Discourse ethics is a theory of moral validity, initiated in the early 1970s by Karl-Otto Apel as a contribution to moral philosophy and subsequently developed within the framework of Frankfurt School Critical Theory by Jürgen Habermas. It is both a moral theory in its own right and a core ingredient of Habermas's critical social theory. Discourse ethics offers answers to central questions of moral philosophy, such as whether claims for the validity of moral judgments can be vindicated rationally and, if so, whether their rationality is universal in reach or confined to a particular socio-cultural context. In the context of Frankfurt School Critical Theory, it addresses the central question of the rationality of critique of a social order. This is the focus of the present essay. I consider discourse ethics primarily from the point of view of the place it occupies in Habermas's critical social theory, which endeavors to elaborate a universalist conception of reason internal to the linguistic practices of modern social orders. By means of this concept of reason, called “communicative rationality,” Habermas seeks to overcome problems that he sees as endemic to the mode of critique practiced by Horkheimer and Adorno in their jointly written *Dialectic of Enlightenment* (Horkheimer and Adorno 2002) and, more generally, within early Frankfurt School Critical Theory (Habermas 1984: 372–399).

**The Place of Discourse Ethics in Habermas’s Critical Theory**

*Dialectic of Enlightenment* tells the story of the development of human rationality from myth to enlightenment—from the earliest human societies characterized by mythical worldviews to modern societies governed by enlightened reason. The story culminates with reason’s relapse into barbarism, exemplified in the first decades of the twentieth century by National Socialism in Germany, Stalinism in the Soviet Union, and consumerism and the “culture industry” in the USA. Horkheimer and Adorno present this barbarism as the result of the degeneration of enlightened reason into a socially pervasive, instrumental rationality. With the advance of capitalism alongside scientific and technological developments, modern social orders become all-encompassing systems of instrumentalizing relations: humans relate to other humans, to the material and natural world, and to their inner selves, primarily as objects to be used for their own ends.

Habermas does not dispute the authors’ characterization of National Socialism, Stalinist communism, and capitalist consumerism as forms of barbarism. His dispute, rather, is with their apparent rejection of Western modernity’s project of rational enlightenment. In his view, Horkheimer and Adorno are too pessimistic with regard to the potential for human emancipation.
inherent in modern societies (Habermas 1981). Moreover, as a consequence of their pessimism, Horkheimer and Adorno’s critique of society is aporetic. They criticize the barbaric nature of purportedly enlightened societies by appealing to an idea of reason that is no longer available to the inhabitants of such societies; for in their account the dominance of instrumental rationality has rendered an alternative idea of reason unthinkable and inexpressible, except, on occasion, through works of art.

In response Habermas insists that modernity is an unfinished project (Habermas 1981). To continue the project, critical theorists must keep open the horizons of modernity through a willingness and ability to learn from history and other cultures (Habermas 1992: 138, 2001: 38–57). He maintains, furthermore, that modernity must generate its own normativity—its own standards for rational critique (Habermas 1987b: 7). The challenge is to elaborate a conception of reason that is at once part of existing social life and reaches beyond it—reason must be conceived of as “immanent transcendence” or “transcendence from within” (Habermas 1996: 5, 17, 2003: 7). Since *Structural Transformation of the Public Sphere*, one of his earliest published works, he has endeavored to show how capitalist modernity itself offers emancipatory modes of human communication (Habermas 1991). In other words, he holds that capitalist modernity contains within it the potential to enable human freedom and a genuinely rational society—one emancipated from authoritarian thinking and behavior and from hierarchical and exclusionary social structures. In the 1960s, his systematic theoretical reflections on this rational potential were formulated as a theory of knowledge and human interests (Habermas 1972). In the early 1970s, he took a different tack, favoring a language-theoretical approach.

This enabled him to move Frankfurt School Critical Theory—his own as well as that of his predecessors—in a new direction. Two aspects of his reorientation of Critical Theory are especially significant. First, his reflections on language led him to develop a formal-pragmatic account of the validity basis of everyday speech. This pragmatic version of the “linguistic turn,” inspired by the psychologist and linguist Karl Bühler, by Ludwig Wittgenstein’s use-theory of language, and by the speech act theory of J. L. Austin and John Searle, allowed him to make a theoretically crucial paradigm shift from a subject-object model of action and cognition to an intersubjective one. In the intersubjective paradigm embraced by Habermas, the traditional focus on relations between subjects and their objects gives way to a triadic model of linguistic communication, in which subjects relate to other subjects intentionally by way of speech acts concerning objects or states of affairs (Cooke 1994; Habermas 1998a).

Second, this intersubjective model of linguistic communication laid the philosophical foundations for the concepts of communicative action and communicative rationality, on which the critical power of his social theory now rests. In both cases, foundations are understood in a non-foundationalist way. In this respect, Habermas embraces the idea of a critical theory of society set out by Horkheimer in “Traditional and Critical Theory” (Horkheimer 1972). A critical theory works in conjunction with both traditional philosophy and the empirical sciences and social sciences, drawing on them for support for its theses. Thus, methodologically, a critical theory is not freestanding but rather challenged or substantiated by “truths” established in other disciplines. In line with this, it does not make *a priori* claims to truth; instead, its validity depends on whether its theses prove their truth in the actual practice of human life. On Horkheimer’s understanding, a critical theory makes claims to truth that are always fallible, open to challenge on the basis of actual human experience. Habermas adopts this methodological approach. In developing his discourse theory, he looks for support from theories of language (above all, Bühler, Wittgenstein, Austin, Searle, and Chomsky) as well as from the developmental psychology of Jean Piaget and Lawrence Kohlberg. Thus, for example, his initial reflections on moral validity were accompanied by essays on the stages of moral consciousness that were influenced by Kohlberg’s work.
As part of his reorientation of Critical Theory, Habermas began to articulate a discourse theory of validity, in which “discourse” is the name for validity-oriented forms of argumentation. His concern initially was with truth and moral validity (Habermas 2009: 208–269). During this period, in tandem with discourse theory, he developed a theory of communicative action (TCA), published in two volumes in 1981 (Habermas 1984 & 1987a). This is a normatively motivated, sociological account of advanced capitalist societies, in which communicative rationality is threatened by the expansion of the functionalist rationality necessary for the operation of financial and administrative systems. Here, moral philosophy takes a back seat to social theory. His main concern in TCA is to elaborate a theoretical framework for the critique of modern societies. In contrast to Horkheimer and Adorno, as he reads them, his basis for critique is not an idea of rationality external to the operation of modern societies, but rather one that is an integral part of it. Nonetheless, Habermas is insistent that he has always regarded moral philosophy as part of his critical social theory, never as a freestanding enterprise. Indeed, he sees this as the fundamental difference between his and Apel's discourse ethics: Apel develops his discourse theory of morality within the discipline of moral philosophy, whereas Habermas adopts an interdisciplinary approach (Habermas 2009: 15).

