THE VIRTUOUS CIRCLE OF DISCOURSE:

WHY HABERMASIAN CRITICAL THEORY IS BLIND TO SOCIAL TRAPS

by

Agustín Alonso Goenaga Orrego

B. A., Instituto Tecnológico y de Estudios Superiores de Occidente, 2007

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

The Faculty of Graduate Studies

(Political Science)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

September 2009

Abstract

This paper raises a critique concerning the limitations that Habermas's Theory of Communicative Action (TCA) faces to engage with topics such as social traps. The main argument is that the developmentalist explanation of ego ontogenesis that supports the possibility of the "discourse principle" reduces the TCA's effectiveness to explain social transformation in less-than-ideal situations. The paper introduces the concept of social traps through Rothstein's account of the recursive dynamics between individual agents and social structures in non-cooperative scenarios, and follows his criticisms of rationalist and culturalist approaches. The same problems that these strands of the literature suffer are present, although for different reasons, in the TCA: wrong assumptions about human action, deterministic views of social reality, and a narrow understanding of social transformation. Along these lines, the paper explores the implications that the Habermasian notion of praxis as discursive competences and autonomy has for a wider conception of agency in cases where discourse is inhibited or disrupted. Moreover, this becomes a real problem due to the circular relationship between, on the one hand, the development of the cognitive and moral competences necessary to participate in discursive practices and, on the other, how these practices foster the very same competences that they require to prevail ("the virtuous circle of discourse"). The combination of these elements raises a number of challenges for the TCA to provide convincing explanations about the way in which social traps operate and, especially, how social transformation can be generated in those situations from the micro-level. Finally, the author concludes with some suggestions to be considered in the development of a critical methodology to observe social traps.

Table of Contents

Acknowledgements

I want to thank my supervisors, Dr. Mark E. Warren and Dr. Maxwell A. Cameron, for their guidance, encouragement, and teachings. In the process of developing the project, I received insightful and critical observations from Dr. Bruce Baum, Priya Bala-Miller, Aubin Calvert, Sean Gray, and Kelsey Seymour. Any errors are solely my responsibility.

This thesis was written with the financial support of the Canadian Bureau for International Education (CBIE) as part of a Canada Government Award.

Finally, my permanent gratefulness goes to my parents for their unending generosity and love.

Introduction

Many of the most pressing challenges facing contemporary developing democracies are related to problems of cooperation and action coordination through democratic procedures: pervasive corruption, mafia organizations, oligarchic enclaves that impinge upon elections and popular representation, lack of trust between citizens of different social groups and between citizens and official authorities, inefficient fiscal systems, free-riders, black markets, and so on. The bottom-line question is how to build in contemporary societies a democratic rule of law —conflict-resolution and collective decision-making through democratically institutionalized legal procedures— when social dynamics are hostile to it. In cases like these, actors are usually entrenched within their private interests and unwilling to cooperate with one another, either because they do not recognize the legitimacy of the ongoing collective regulatory instruments (the current legal and judicial institutions) or because they do not trust that their counterparts will live up to the norms embedded in those agreements.

So far, Critical Theory has focused on discussing the first of these two options, that is, the consequences of challenging hegemonic normative frameworks and the way in which power relations intervene in the construction and perpetuation of those norms and values. Jürgen Habermas in particular has advanced a complex theory about the role of normativity in law and how it can be constructed through discourse in an ideal speech situation (1984; 1987; 1996). The Habermasian project for Critical Theory provides a normative grounding, as inclusive as possible, for the instruments of collective regulation embodied in law in post-conventional societies. This contribution represents an ambitious and promising attempt to address the first of the reasons for which the democratic rule of law might be challenged, stimulating for the past thirty years a burgeoning literature on the topic (Bohman, 1994; 2007; Dryzek, 1994; 2002; Elster, 1998; Goodin & Niemeyer, 2003; Gutmann & Thompson, 2004; Habermas, 1996; Held, 2006; Niemeyer & Dryzek, 2007; Warren, 1992; 1993; 1995). Nevertheless, Critical Theory has until now given little attention to the second of the aforementioned obstacles to coordinate collective action, and, instead, other approaches have taken the lead in the study of "social traps".

