things, and the decision-making it implies, might bring their own contribution to the radical re-evaluation.

We have no better reason to opt out of the language of action and responsibility than to stick with it, and we have things and persons to care about. This will bring significant elements of answer to radical re-evaluations, without which an otherwise valuable exercise becomes an idle spiral of despair.

Dennett’s philosophical scrutiny, then, simply vindicates this ordinary, practical attitude, with the proviso that it excludes finding, somewhere within the human condition, metaphysically special properties.

Taylor does not seem to acknowledge those crucial factors. The passage above evokes a situation where certain evaluative judgments must be issued and validated before the practical business can get going — as if he assumed that one’s answer to practical questions about how one should act, what decisions one should take, presupposes that one has at least the beginning of an answer to fundamental evaluations concerning what has intrinsic worth, and so could constitute a non-derived finality.
By contrast, Dennett suggests that for committed beings with limited
cognitive capacities and limited time, who must therefore always make decisions
under time pressure, certain conversation-stoppers appear as unforsakeable.
From the moment we accept that there is nothing better to do than to go on living,
practical concerns force a pro tanto resolution of issues of moral meaningfulness,
which we might be able to clarify as we go on living.
CONCLUSION

I endorse Dennett’s overall doctrinal framework. Dennett is right to specify at the outset that, for purposes of “serious” descriptions and accounts, we should use a rationally acceptable language regimented into four levels of discourse, which excludes skyhooks and essentialist thinking, but admits of non-referential terms and vocabularies. It is a great quality of his framework that it provides, in virtue of a sufficient characterization of the relations between levels of discourse, credible and satisfactory answers to the placement question.

The ontological interpretation of terms is to be decided on a “good reasons” basis, but the reasons we shall provide in support of our verdicts are at least as significant of the verdicts themselves, which do not come down to a real/unreal dichotomy, but rather to a variety of possibilities.

Epistemologically, Dennett is right to accord more importance to warranted assertability than to truth itself, which we can only take “with a grain of salt”. Adequate standards need not be ultimately grounded in extra-human principles. They may very well be grounded in our dispositional constitution.

Within this framework, Dennett has built a solid, evolutionary theory of human nature, which indirectly shows how it underdetermines the standards by reference to which we morally assess conduct, sentiments, and judgments.
In rational criticism, humans deploy and develop their autonomy, this capacity they have to appropriate the reasons why they act as they do and hold the judgments they hold. We have no “magical power”, and we are not endowed with the Conscious Active Self that ordinary experience suggests to us, and that various philosophical doctrines have erroneously vindicated. But we have an extraordinary capacity for maximizing our autonomy, and this gives us free will, and makes of the space of reasons opened by rational criticism a space of freedom. The capacity for rational criticism marks the beginning of the “our” history — the history of language-using human beings — and it has developed historically into a variety of cultures. As I see it, Dennett is proposing a naturalist view of Hegel. Dennett insists on the two faces on underdetermination: all that conditions and constrains our condition, and all the expandable autonomy and self-control that saves us from determination.

Dennett’s metaphysics of experience was not discussed in this dissertation, but it is compatible with Wiggins’. It emphasizes the role of evolution as it makes the point that secondary qualities and affective properties on the one hand, and on the other hand the functional systems, dedicated to their detection, that humans are endowed with are «made for one another» (Dennett 1992, p.378), because they originate in the relation between «mutually adjusted systems» (Dennett 1992, p.376): on the one hand, our evolutionary ancestors who had interests in perceiving certain things; and on the other hand, animal and vegetal species that had interests in being perceived. This, then appears as the
evolutionary version of Wiggins’ “equal and reciprocal partners” view of the experience of non-primary properties.

Wiggins’ framework may deliver a similarly adequate view. But for that, it would have to be the case that we, interpreters of discourse and assessers of judgments, who avail ourselves to the Principle of Coherence and the Principle of Humanity to ascribe propositional attitudes to one another and propositional content to those same propositional attitudes, we who participate upon those conditions in the practice of giving and asking for reasons, we come, on a piecemeal basis, to ban those concepts which are used to play regress-stopping explanatory roles precisely at those junctures where our concerns and motives would be better served by finite regresses that peter out into insignificance. In other words, we would have to rediscover, on a piecemeal basis, in the context of “radical interpretation”, the qualities of Dennett’s rationally acceptable language, which we might as well adopt at the outset, with all its specifications, including one which will provide us “automatically” with credible answers to the placement question.

As it appears from the respective moral anthropologies of Wiggins and of Dennett, one of the strengths of this approach is that it combines phenomenological and genealogical investigations, thus showing how our moral lives are underdetermined by the dispositional constitution of our species.
Dennett’s anthropology of moral life focuses on our ordinary sense of morally responsible, free agency, and on moral decision-making.