Habermas's initial reflections on moral validity were prompted by the sociological observation, influenced by Émile Durkheim, that modern social orders are ultimately maintained and reproduced only through the force-free recognition by their inhabitants of the moral validity of their basic institutions and structures. This is the topic of *Legitimation Crisis* (Habermas 1976). It is also the conclusion he draws from his discussion of the “linguisticization of the sacred,” a key component of his theory of modernity. It offers a genetic account of the modern worldview as a historical process in which “the sacred” comes to reside in everyday linguistic practices and is, in this sense, secularized (Habermas 1987: 77–111). In these works, Habermas uses the concept of morality in an undifferentiated way. In his subsequent writings on discourse ethics, as we shall see, he distinguishes morality first from ethics and then from political legitimacy. Moreover, the theory of moral validity he sketches in these earlier works is primarily a sociological account of the moral basis of social order. It is important to see that the sociological account is at once explanatory and normative: it both explains why social orders are threatened when their moral basis is eroded and offers a vantage point for critique of social orders that fail to supply their inhabitants with the moral motivation to obey structures of authority. This links it with his later program of discourse ethics, which is at once an explication of what it means to act morally and a justification of a universalist moral principle that defines the “moral point of view” (note, however, that in his discourse ethics Habermas insists that the moral point of view is not in itself motivating and relies on moral socialization processes).

In order to grasp the status of Habermas's discourse ethics, therefore, it is crucial to recognize three things. First, it is not a stand-alone enterprise, but rather part of a broader discourse theory, itself part of a more encompassing theory of communicative action, which, in turn, is located within a theory of modernity (Habermas 1990a: 116). Second, its justification is not a task for philosophy alone but involves multiple lines of argument, from various disciplines, all of which are open to challenge on the basis of new empirical findings and theoretical advances. Third, explanation/explication and justification cannot be neatly separated in Habermas's theory but are entwined in complex ways.

The dual explicative and justificatory character of Habermas's discourse ethics is not readily discernible from the title of his seminal essay on the topic, which appeared in 1983 in the wake of TCA. In “Discourse Ethics: Notes on a Program of Philosophical Justification” (DE), he describes his project as a concern to address the philosophical question of what it means rationally to *justify* moral norms and principles (Habermas 1990a: 43–115). It should
be noted, however, that the answer he gives is congruent with his *explication* of moral validity in the earlier sociological writings. There he argued that

we cannot explain the validity claim of norms without recourse to rationally motivated agreement [...] The appropriate model is [...] the communication community [Kommunikationsgemeinschaft] of those affected, who as participants in a practical discourse test the validity claims of norms and, to the extent that they accept them with reasons, arrive at the conviction that in the given circumstances the proposed norms are ‘right’.

(Habermas 1976: 105)

Indeed, the main ingredients of discourse ethics were already in place in his earlier, more evidently explicative works. In writings prior to DE, Habermas explicates moral validity as a *cognitivist* ethics, in which moral norms are held to have a relation to truth; as a *diagonal* ethics, in which the validity of norms rests on the supposition that they could be vindicated discursively by an agreement reached among participants in real argumentations; and as a *procedural* ethics, in which a validly conducted argumentative procedure determines the validity of the outcome (Habermas 1976). Nonetheless, there is a crucial difference between the earlier account of moral validity and his subsequent formulations.

In the earlier account, he expressly omits a maxim of universalization from his explication of moral validity. In DE, by contrast, a principle of universalization (U) is specified as the distinctive principle, governing the search for the right answer by the participants in moral argumentation. In his earlier sociological writings, any such moral principle is dismissed as superfluous: “A cognitivist linguistic ethics [Sprachethik] has no need of principles. It is based only on fundamental norms of rational speech that we must always presuppose if we discourse at all” (Habermas 1976: 110).

**The Development of Habermas’s Discourse Ethics**

From the beginning of his discourse ethics program, Habermas has described himself as following Kant in the attempt to answer the question of what it means to act rightly in a moral sense. Like Kant, he limits morality to the class of universally justifiable normative judgments, leaving aside matters of “the good life.” Thus, he demarcates ethics, as the doctrine of the good life, from moral theory as an account of the validity of universal norms and principles (Habermas 1990a: 196–197). Given his sharp distinction between ethics and morality, “discourse ethics” is a misleading name for his program, since what he proposes is, in his own terminology, a theory of moral validity. There is something to be said for preferring the name “discourse morality” (Baynes 2016: 99).