Indeed, problems of action coordination and cooperation are usually intertwined with conflicts about norms, knowledge, and interests tilted by asymmetrical power relations. Therefore, these problems need to be taken into account if Critical Theory is to maintain its emancipatory objectives. In this paper I will refer to Critical Theory, mostly through Habermas's Theory of Communicative Action (TCA), not as a general theory but as "a method of analysis deriving from a nonpositivist epistemology" (Antonio, 1981:332). The question that I will try to raise concerns the limitations that the TCA, as the mainstream thread of Critical Theory, faces to engage with topics such as social traps. Moreover, what are the implications that these limitations have for the broader project of Critical Theory? Even if the critique that Horkheimer, Adorno, and Marcuse started in the 1930s and 1940s was focused on challenging the reified rationality of advanced capitalist societies (Couzens Hoy and McCarthy, 1994; Horkheimer and Adorno, [1944] 2002; Marcuse, [1964] 1991; Nagel, 2008; Wiggershaus, 1994), to ignore situations such as social traps because they mostly happen in developing countries would be today, under the global interconnectedness of the world, a fallacious excuse. Furthermore, the dynamics inherent of social traps are present even at the core of the advanced capitalist societies once we observe how stratified layers of social integration, along the lines of gender, race, ethnicity, and class, become recurrent obstacles for action coordination between different demographic groups. The path that Critical Theory took with Habermas has increasingly moved away from these concerns. His attempt to escape from the iron-cage of instrumental rationality by developing a critique of ideology based on communicative action has paid the price of turning itself blind to forms of self-imposed coercion and "self-frustration of conscious human action" other than technocratic domination.[1] Under these premises, to argue that Habermasian Critical Theory is blind to social traps is

[1] It would certainly be naïve to expect a theory to solve every problem and be effective in every imaginable situation. However, the central premise of Critical Theory has been to push knowledge towards emancipatory purposes. According to Raymond Geuss (1981),

> "the Frankfurt account of the essential distinguishing features of a 'critical theory' consists of three theses:
>
> 1. Critical theories have special standing as guides for human action in that:
>   a. They are aimed at producing enlightenment in the agents who hold them, i.e. at enabling those agents to determine what their true interests are;
>   b. They are inherently emancipatory, i.e. they free agents from a kind of coercion which is at least partly self-imposed, from self-frustration of conscious human action.
> 2. Critical theories have cognitive content, i.e. they are forms of knowledge.
> 3. Critical theories differ epistemologically in essential ways from theories in the natural sciences. Theories in natural science are 'objectifying'; critical theories are 'reflective'" (Geuss, 1981:2).

not just an unfair attempt to raise the bar too high in measuring the explanatory (and emancipatory) power of the TCA as a critical methodology for social research (Antonio, 1981; Prasad, 2005; Zanetti, 1997). Social traps represent one of the most common factors that hinder agents from determining, in the words of Raymond Geuss (1981), “what their true interests are”. If this is the case, the TCA’s inability to engage with action coordination problems in general, and particularly in the developing world, narrows its scope as a Critical Theory and makes it vulnerable to the critiques of subaltern (i.e., Nagel, 2008; Willet, 2001) and feminist scholars (i.e., Fraser, 1985).

David Held pointed out in 1981 that Habermas had failed so far to adequately respond: “To whom is critical theory addressed? How, in any concrete situation, can critical theory be applied” (Held in Livesay, 1985:68)? I sustain that an unacknowledged ethnocentrism would become evident if we try to answer those questions from Habermas’s own writings. As Nancy Fraser puts it, this strategy “necessitates that one read the work in question from the standpoint of an absence; that one extrapolate from things Habermas does say to things he does not” (1985:98-99). Therefore, I suggest following Niklas Luhmann in order to find a starting point from which to analyze the Habermasian project:

> [T]he key formula [the Habermasian discourse principle] states: “Those norms for action are valid, to which all potentially affected persons could agree as participants in a rational discourse.” Every concept of this maxim is carefully explained with the exception of the word “could,” through which Habermas hides the problem. […] The master and the invisible hand will not be replaced. But who determines, and how does he do so, what *could* find rational agreement? How does this decisive operation, on which everything in the postmetaphysical age depends, become juridified? As a result, it also remains unclear, on all levels of the argument, how the conjunctive becomes an indicative, how the potential becomes a reality, or, for example, how power “comes forth” out of the freely discussing civil society, which does not of course exist (Luhmann, 1998:164-165).