His compatibilist analysis of our ordinary sense of free, responsible agency mostly vindicates it, as it shows that we are fatality-avoiders, endowed with an autonomy whose maximization has no clear limit. This capacity is part of our dispositional constitution, which generally conditions and constrains our action and cognition, but our capacity for autonomy implies that our paths are certainly not determined at birth. At the same time, Dennett’s analysis guards us against the illusion of the Conscious Active Self, and the postulation of «a magical power none of us could have» (1984, p.83).

With his analysis of moral decision-making, Dennett adds an interesting consideration to the underdetermination thesis: we have cognitive limits, time limits, and we cannot help but be committed. Given that we have free will, we continuously have to make decisions under time pressure. It is in this context that we must consider deliberation-stopping maxims. If I am right, those conversation-stoppers play in Dennett’s moral theory a role equivalent to that played by unforsakeable concerns and motives in Wiggins’.

Moreover, Dennett’s consideration of our practical predicament as limited beings throw some light on the issue, set up by Wiggins, of moral meaningfulness. Since we must continually make decisions under time pressure, we have no choice but to take certain things as what we might call “pro tanto ultimate ends”, and see what happens as we go, fully knowing that our sense of
what is intrinsically worthwhile, and of what confers an ultimate sense of point to
our experience, action and existence, is eminently revisable.

Wiggins’ Humean theory of human nature completes Dennett’s, adding to
it a deep layer of Humean scholarship and new insights. What I have called the
“Dennettian linguistic extension” of the evolutionary theory of human nature is the
locus of the most promising connections. Both philosophers validate Hume’s
account of natural virtues and artificial virtues. Dennett emphasizes the Humean
insight that rational criticism allowed humans to appropriate the reasons why they
act as they do and hold the judgments they hold, as part of our indefinitely
expandable capacity for autonomy. This is well completed by Wiggins’ emphasis
on the intersubjective strand of Hume’s thought, which suggests that our moral
standards are neither subjective, nor objective, but intersubjective.

Wiggins’ conception of the unforsakeable, with its fourfold classification of
concerns, motives, etc. is richer than Dennett’s short list of conversation-
stoppers, but it is well completed by Dennett’s emphasis on moral decision-
making, and his considerations about our practical predicament as limited beings.

As regards responsibility and free will, I side with Dennett. Wiggins’
considerations about there being an unclaimed ground between causal
determination and randomness are interesting, but not directly relevant to moral
life.
Wiggins’ “weak cognitivism” about moral judgments relies on his analysis of what he presents as our ordinary concept of truth, and on the principles of the Davidsonian theory of the interpretation and critical assessment of judgments. I find nothing to invalidate this theory outright, but it requires many unverifiable assumptions about the process of interpretation and critical assessment, and about our ordinary notion of truth. Comparatively, I find more attractive Dennett’s pragmatist epistemology, with its emphasis on warranted assertability and its downplaying of the notion of truth. At the level of moral discourse, of course, our standards of justification do not have much to do with predictive success. They are, simply, our ordinary moral standards, i.e. our unforsakeable concerns, motives, etc. Taking the truth of moral evaluative “with a grain of salt” will not confine us within the “vulgar subjectivism” which Wiggins rightly indicts as «what has given subjectivism a bad name» (2002 [1987c], p.208). We can take truth *cum grano salis* and acknowledge that moral standards are intersubjective.

In my view, the most important weakness in Dennett’s framework concerns his theory of interpretation (the theory of intentional systems), which involves too instrumentalist a conception of beliefs. Dennett’s view that there is nothing more to the beliefs entertained by someone than the predictive power that belief ascription can afford, is not convincing. On this count, Robert Brandom has a more realist theory, according to which belief content is determined by the articulation of deontic statuses (commitments and entitlements). Although this is
not the place to argue for this proposal, I take Brandom’s theory to be naturalizable and compatible with Dennett’s framework.

If I am right, we can “forsake” Wiggins’ framework, centered as it is on a Davidsonian theory of interpretation, adopt Dennett’s rationally acceptable language and its attendant pragmatist epistemology and supple, comparative ontology. Then we can integrate within Dennett’s framework Wiggins’ Humean theory of human nature; his view of the sphere of the unforsakeable; and his in-depth, insightful reading of other philosophers and traditions, including his critique of reductionism which, as I hope to have shown, does not apply to Dennett.

Dennett barely discusses the issue of moral meaningfulness as such. But with his framework and his moral anthropology, duly completed by that of Wiggins, we may acknowledge how our conception of the good is underdetermined by human nature, and assess whether some things are apt to make life morally significant. Then we will have a naturalist sense of Camus’ sentence: “Quelque chose a du sens, enfin, que nous devons conquérir sur le non-sens.” (Camus, *L’homme révolté*)
BIBLIOGRAPHY