Like Kant’s moral philosophy, Habermas’s discourse ethics is deontological, cognitivist, universalist, and formalist. It is *deontological* in the sense that it attributes an imperative, binding force to moral norms analogous to the unconditional character of truth claims. It is *cognitivist* in the sense that it answers the question of how to rationally justify normative statements, whereby rationality has a context-transcending meaning and is construed in terms of universalizable interests, abstracting from particular needs, desires, and value orientations. It is *universalist* in the sense that norms and principles, when morally valid, do not merely reflect the intuitions of particular individuals, groups, cultures or epochs, but hold good universally. Thus, his moral principle “is not just a reflection of the prejudices of adult, white, well-educated, Western males of today” (Habermas 1990a: 197). Finally, it is *formalist* in the sense that the distinction between valid and invalid norms is not made on the basis of their particular content but is rather decided by a formal principle of universalization.
Discourse Ethics

In DE, Habermas offers the following definition of his principle of universalization (U): “All affected can accept the consequences and side-effects its general observance can be anticipated to have for the satisfaction of everyone’s interests (and these consequences are preferred to those of known alternative possibilities for regulation)” (Habermas 1990a: 65). There is an evident analogy here with the Kantian principle of universalization, the categorical imperative, especially with its first formulation: “Act only in accordance with that maxim through which you can at the same time will that it become a universal law” (Kant 2002: 137).

While acknowledging the close affinities between Kantian moral philosophy and his own discourse ethics, Habermas insists on significant differences between the two approaches (Habermas 1990a: 203–204). First, discourse ethics is dialogical in the sense outlined earlier: norms are valid if they could be vindicated by an agreement reached among participants in real argumentations (guided by idealizing suppositions); by contrast Kant assumes that individuals can test the validity of their maxims of action “monologically,” in isolation from others. Second, it is a de-transcendentalized version of Kantian ethics in two respects. To begin with, it de-transcendentalizes reason. It gives up Kant’s dichotomy between an intelligible realm comprising duty and free will and a phenomenal realm comprising inclination, subjective motives, and political and social institutions. By contrast, discourse ethics posits a relation of productive tension between the intelligible and the phenomenal—between what Habermas refers to as “immanence and transcendence” or “facticity and validity.” More precisely, discourse ethics posits a tension between actual human behavior within social and political institutions, on the one side, and, on the other side, certain presuppositions, which participants unavoidably make when they engage in argumentation (discourse); these are idealizing in the sense that they reach beyond—transcend—actual human practices. Argumentation, in turn, is a form of communication embedded within—immanent to—the everyday communicative practices of the inhabitants of modern societies.

In addition, its method is de-transcendentalized. It replaces Kant’s transcendental deduction of the moral principle with a formal-pragmatic argument based on the rational reconstruction of necessary presuppositions of argumentation in general. It may be noted that this methodological difference constitutes a further significant point of disagreement between Apel and Habermas. Apel’s discourse ethics favors a strong form of transcendental analysis, making a claim to ultimate justification for the moral rules identified by the moral philosopher, thereby establishing a secure basis for unerring moral knowledge (Habermas, 1990a: 94–98; Apel 1980). By contrast, Habermas describes the status of his analysis as relatively weak, insisting on the hypothetical and fallible character of his theses, even though their substance is universalist and thus “very strong” (Habermas 1990: 116).

In DE, Habermas’s justification of moral validity proceeds in two steps. Step one introduces (D), the discourse principle. “Only those norms can claim to be valid that meet (or could meet) with the approval of all affected in their capacity as participants in a practical discourse” (Habermas 1990a: 66). (D) is based on a rational reconstruction of the presuppositions underpinning participation in argumentation. Presuppositions of this kind do not have regulative force even though they point beyond actually existing social conditions in an idealizing manner; they do not impose obligations to act rationally, but rather make possible the practice that participants understand as argumentation (Habermas 1993: 31). His thought here is that participants in argumentation cannot in good faith describe themselves as engaging in argumentation unless they make certain (usually counterfactual) assumptions. Put slightly differently, they must attribute to themselves and to others certain normative commitments. The—not exhaustive—list of unavoidable assumptions (normative commitments) includes the following: all relevant arguments and affected persons are included in the discussion, each participant is granted equal opportunity to contribute to the discussion,
each participant is truthful, the discussion is free from all force except that of the better argument, and all participants are concerned with reaching agreement through the exchange of arguments on the answer to the question under discussion (Habermas 1993: 31). These unavoidable suppositions amount to the projection of an ideal communicative situation, in which a perfect argumentative procedure would ensure the validity of the argumentation’s outcome. Following widespread criticism of his idea of an “ideal speech situation,” Habermas has warned against misunderstandings. It is not a condition that could ever actually be realized by human beings but rather a “methodological fiction” or “thought-experiment,” which helps to explicate the meaning of validity in a context-transcending, universally valid sense (Habermas 1996: 323).

Step two of his justification of moral validity introduces (U), the principle of universalization. As we have seen, this expresses the Kantian intuition that moral agents can justify their decision only through appeal to generalizable interests. (U) is the “rule of argumentation” that determines what counts as a rational agreement in moral discourses. How does Habermas arrive at (U)?

In DE, Habermas contends that that the unavoidable suppositions of argumentation, as summarized by (D), combined with the knowledge of what it means to justify an action norm, provide a transcendental-pragmatic derivation or induction of (U) (Habermas 1990a: 92–94). Inductive reasoning, by contrast with deductive logic, allows for the possibility that the conclusion is false, even if all the premises are correct. Thus, he describes (U) as a “bridging principle” analogous to principles of induction in the empirical sciences, which bridge the gap between particular observations and general hypotheses (Habermas 1990a: 63). In an essay published in the same volume as DE, he compares it to reflective equilibrium in John Rawls’s Theory of Justice: “a reconstruction of the everyday intuitions underlying the impartial judgment of moral conflicts of action” (Rawls 1971; Habermas 1990a: 116).

Criticisms of Habermas’s Discourse Ethics

DE has been the subject of hard-hitting criticism from several angles. The most common criticisms are: Habermas’s justification of (U) fails (Wellmer 1991; Finlayson 2000; Gunnarsson 2000; Lafont 2003); that he confuses moral validity and democratic legitimacy (McCarthy 1991; Wellmer 1991; Lafont 2003); that its formalist, proceduralist approach to moral validity is unable to accommodate important kinds of moral concerns, experiences, and intuitions (Taylor 1991; Benhabib 1992); and that it is a bad explication of moral validity because it fails to capture the sense of absolute necessity attached to moral validity on the cognitivist account proposed by Habermas (Wellmer 1991; Lafont 2003; Cooke 2013).