By referring to the unexplained possibility that the word *could* places in Habermas’s account of law, Luhmann hints that his theory might be built upon unacknowledged assumptions which once made explicit can show its limitations. In this paper I want flesh-out some of the assumptions concealed by the conditional form in Habermas’s discourse principle and explain how they interfere in the TCA’s explanatory leverage of social traps. If the "ideal speech situation"[2] provides a normative cornerstone for the TCA and the theory of law based on deliberative procedures, it is at the cost of having little to say about less-than-ideal situations such as social traps.

In order to advance this argument, I will first introduce the concept of social traps as it has been treated by cognitive rational choice, historical institutionalism, and particularly Bo Rothstein's institutionalist approach (2005). Rothstein's critique of rationalist and culturalist accounts of cooperation problems will help me draw —in the fourth section of this paper— a parallel with some of Habermas's limitations to explain social transformation in less-than-ideal situations, making the TCA face similar shortcomings when observing social traps.

In the following section I will introduce the TCA and then turn to Anthony Giddens's critique, where he argues that Habermas's normative idealism makes his account of Critical Theory too detached from *praxis*. The absence of a strong theory of praxis as human's "conscious creative activity"[3] hinders the ability of the TCA to accurately observe situations like social traps, since it tends to conceal them under a deterministic explanation based on structural conditions in which individual action has little relevance. From Habermas's perspective, negative structural conditions hamper motivational, cognitive and moral development; they reproduce patterns of distorted communication (communication biased by unequal power relations), and inhibit the materialization of agency as communicative action. This determinism is thus incapable of explaining the role of human action in cases of social transformation. The question that is then raised is whether the TCA would need to consider other forms of agency, which not necessarily

[2] Habermas described the characteristics of the "ideal speech situation" in the following terms: "Participants in argumentation have to presuppose in general that the structure of their communication, by virtue of features that can be described in purely formal terms, excludes all force—whether it arises from within the process of reaching understanding itself or influences it from the outside—except the force of the better argument (and thus that it also excludes, on their part, all motives except that of a cooperative search for the truth). From this perspective argumentation can be conceived as a reflective continuation, with different means, of action, oriented to reaching understanding" (Habermas, 1984:25).

[3] I use the term *praxis* to refer to the individual ability of human beings to actively modify their life conditions. The term *praxis* comes from Marx's *Theses on Feuerbach* (Marx, [1888] in Tucker, 1978), where he insisted that thinking was intertwined with every other human activity: "For Marx, what made human beings different from any other species of animate life was their capacity for conscious creative activity —for practice or *praxis* as he called it— a concept which he used to embrace both thought and life" (Kitching, 1988:26).

occur through discursive practices, to be able to provide a better explanation of social change in less-than-ideal scenarios.

Thirdly, I will discuss what I have called the *virtuous circle of discourse* in Habermas's work, that is, the circular argument claiming that discourse requires specific moral and cognitive competencies in individuals for communication to remain undistorted, and, at the same time, that those skills can only be developed through participation in discourse. At this point I will argue that this circularity reproduces a similar logic to the one operating within social traps, where trust and cooperation are recursively constituted or hindered.