Habermas has made some modifications to his initial formulation of the theory and provided helpful clarifications; together, they address many of these criticisms. One important modification is his reconfiguration of discourse theory as a theory of interconnecting discourses—as a kind of network theory. He takes a first step in this direction in Justification and Application (Habermas 1993). Here he expands the category of discourse to include ethical discourses, which are concerned with questions of the “good life,” and pragmatic discourses, which are concerned with prudential questions of how to act in a specific context (Habermas 1993: 1–18). Up to then, Habermas had reserved the term “discourse” for forms of argumentation in which participants necessarily suppose the approximate satisfaction of idealizing conditions relating to access, conduct, and the common search for the single right answer. Only discourses concerned with questions of truth (“theoretical discourses”) and those concerned with moral validity (“moral-practical discourses”) were discourses in the strict sense. Other argumentative forms, such as aesthetic deliberation, were characterized as “critique” (Habermas 1984: 42). He has now abandoned this terminological restriction,
Discourse ethics referring to ethical-existential and pragmatic argumentations as discourses (Habermas 1993: 1–18). Shortly after this first expansion of the term "discourse," he expanded his terminology yet again to include legal-political discourses, in which ethical, moral, and pragmatic questions are interconnected, and discourses of application, which seek to determine how abstract moral principles and norms should be applied in particular cases (Habermas 1996). It should be noted, however, that his original conceptual distinction persists within the broader category of discourses. On one side, there are discourses concerned with thematizing pragmatic, ethical-existential, ethical-political or legal-political matters, or with applying laws, ordinances, and policies appropriately through reference to context-specific norms. On the other side, there are discourses concerned with justifying the truth of propositions and of decontextualized moral principles or rules through reference to an idea of unconditional validity. Only discourses concerned with moral validity or with propositional truth are cognitivist in the sense that they make claims to universal rational validity in a context-transcending sense.

This expansion of the category of discourse helps to address the criticism that discourse theory has nothing to say about ethical questions, which are often experienced as more pressing and more difficult than questions of moral justification in the narrow sense, and seem at least equally in need of argumentative probing. In the modified version of discourse theory, ethical validity claims, too, may be the subject of discourses. Furthermore, Habermas clarifies that ethical questions, like moral questions, carry a sense of obligation and may have a context-transcending reference point (Habermas 1993: 5). The relevant distinction between moral and ethical questions, therefore, is one of universality, not of logical form (cf. Heath 2014: 845). More precisely, the relevant distinction is between argumentations that are tied conceptually to the idea of universal rational agreement (moral discourses) and those that are not (ethical discourses). According to Habermas, ethical discourses may be concerned with context-transcending ideas of the good for human beings, but they do not rest on the idealizing supposition that a rational consensus as to the single right answer is achievable; this is because the modern de-transcendentalization of reason, combined with the fact of value pluralism, excludes the possibility of a discursively reached agreement regarding the good for human beings.

Habermas builds on his expanded category of discourse to reconfigure discourse theory as a theory of interconnected discourses (Habermas 1996). This enables him to avoid the criticism that he conflates moral validity with political legitimacy. He now refers to an overarching discourse principle, which explains in a general way the point of view from which action norms can be justified impartially (Habermas 1996: 107). The discourse principle provides the basis for a differentiated account of types of practical argumentation, in which discourses are distinguished from bargaining procedures and in which various sub-types of discourses are specified. This permits a distinction between the principle of democracy and the strictly moral principle, enabling a corresponding distinction between democratic legitimacy and moral validity. As before, he defines moral validity in terms of a principle of universalization (U). Democratic legitimacy, by contrast, requires only that all citizens agree on the validity of the norms, principles, laws, ordinances, and policies that are at stake in a given process of deliberation. He formulates the principle of democracy as follows:

"Only those statutes may claim legitimacy that can meet with the assent of all citizens in a discursive process of legislation that in turn has been legally constituted."

(Habermas 1996: 110)

The main differences between the principles of democracy and morality are that they have different reference groups and that they regulate different matters in a different way. The
moral universe is unlimited in space and historical time, encompassing all natural persons in their life-historical complexity; a democratic (legal-political) community, by contrast, protects the integrity of its members only insofar as they acquire the status of bearers of individual rights (Habermas 1996: 451–452). Furthermore, moral norms regulate interpersonal conflicts from the point of view of impartiality and are strictly bound by the principle of universalizability; democratic decisions and norms, by contrast, draw on moral, pragmatic, and ethical considerations in order to give binding force to collective goals and programs (Habermas 2008: 93). He rejects what he calls moralistic misunderstandings of the democratic principle of legitimacy on account of their subordination of law to morality (Habermas 1996: 229–232). He insists that “[t]he democratic lawmakers procedure must exploit the rational potential of deliberations across the full spectrum of possible aspects of validity, and by no means merely the moral aspect of the universalizability of interests” (Habermas 2008: 93). A problem here is that Habermas does not make clear whether legal-political norms, despite their context-specificity, have the cognitive meaning he attributes to moral norms (Cooke 2013). On the one hand, he evidently sees democratic politics as oriented toward truth (Habermas 2008: 143–144). In line with this, he interprets the principle of majority rule in terms of a search for truth among citizens; he also criticizes Rawls for suspending the truth question and adopting, instead, the category of the “reasonable” (Habermas 1996: 179, 1998b: 49–73). On the other hand, consistency would require him to deny a cognitive meaning to legal-political norms. For, just like ethical discourses as he construes them, participants in legal-political discourses do not anticipate universal agreement. This is due to the context-specificity of the discussions. The agreement sought is not one among all humans but rather among a demarcated group (“citizens”). Furthermore, legal-political discourses appeal to ethical and pragmatic reasons as well as moral ones. As mentioned, Habermas holds that no universal agreement on ethical reasons is possible under conditions of modernity—such reasons are reasons only for particular groups in particular contexts. Pragmatic reasons, too, lack a reference to universal agreement for they answer the question of what a particular individual or group ought to do at a particular time in response to a particular challenge.

Habermas’s reconfiguration of discourse theory as a network theory of discourses also enables him to address objections to its formalism and proceduralism. Feminist writers in particular have expressed concern that discourse ethics leaves aside many kinds of questions that are morally relevant; furthermore, that it is insensitive to particular needs, aspirations, and life-experiences. Since moral discourses are not just one component in a network of interconnected discourses, neither their narrow focus nor their high level of abstraction is quite so troubling. Worries about their abstraction are further allayed by the inclusion within the discourse network of “discourses of application.” Habermas fully acknowledges the highly abstract character of moral norms and, drawing on the work of Klaus Günter, emphasizes that they must be applied in historically specific sociocultural contexts in response to actual disputes and conflicts (Habermas 1990b: 95–97, 2009: 21; Günter 1993). Furthermore, he points out that his account of moral validity is not a complete account of what it is to be a moral agent; nor is it a theory of moral motivation. It relies on the internalization, through socialization, of certain morally relevant competences, in particular, the willingness and ability to adopt the perspective of the other (“ideal role taking”) and, in general, a willingness to act morally, together with a basic knowledge of what this means (Habermas 2009: 17–19). He also draws attention to the link between morality and solidarity, since moral norms and principles, though highly abstract, speak to the vulnerability and neediness of human beings (Habermas 1990: 199–203, 2009: 17; cf. Rehg 1994).

Since DE, Habermas has modified and clarified the status of his moral principle (U). This helps him to deal with objections directed against discourse ethics as a program of
philosophical justification. Most of these objections relate to his claim in DE that (U) is justified by induction: His assertion that (U) can be justified inductively on the basis of a rational reconstruction of the general presuppositions of discourse (D), combined with the knowledge of what it means to justify an action norm. In line with the weak status he attributes to his strong theses, Habermas has made clear that he understands such knowledge in a weak sense: moral agents merely need to know that moral justifications consensually resolve disputes concerning the rightness of normative statements (Habermas 1993: 32–33). However, this justificatory strategy is less weak than he suggests. Closer consideration reveals that it relies on normative commitments that go beyond the commitments necessarily undertaken by participants in argumentation. To begin with, it relies on normative commitment to a particular view of moral conflict resolution: the view that moral disputes should be resolved consensually. This is a commitment that goes beyond knowing what it means to engage in argumentation. Habermas maintains, correctly, that the very concept of argumentation implies that the disputing parties, once they enter into argumentation, necessarily see themselves as engaged in the search for the right answer to the matter in dispute. But the claim that moral disputes should be resolved consensually makes more than this conceptual point. Without conceptual inconsistency, participants in moral argumentation could, for example, take the view that certain individuals or groups have privileged insight into what is morally right, which may not be recognized by everyone concerned, and that their insight should determine the outcome.

This worry is compounded by later formulations of (U), where the equality requirement, too, seems to go beyond normative commitments necessarily undertaken by participants in argumentation. In various publications following DE, Habermas writes that the foreseeable consequences and side effects of the general observance of a norm for the interests and value-orientations of each individual must be jointly acceptable to all concerned (e.g. Habermas 1998b: 42, cf. 2009: 16, emphasis in original); moreover, he holds that the corresponding practice must be in the equal interest of all (Habermas 1998b: 36). Indeed, these normative commitments soon become part of his account of knowing what it means to justify an action norm. In writings subsequent to DE he states that this is a matter of knowing that such norms “regulate problems of communal life in the common interest and thus are ‘equally good’ for all affected” (Habermas 1993: 33). Evidently, therefore, Habermas's justification of (U) relies on commitments to historically and socioculturally specific, substantive values of equality and consensual moral conflict resolution. In other words, Habermas's justification of (U) relies on ideas of equality and moral conflict resolution that are not universal, but rather the result of particular historical socialization processes; furthermore, within these socialization processes, they may compete with other socially established ideas of equality and moral conflict resolution. Or, put differently again, his justification of (U) presupposes the validity of his thesis of the linguistification of the sacred and, more generally, his theory of modernity.

There is some evidence that Habermas has conceded this point and modified his position accordingly. In an essay published in the mid-1990s, he acknowledged, at least implicitly, the dependency of discourse ethics on his linguistification thesis and the particular account of modernity to which it gives rise (Habermas 1998b: 45). Furthermore, in the same essay he described his justificatory strategy as abduction rather than induction (Habermas 1998b: 42). “Abduction” is Charles Saunders Peirce’s term for a kind of nondeductive inference that takes place at the initial stage of the process of inquiry, when inquirers arrive at their best guess about which hypotheses to select and subject to inductive testing. In another more recent essay, Habermas states explicitly that he sees (U) as just one of several available moral principles (Habermas 2009: 16). This raises the question of how to decide which principle is the best one. In his essay from the 1990s, Habermas proposes a pragmatic test, subject to the
condition that the chosen moral principle is able to select moral norms that could command universal agreement. He writes that “participants themselves will perhaps be satisfied with this (or a similar) rule of argumentation as long as it proves useful and does not lead to counterintuitive results. It must turn out that a practice of justification conducted in this manner selects norms that are capable of commanding universal agreement—for example, norms expressing human rights” (Habermas 1998b: 43). Notice that “usefulness” is a pragmatic standard of evaluation: a judgment or action is valid if it works—if it is deemed to be an effective solution to a problem in a given context. (Recall Habermas’s category of pragmatic discourses, which are concerned with prudential questions of how to act in a specific context.) However, given Habermas’s earlier insistence that moral validity claims are cognitive in the sense of having a relation to a universalist, context-transcending idea of truth, and given the importance of their truth-relation for his critical theory of society, a pragmatic standard of this kind could at best function merely as a preliminary step in choosing moral principles as candidates for further rational evaluation; it could not be used to determine the validity of moral principles such as (U). Moreover, in choosing certain moral principles and rejecting other ones, the moral theorist would have to be alert to possible ethnocentric and other biases. In fact, Habermas makes clear that in eventually deciding on one moral principle rather than another, “usefulness” is not the decisive factor, but rather its ability to generate norms capable of commanding universal agreement. The difficulty, as we have seen, is that he has not provided a satisfactory justification for this stipulation.