In the fourth section, I will recapitulate the whole argument about why the TCA faces crucial obstacles to observe social traps. I claim that the developmentalist approach behind the discourse principle has three consequences on the TCA: (1) The evolutionary theory that explains the achievement of moral autonomy and a universal ethics of speech ends up offering a deterministic view of cognitive, moral, and interactive development that occurs independently from individual action. As a result, the TCA only recognizes human praxis as discursive practices. (2) The strong distinction between instrumental and communicative action does not consider a concept like "practical consciousness" (Giddens, 1984) or "subjective rationality" (Rothstein, 2005). Such a concept would allow for a more inclusive conception of praxis and a better description of social reproduction and *especially* social transformation. (3) The circularity in Habermas's account of the two functions of deliberation reproduces the reciprocal influence between social structures and individual action. However, by grounding his theory of law upon this circular argument based on ideal situations, Habermas obscures the fact that the same recursive dynamics operate in an opposite direction in less-than-ideal cases (such as social traps), where self-reinforcing conditions systematically hinder action coordination through consensus. Finally, in my conclusions, I will use the previous discussions to present a few issues that a critical methodology should consider in the study of social traps.

Introducing Social Traps

In this first section, I will introduce social traps as situations where the recursive dynamics between structures and agents are particularly significant. Bo Rothstein's (2005) approach to the topic will be useful for these purposes since his argument is based on two main components: (1) an explanation of how social traps are perpetuated through the reciprocal influence between social structures and individual actions, and (2) the role of political institutions as linking mechanisms between them. I will concentrate on the first one of these elements since it speaks directly to some of the problems that are present in Habermas's TCA (which I discuss in the following sections). Along these lines, I will begin by addressing the shortcomings that rationalist and culturalist approaches face in their explanations of social traps. These limitations are central to my argument and I will return to them in my conclusions, since they are also present ―although for different reasons― in the TCA. The second and third parts of this section will focus on Rothstein's argument and the recursive dynamics of agents and structures.

Rationalist and culturalist approaches to social traps

The psychologist John Platt (1973) coined the term "social traps" to describe the kind of situations where cooperation among self-interested, utility maximizing actors is hindered. Nevertheless, the topic has been central among political scientists for the last forty years, and has been addressed under several different names: "*Provision of Public Goods*, *Problem of Collective Action*, *Tragedy of the Commons*, *Prisoners' Dilemma*, and *Social Dilemma* are but a few" (Ostrom in Rothstein, 2005). All these labels indicate circumstances where cooperation among self-interested actors is hindered:

> According to the logic of the social trap, even people with clear preferences for "fair play" will continue their disloyal behavior because they believe, and for good reason, that almost "all other people" are going to keep playing dirty. And, again, this is not because most other people are actually evil and fundamentally disloyal, but because they expect that everyone else will cheat. Changing the situation is thus a matter of changing the worldview of large groups of citizens about the kind of society they live in and how people might conceivably act in that society (Rothstein, 2005:8).

Social traps have frequently been approached through either rationalist or culturalist standpoints that face three major limitations to provide a convincing explanation of social traps. Rothstein (2005:29-30) lucidly points at these interrelated problems: a) both traditions are based upon unrealistic assumptions of human action, b) they are deterministic in their conclusions, being incapable of offering a convincing theory of *praxis*, and c) precisely because of their deterministic imprint they cannot explain social transformation.

"Rationalists begin with assumptions about actors who act deliberately to maximize their advantage... Analysis begins at the level of the individual and culminates in questions about collective actions, choices, and institutions" (Lichbach & Zuckerman, 1997:6). However, this emphasis on utility-maximization adopts an "anemic or thin version of intentionality and interests" (Lichbach in Rothstein, 2005:30). The consequence of such an assumption is that this approach overlooks important variables regarding the way in which individuals deal with uncertainty and incomplete information, how norms and social contexts influence their decision-making processes, and the ways in which strategic calculations about each other's future actions determine their social behaviour in daily life.

On the other hand, culturalists tend to see agents as puppets whose lives are determined by the overwhelming tide of cultural and structural conditions. From this standpoint, culture would be the only explanation for individual action:

> Culture, from this perspective, "is a worldview that explains why and how individuals and groups behave as they do" […]. Obviously, there is hardly any room for such things as intentions, strategic action, not to say deliberative choice, within this perspective. Once the world has been "culturally constructed" for them, individual agents are no longer "agents" in any meaningful sense of the word (Rothstein, 2005:30).

In both approaches, the reductionist view of human action (individual agency only as purposeful-rational action or as an inconsequential pawn of cultural frameworks) leads to a deterministic view of social phenomena. In the context of social traps, the end of the road for both trends would certainly point at the impossibility to escape once a group falls into those dynamics. Although empirical evidence seems to support such a pessimistic prospect showing that social traps tend to endure and have long-lasting effects (North, 1990; North, Summerhill & Weingast, 1999; Putnam, 1993), Rothstein's case study of the labour conflict in Sweden in the 1920s (2005) —along with others (i.e., Ostrom, 1990)— debunks those approaches by looking at a case in which social transformation occurred and opened a way out of the social trap. For rationalist and culturalist approaches, there is very little room for agency, strategy, and choice once a system of incentives or a belief system, respectively, is in place, therefore such an assumption makes them incapable of explaining cases of social transformation.

Bo Rothstein's institutionalist argument

Aware of these criticisms against the dominant literature on social traps and following the tradition of historical institutionalism (North, 1990; Putnam, 1993), Rothstein (2005) has built an argument that solves many of the shortcomings of the previous approaches, particularly because it takes into account the recursive interaction between social structures and individual actors, refusing to give primacy to either rationalism or culturalism (Rothstein, 2005; Ostrom, 2008). On the contrary, he underlines the necessity of combining individual and collective historical perspectives to be able to advance a powerful explanation of social traps. According to him:

> the theory about social traps allows us to link two approaches in the social sciences that are usually widely disparate: those which stress the importance of *historically established social and cultural institutions and norms*, and those which emphasize the importance of *human strategic actions and choices* (Rothstein, 2005:14).

Rothstein's argument is grounded on the combination between, first, the role of culture that individuals pragmatically use as a "tool box" through subjective rationality and, second, an institutional component embodied in an in-depth study of how the universal welfare state affected labour politics in Sweden in the 1920s (Ostrom, 2008:137). He sees this approach as a commitment to a methodological individualism capable of providing a descriptive account of subjective rationality by incorporating the agents' limited computational capacity, the influence of emotions on their behaviour, and the impact of cultural and social contexts on preferences (2005:36). This methodological strategy should then be able to simultaneously observe the circular dynamics of social traps and the possibility for social change.

> New political actions emerge from cognitive mechanisms (although very often not as intended by actors). In other words, political change is *actor based* and hence occurs only when agents take notice and act upon structural changes, other people's actions or new opportunities. However, mechanisms need not be coupled only to instrumental rationality. On the contrary, how mechanisms work in different social settings should be an open empirical question. They may equally be based on emotions, problems of information, or ideology. The implication is that a focus on causal mechanisms makes it necessary to specify how we see the relation between rationality and culture and how they are related at the individual (micro-) level (Rothstein, 2005:34-45).

From this point of view, actors orient their behaviour in relation to the images, beliefs and expectations about how others will behave. In order to understand this kind of strategic behaviour it is necessary to take into account how those perceptions enter into human consciousness and shape political action (Rothstein, 2005:14). An understanding of rationality as a simple instrumental calculation of utility-maximization becomes untenable once we accept that self-interest and rationality are context dependent (Mantzavinos, North & Shariq, 2003; North, 1990). Moreover, empirical research has shown that such an assumption has little power to make predictions on future human conduct (Rothstein, 2005:35). Social traps then cannot be observed as an example of irrational actors undermining their own self-interests due to their non-cooperative behaviour, since what counts as rational action in those situations is determined by the individual perceptions of *who the other is* and *whether he can be trusted or not*. "This assessment may come from many different sources, such as personal knowledge about the individuals in question, culturally determined stereotypes, or memories of how the actors have acted in similar situations in the past" (Rothstein, 2005:15). If those factors affect individual action, it is impossible to determine whether cooperating or cheating is the rational option until a political anthropology manages to grasp the context in which individual action is embedded (Bates, de Figueiredo & Weingast, 1998).