“Intuitive plausibility” fares no better as a standard for determining the validity of moral principles such as (U). While it too may have a place at the preliminary stage of selecting likely candidates, its limits become obvious when we consider its dependency on sociocultural interpretations of what it is to be human and what would constitute a good society. Habermas himself draws attention to this. He confronts the suspicion of possible ethnocentric biases, acknowledging that (U) may reflect a socioculturally specific conception of the good (Habermas 1998b: 43, 2009: 16). He claims to be able to dispel the suspicion through appeal to knowledge of what it means to engage in the practice of argumentation as such, together with knowledge of what it means to justify an action norm (Habermas 1998b: 43). But, to repeat: Habermas’s explications of what this entails involve normative commitments to particular conceptions of equality and moral conflict resolution that go beyond those necessarily made by participants in any argumentative practice. It is unsurprising, therefore, that some commentators find this response to the ethnocentric objection unsatisfactory, for it seems simply to reiterate the justificatory steps that have bothered them from the outset (Gunnarsson 2000: 120–123).

Other commentators see Habermas’s modifications and clarifications of the status of (U) as evidence of a significant departure from his original justificatory program (Finlayson 2000). However, our earlier discussion of the status of discourse ethics allows us to see them as in line with it. We will recall that discourse ethics was never conceived as a stand-alone enterprise but as part of a more general discourse theory within a broader theory of communicative action, in turn located within a theory of modernity. As a result, justification and explanation/explication were never neatly separable in Habermas’s theory, but entwined in complex ways. We will recall, too, that he never regarded his justification of moral validity as a task for philosophy alone, but as involving multiple lines of argument, from various disciplines, whose claims to validity are understood as inherently open to rational challenge. It could be argued, therefore, that a proper grasp of the status of Habermas’s discourse ethics deflects the criticism that his argument in DE fails to prove the universal validity of (U), since the argument in DE is not intended as a stand-alone justification of morality. Certainly, Habermas’s claims that (U) can be justified inductively or even abductively are not helpful in this regard. They distract from his point that the status he claims for his strong theses is
weak; they also close off avenues for an explicitly self-reflective discourse theory of morality, which would make no final claims as to what determines the validity of moral norms and allow room for rational contestation of the substantive values that (U) incorporates.

**Moral Validity Revisited**

I have suggested that with the help of certain modifications and clarifications Habermas has been able to deflect many of the criticisms that have been directed against his discourse ethics; furthermore, that certain criticisms of his justification of (U) may be due to a failure to grasp the manner in which philosophical justification and social theory are interconnected in Habermas's project and, in particular, the ways in which justification and explanation/explication are, and have always been, entwined. But discourse ethics remains vulnerable to the accusation that, in its formulations to date, it invites a strong reading of its justificatory claims, despite Habermas's assertion that the status claimed for his strong theses is weak. Furthermore, it smuggles substantive normative commitments into “knowing what it means to justify an action norm”, discouraging critical reflection on the validity of these commitments. In this concluding section, I revisit the objection that he has not provided a satisfactory account of the validity of the principle (U), which determines the validity of moral norms. However, rather than insist that he provide some inductive, abductive or other kind of justification of it, I take seriously his assertion that the status claimed for his argument is weak, together with his remark that (U) is just one of several available moral principles. Accordingly, I emphasize the importance of critical reflection on the very constructivist terms in which his discourse theory of morality is formulated: on its guiding idea that moral validity is defined in terms of discursively achieved universalizability. On a strong reading of the status of its justificatory claims, Habermas’s discourse theory of morality closes rather than opens the horizons of modernity, for it disallows critical reflection on the validity of its own constructivist approach. The strong reading leaves no conceptual space for critical theory to learn from religious and ethical beliefs, practices, and traditions that do not affirm Habermas's particular view of secular normativity, but embrace conceptions of truth or the good that are not discursively generated. The strong reading is not just invited by Habermas's apparent concern to justify (U) inductively or abductively; it is also invited by his theory of modernity. For this reason, I advocate a weaker reading of the status of both his discourse theory of morality and of his theory of modernity.

We will recall that his theory of modernity offers a genetic account of human history as a gradual process of linguistification. In this account, in the passage from mythical to modern worldviews the authority of the sacred is gradually replaced by that of secular moral norms. For Habermas, moral norms are secular in the sense that they appear only to ideas of normativity that are humanly constituted in processes of communicative action. In other words, the authority of norms and principles is entirely constructed by their subjects through the exchange of reasons in argumentation. This is why he holds that modernity must generate its own normativity. However, the thesis is stronger than his linguistification thesis warrants and is not entailed by the core elements of his account of modernity. The salient features of modernity, in his account, are a belief in the power of reason to question all established authorities (non-authoritarianism), together with a widespread commitment to the values of inclusion and equality. Non-authoritarianism does not depend on a belief that the normative basis of the authority of morality (or religion or political leaders or the law) is entirely a human construction. Nor does commitment to values of inclusion and equality depend on this belief.

Habermas's conception of moral validity is constructivist in the specific sense that its cognitive power is produced through the exchange of reasons in argumentation. In this
conception, moral norms and principles are not just tested in discourse; they are generated by discourse. The concept of moral validity is entirely discursive: moral validity is defined as an agreement reached argumentatively under idealized communicative conditions, in which conditions such as inclusion, fairness, truthfulness, absence of all force except that of the better argument, and concern for the single right answer have been met. It does not matter that such a condition is a “methodological fiction”: as we have seen, Habermas himself emphasizes that the “ideal speech situation” is not a condition that could ever actually be achieved. What matters is that the very concept of moral validity is defined in terms of this idealizing projection. The “ideal speech situation” is a conceptual thought-experiment. For the purposes of conceptualizing moral validity, it calls on us to imagine a social condition in which disputing parties arrive at norms and principles that are morally valid in an unconditional—absolute—sense.

In Habermas’s original formulation of discourse theory, both the concepts of propositional truth and of moral validity were defined in terms of a discursively reached agreement (Habermas 2009: 208–269). However, for a long time the constructivist character of these concepts did not play a significant role in his theory. He did not draw attention to it and many of his texts are ambiguous in this regard. For instance, nothing in DE appears to turn on the question of whether the normativity of (U) is produced in discourse: it seems immaterial whether (U) is an argumentative rule for testing the validity of moral norms or an engine for generating moral validity. Similarly, in the passage cited in the previous section, where Habermas makes the justification of (U) subject to a pragmatic test of usefulness and intuitive plausibility, he requires merely that the valid outcomes of moral discourses command agreement; he does not say that their very validity is constituted by the discursive process. Put differently, the passage leaves open whether discursive agreement is merely a necessary condition of the validity of human rights or a necessary and sufficient condition. While for many years Habermas seemed content to leave this question open, he now explicitly affirms the constructivist character of moral normativity. He contrasts the concept of moral validity with the concept of truth, which on his revised understanding lacks precisely this constructivist character (Habermas 2003: 237–275).

From the 1980s onwards, in response to critics of his discourse theory in general, Habermas began to revise his theory of propositional truth. He gradually distanced himself from his definition of truth as the outcome of discursive procedure. He replaced it with an idea of truth that is justification-transcendent, in the sense that it cannot be made to coincide even with the concept of ideal justification or “warranted assertibility” (Habermas 2003: 247–248). He now conceives of truth as a circular process in which argumentation fulfills the role of trouble-shooter with regard to everyday behavioral certainties that have become problematic. In such cases, the propositions in question become the subject of specialized discourses, in which experts probe the evidence-based arguments supporting or challenging them. The results of these argumentations are fed back as “truths” into everyday life. Strictly speaking, however, they constitute not truth, but judgments as to warranted assertibility, for they are inherently fallible, vulnerable to challenge when they no longer serve their pragmatic function in everyday behavior and practice. In his revised account of propositional truth, therefore, argumentatively reached agreement merely points toward truth in an unconditional sense (Habermas 2003: 252–256). At the same time, truth and justification remain internally connected: a discursively reached agreement authorizes truth (Habermas 2003: 258). Moral validity, by contrast, lacks the justification-transcendent character of the concept of truth (Habermas 2003: 256–261). A discursively reached agreement warrants the rightness of moral norms and principles; it does not merely authorize their rightness: “[i]dealized warranted assertibility is what we mean by moral rightness…it exhausts the meaning of normative rightness itself” (Habermas 2003: 258, emphasis in original). In this conception,
Discourse ethics

the moral world is “made by us”: the realm of morality is itself generated in discourse. By contrast with truth, which relates to an objective world deemed to have some independence of human agency, the very domain of moral validity is humanly, indeed argumentatively produced (Habermas 2003: 262).

Critics see this as a bad explication of moral validity, questioning whether Habermas’s account of moral validity captures the sense of absolute necessity he attaches to morality (Wellmer 1991; Lafont 2003). They contend that a non-constructivist conception of moral validity would enable him to provide a more satisfactory account of the unconditionally binding character of moral norms and principles. I share their view (Cooke 2013). However, my present focus is somewhat different. My objection here is that Habermas presents his constructivist account as though it were entailed by his linguistification thesis, impeding critical reflection on its validity. A theory closed in this way is unable to learn from its encounters with rival conceptions of context-transcending validity; similarly, it is unable to allow for mutual learning between “postmetaphysical” thinkers, who share Habermas’s constructivist account of validity, at least in the domain of practical reason (Habermas 1992, 2017), and those for whom the ultimate source of moral or ethical validity is non-linguistic, even non-human.

Habermas recognizes the multiple kinds of normativity that are operative in the communicative practices of modern societies (rules of grammar, rules of style, religious prescriptions, legal ordinances, moral commands, aesthetic principles, ethical guidelines, and so on). However, he holds that only moral norms and principles have a cognitive meaning, in the sense of having a relation to truth. We have seen that he grants a possible context-transcending reference point to ethical validity claims (a reference to some subject-transcending idea of the good). We have also seen, however, that he denies the possibility of a universal, discursively reached rational consensus as to the validity of ethical claims. The same holds for religious validity claims: he acknowledges their context-transcending reference point, but does not see them as open to discursive vindication (or even thorough-going discursive examination) (Habermas 2008: 129; Cooke 2013). From the point of view of Habermas’s critical theory, therefore, ethical and religious utterances have no cognitive meaning for they lack a relation to truth. This has worrying implications for the ability of his theory, and those who share its constructivist interpretation of context-transcending validity, to learn from ethical and religious beliefs, practices, and traditions (Cooke 2016).