Two ways of understanding subjective rationality: North vs. Swidler

There are two ways in which the notion of subjective rationality can be understood. The first one, following North and the cognitive rational choice tradition, sees it as the result of "incomplete processing of information" by individuals when trying to decipher the environment. The incomplete feedback of information after individuals interact with the prevailing institutions of society leads "to ideas, ideologies and dogmas" which play a major role in human beings' choices (North, 1990:22-23). Therefore actors usually make the most rational choice for their perceived interests under conditions of incomplete information. Moreover, the solutions to cooperation or exchange problems achieved in the past carry over into the present (1990:37) and become recurrent in how people interact. In short, previous institutions, perpetuated through formal or informal incentives and constraints, determine what counts as rational action through the formation of "subjective models of reality" (1990:112).[4] This is the reason why, for North, social transformation can only occur incrementally and is tenuously linked to human action. Change rather consists of "marginal adjustments to the complex of rules, norms, and enforcement that constitute the institutional framework (1990:83)."

The second account of subjective rationality, the one embraced by Rothstein, understands it as the result of knowledgeable agents that possess the practical skills to operate pragmatically in specific social contexts. These actors face the problem of fragmentary information that reinforces beliefs and perceptions about other actors based on past experiences, but also have a margin of action for strategic maneuvering and social transformation. Rothstein borrows this view of "culture as a tool-box" from Ann Swidler:

> A starting point in this approach is that people *know much more culture* (signals, stories, symbols, rituals, etc.) than they actually use. Secondly, there is variation in how they make use of the cultural repertoire that is available to them and they "select within that repertoire what works at the moment" […]. Thirdly, people also differ in "how seriously they take their culture and how richly they deploy it" […]. Lastly, Swidler argues that people sustain a lot of contradictory or uncoordinated cultural codes in their repertoires (Rothstein, 2005:37-38).

I will elaborate on the consequences of these two conceptualizations of subjective rationality in the fourth section of this paper, when I will argue that the Habermasian distinction between instrumental and discursive rationality falls into a similar determinism as the one present in North's work, where the incomplete processing of information (which is not directly translatable to Habermas's notion of distorted communication but operates in an equivalent way) highly reduces the possibility of individual agency.

[4] Although North uses the term "subjective models of *reality*", the oxymoron "subjective models of *rationality*" might be more accurate to portray the challenge to a univocal rationality that ignores how the context conditions responses and decisions.

The recursive interaction of structure and agency in social traps

From the explanation of social traps based on the strategic behaviour of actors guided by subjective rationality, Rothstein claims that we can flesh-out four fundamental characteristics of the social world:

(1) that actors behave strategically —"what people do depends on what they believe others are going to do" (Rothstein, 2005:13)—;

(2) that "individual rationality may very well be collective irrationality" (2005:13), that is, individual actors often cannot or choose not to take into account the unintended consequences that will occur in the societal or group levels as a result of their private self-interested actions;

(3) that "whether or not an action is rational cannot in these types of situations be determined solely by reference to one's individual preferences, but is rather determined by the social context" (2005:13), especially the "collective memory" (2005:37) that determines from past experiences whether the others are trustworthy or not;

(4) that "we cannot rationally forget" past betrayals and deceitful behaviour even if that would increase everyone's will to cooperate (2005:13).

Rothstein is recognizing here the recursive mechanisms that connect individual actions with social structures, arguing that "social and political structures are the result of the aggregation of individual behaviors" while, at the same time, "the particular structural position (or configuration) of agents *vis-à-vis* each other affects the resulting individual behavior" (2005:39). The idea that he tries to advance is that "structural explanations can be reduced to problems of aggregation of individual-level explanations" (2005:39). However, it is essential to clarify that individual behaviour in groups cannot be equated to the behaviour of isolated individuals:

> The fact that actors are in a group is what fundamentally changes the structure of their interaction, their perception of the chances of success, their perception of their safety, their perceptions of gains and losses and their distribution, their emotions, and their concerns of fairness, etc. Most importantly, what changes is their view (their belief system) about what they can expect of the other agents in that group. Will they cooperate or are they more likely to cheat? Can they be trusted or not (Rothstein, 2005:39)?

In other words, the connection between structure and agency manifests itself in social traps in the following way: social structures —such as belief systems or shared perceptions and assumptions of the others— shape individual action by modifying the expectations about future cooperation between actors: will they cooperate or will they cheat? This in turn determines individual action, reinforcing the perceptions of each other that are built into and sustain those belief systems. Social structures are then reproduced every time an actor decides to cooperate or to cheat according to the expectations created by those prevailing structures.