In Habermas’ critical social theory, learning means socio-cultural learning and has a strong cognitive sense. It is a movement in the direction of truth or moral rightness. Participants in processes of socio-cultural learning are required to engage with their interlocutors as partners in the search for answers to questions that are true or morally valid. In other words, learning is conceived as mutual learning, which has the strong cognitive meaning he attaches to claims to propositional truth and to moral validity. If we probe this conception of learning, we can see that it calls for a shared understanding of the meaning of learning and hence, a shared conception of truth or moral validity. If participants in argumentation have fundamentally different conceptions of context-transcending validity, and by extension learning, they will not be able to see the outcome of their deliberations as mutual learning; at best, they will be able to say that they have learnt something of value for themselves. Think of an argumentative exchange between two parties who disagree on the question of whether freedom of religion is a universally valid moral principle. One party’s view of moral validity is utilitarian: she holds that a moral norm is valid only if it maximizes happiness. The other party’s view is Kantian: he holds that a moral norm is valid only if he could will that it becomes a universal law. In their argumentative exchange, perhaps over time, both parties might change their views with regard to the universal moral validity of the principle.
of freedom of religion; they might even change their views on the validity of a certain understanding of utilitarianism or of Kantian morality. Indeed, the substance of their views might converge in the end—for example, they might end up agreeing that freedom of religion is a universally valid moral principle. However, none of this is sufficient for the result to count as mutual learning in the strong cognitive sense in which Habermas understands learning. In order for it to count as mutual learning in this strong cognitive sense, the two parties, by way of their argumentative exchange, would also have to learn something together with respect to the very concept of moral validity. The same holds for argumentative exchanges between those who share Habermas's constructivist understanding of moral validity and those who think of moral validity as having some independence of discursively reached agreement. It holds even when all parties are committed to modern norms of non-authoritarianism, inclusion, and equality. Indeed, it holds even when all parties share Habermas's view that discursive justification is a necessary condition of moral validity. What divides participants in our imagined argumentative exchange is that some of them think of discursive justification as indicating or authorizing moral validity rather than as constituting it. This may be due to their religious beliefs, but it could also be due to a non-religious "metaphysical" understanding of the goods orienting our lives as human beings, together with a view of ethics and morality as interconnected (Taylor 1989). For the parties in the argumentative exchange to regard the outcome as mutual learning in Habermas's strong cognitive sense, they would also have to engage reflectively with the arguments for a constructivist understanding of moral validity vis-à-vis a non-constructivist understanding, and hold that they had learnt something about the strengths and weaknesses of the respective arguments. In other words, in order for the participants in an argumentative exchange to conceive of the outcome as mutual learning in the strong cognitive sense in which Habermas understands learning, they must also seek a common understanding of what moral validity means. But this implies a readiness on the part of those who share Habermas's constructivist view of moral validity to learn from those who do not, for example, from religious believers who hold that the ultimate source of the validity of moral norms is not human, but divine. Habermas's account of modern normativity, as generated from within human practices, seems to rule this out by fiat.

Not surprisingly, therefore, learning from religion, as he understands it, is a matter of appropriating the propositional content of religious teachings within a staunchly secular (though not "secularist") framework. In his recent writings on law and democracy, he speaks of "critical appropriation" of the contents of religious beliefs, practices, and traditions, of a methodological atheism/agnosticism with regard to the contents of religious traditions and of "salving" these contents. (Habermas 1991, 136–139, 1992: 14–15, 2008: 209–248).

In line with this, he calls for secular translations of religious utterances. He views the major world-religions as semantic reservoirs, which secular modern societies may draw on productively to enrich their moral vocabularies; however, the religious content in question must first be translated into a secular language in order to make it accessible to all members of society, irrespective of religious belief. His concern is not just accessibility: the underlying point is that only secular translations of religious utterances are open to thorough-going discursive examination and validation, since only secular translations have a relation to truth in the postmetaphysical sense embraced by Habermas. This makes learning from religion an exercise in which the postmetaphysical framework he deems appropriate for contemporary critical social theorists (and secular citizens) is immunized against rational challenge and revision.

Habermas seems committed to the view that only a constructivist understanding of normativity is appropriate for the inhabitants of modernity. Modernity must generate its own normativity—otherwise it will undo the historical learning process, which has enabled the rational contestation of established authorities and led to a widespread commitment to universalist values of inclusion and equality. I see this as a closing of the horizons of modernity.
Since he insists, against Horkheimer and Adorno, that modernity is an unfinished project, he cannot disregard this objection. If critical social theory is to keep open the horizons of modernity, it must be open to learning from religious traditions (and ethical worldviews), not just on the level of moral content, narrowly understood, but also with respect to its own post-metaphysical orientation. For this, it will have to adopt a critically reflective attitude to its particular constructivist conceptualization of context-transcending validity in the domain of practical reason (Cooke 2016).

There is a further reason, also internal to Habermas’s project, to urge the need for a critically reflective attitude to the constructivist terms in which he formulates his conception of moral validity: doing so would reopen the path for the more extensive critical task he gave to discourse ethics in his earlier writings. In these writings, we will recall, the social order as a whole was deemed to have a moral basis; in consequence, every aspect of social life was open to challenge on moral grounds. From DE onwards, the scope of morality became much more limited; it was no longer viewed as a critical tool for evaluating judgments and actions relating to the good life for humans, and the kind of society that would enable such a life. Questions of the good life and good society became the domain of ethical reasoning. But this, lacking the robustly cognitive conception of context-transcending validity that Habermas attributes to moral reasoning, is not suitable for the purposes of thoroughgoing critique of a given social order. What kind of reasoning is appropriate? Habermas’s answer is not clear. Certainly, he is clear that a critical theory of society in the Frankfurt School Tradition cannot lightly dispense with a cognitively construed, context-transcending conception of reason: his theory of communicative rationality is intended to meet exactly this challenge. What remains unclear is how he understands the kind of validity claim raised by critical social theory for its utopian projections of a social order that would instantiate the emancipatory promise of modernity and avoid its pathologies. Since these are projections of the good for humans, it would make sense to characterize them as ethical validity claims.

As things stand, however, this path is not available to Habermas unless he gives up his commitment to a cognitively construed, context-transcending conception of critique. The alternative is to give up his thesis that a discursively achieved, universally binding, rational agreement defines the concept of context-transcending validity in the domain of practical reason, together with the sharp distinction between morality and ethics that follows from this thesis. For if Habermas is correct that the validity of ethical claims cannot be construed as a universally binding, rational agreement reached in a discursive procedure, the required conception of moral/ethical validity could not be formulated in the constructivist terms in which he formulates his idea of moral validity. A non-constructivist account, or different kind of constructivist account, could allow for a conception of moral/ethical validity with cognitively construed, context-transcending power, but one in which agreement reached in an idealized argumentative procedure is not a definition of moral/ethical truth. Thus, the key question, as I see it, is not whether Habermas succeeds in demonstrating the universal validity of his moral principle (U); rather it is whether his discourse theory of morality provides the basis for a conception of rational critique of an existing social order that claims validity in a context-transcending sense, while avoiding ethnocentric, cultural, gender, and other biases.

References
MAEVE COOKE


DISCOURSE ETHICS

Further Reading