The circularity that links the social structures with individual agency is the same circularity that explains why social traps are situations from which it is extremely difficult to escape. Social trust cannot be built overnight. On the contrary, the collective memory of untrustworthy and deceitful behaviour is hard to erase and tends to reinforce those negative assumptions and expectations about the others. Moreover, social traps become an example of "stable but inefficient equilibria", where although everybody suffers from the lack of cooperation, past experiences make non-cooperative behaviour the most rational option for individual actors.

> Empirical evidence shows that perceptions of *the others* are highly stable and difficult to change. These cognitive or mental maps are often included in enduring cultural socialization processes, where they strongly characterize the worldviews of individual actors with respect to things such as the honesty and competence of state institutions or whether it is reasonable to trust other people in general or specific groups of people (Rothstein, 2005:21).

According to Rothstein, in order to gain a better explanatory leverage of social traps, we need then to consider a broad repertoire of factors that drive human action (2005:36). This perspective must go beyond the utility-maximizing paradigm of instrumental rationality and include those factors that result from social and political structures —i.e., cultural traits, systems of beliefs, institutional (formal and informal) incentives and constraints. A full understanding of how institutions and subjective rationality link particular social structures with certain patterns of individual action would provide a more accurate explanation of why actors choose to perpetuate social traps by entrenching themselves in non-cooperative behaviour. Such an account requires a consideration of agency that goes beyond functionalist notions of purposeful action:

> "Agents are not perfectly rational and fully informed about the world in which they live. They base their decisions on fragmentary information, they have incomplete models of the process they are engaged in, and they may not be especially forward looking. Still, they are not completely irrational: they adjust their behavior based on what they think other agents are going to do, and these expectations are generated endogenously by information about what other agents have done in the past" (Peyton-Young in Rothstein, 2005:37).
>
> What people "think other agents are going to do" is of course something they learn (or make inferences about) from the culture in which they live. Similarly, "what other agents have done in the past" must be seen as agents in a society sharing some sort of "collective memory" about each other as individuals and groups (who are the Serbs, the politicians, the police, the Catholics… and to what extent can they be trusted?) (Rothstein, 2005:37).

Finally, the second component of Rothstein's argument addresses the role of institutions as connecting mechanisms between structures and agents. Institutions are important for Rothstein since "they present incentives, they induce strategy because they make it plausible to calculate what the other agents are likely to do, and, in some cases, they influence ethics and norms" (2005:42). It is at this point where, for our purposes, the most important controversy in Rothstein's work appears. "Are we to understand political institutions as any kind of repetitive behavior that influences political processes or outcomes? Or should we reserve the term 'political institutions' for formal rules that have been decided upon in a political process" (2005:40)? The question about whether to provide a wide or a narrow definition of institutions certainly shapes the trajectory of Rothstein's research and produces the shortcomings that a critical method might overcome. This is not the place to explore such an issue since this paper's objective is not to outline a critical method for social traps, nonetheless I will briefly return to this in my conclusions to suggest some features that a critical observation of social traps would need to consider.

Praxis is only discourse? Only discourse is praxis?

In this section I will first introduce the TCA, emphasizing the concept of the lifeworld and its role in Habermas's explanation of social reproduction. I then follow Giddens in his critique of two key distinctions in Habermas's work: first, the one between instrumental and communicative action, and, second, the one between the lifeworld and the system. I will devote the second part of this section to explain why the lifeworld as an intangible set of rules, resources and expectations cannot be empirically distinguished from the systems that contain it, not even if we recognize the emergence of a formally institutionalized public sphere. This, I will argue later on, is one of the central obstacles that Habermasian Critical Theory faces to address social traps, since it fails to incorporate human agency as a day-to-day factor that intervenes, most of the time in an undiscursive fashion, in processes of social change.

The lifeworld and the Theory of Communicative Action

Since his writings of the late 1960s (Habermas, 1971 [German edition in 1968]; Outhwaite in Ritzer, 2003:229), Habermas tried to provide a normative grounding for Critical Theory that would avoid a substantive foundation (such as Marcuse's distinction between false and real needs, for example) or some kind of "first philosophy" (Marcuse, [1964] 1991; Livesay, 1985:74). In order to achieve this, Habermas developed the notion of the ideal speech situation as the normative cornerstone of the Theory of Communicative Action. By grounding communication upon the ideal speech situation and a "universal pragmatics"[5], Habermas offered a conception of "truth" that looked at itself as "one among several 'validity claims' that can be redeemed in discourse" (Giddens, 1977:143). These concepts became essential to his theory of communicative rationality, since they seemed to help him avoid a transcendental basis of knowledge. Communicative rationality would, in turn, contribute to the construction of a theory

[5] "The task of universal pragmatics is to identify and reconstruct universal conditions of possible understanding […]. In other contexts one also speaks of 'general presuppositions of communication' but I prefer to speak of general presuppositions of communicative action because I take the type of action aimed at reaching understanding to be fundamental. Thus I start from the assumption […] that other forms of social action […] are derivatives of action oriented to reaching understanding "(Habermas, 1979:1).

of human action that “does not posit a self-sufficient subject, confronting an object world, but instead begins from the notion of a symbolically-structural life-world, in which human reflexivity is constituted” (Giddens, 1987:236). Through this strategy, Habermas tried to provide an alternative to the positivist-functionalist emphasis on purposive-rationality which had generated a loss of moral meaning and a diminution of freedom in advanced capitalist societies. Critical Theory, through the TCA, could then pursue its objective of expanding the ability of self-reflection among human subjects to transform asymmetrical conditions of power and domination.

During more than fifty years as a philosopher, Habermas has explored two different tactics to provide a foundation for Critical Theory. The “turn to language” in the 1970s marked the shift from one argument to the other, distinguishing two phases in Habermas’s career (Giddens, 1987:226; Craib, 1992:232-234; Outhwaite in Ritzer, 2003): first, an epistemological approach that culminated in *Knowledge and Human Interests* in 1968 (the English version appeared in 1971); and, from then on, the attempt to ground Critical Theory upon language and communication. The epistemological argument was based on drawing classifications of rationality, knowledge, and science that would correspond to the “elements of the human self-formative process”: labour, interaction and authority/power (Habermas, 1971:196), that is, “the specific fundamental conditions of the possible reproduction and self-constitution of the human species” (Giddens, 1977:138). The purpose of this discussion was to follow the Frankfurt School’s critique of positivism by showing how its view of science had detached knowledge from interests: “‘Scientism’ means science’s belief in itself: that is, the conviction that we can no longer understand science as one form of possible knowledge, but must rather identify knowledge with science” (Habermas, 1971:4). This exclusionary view of knowledge conformed “to only one type of knowledge-constitutive interest, that in the prediction and control of occurrences, or ‘technically exploitable’ knowledge” (Giddens, 1977:139). Along these lines, if this kind of knowledge was associated with labour, instrumental action and the empirical-analytic sciences, interaction would correspond to the search of mutual understanding characteristic of the historical-hermeneutic sciences. For him, both of them claim to separate knowledge from interests and share “the methodological consciousness of describing a structured reality within the horizon of the theoretical attitude" (Habermas, 1971:303). This is the reason why a critical social science should "determine when a theoretical statement grasp invariant regularities of social action as such and when they express ideologically frozen relations of dependence that can in principle be transformed" (Habermas, 1971:310). This critique of ideology should then be concerned with power relations and be able to self-reflect about its own interests.

As Outhwaite says, "Habermas […] then came to feel that the trichotomy of empirical, hermeneutic, and critical sciences was too simplistic, especially in that reflection in the philosophical sense did not necessarily mean emancipation in practice" (Outhwaite in Ritzer, 2003:232). Therefore, in his second period, Habermas took a linguistic turn to pursue the same objectives but through a different strategy: a theory of communicative action based on the analysis of linguistic communication.

> Only a rational agreement which excluded no one and no relevant evidence or argument would provide, in the last resort, a justification of the claims we routinely make and presuppose in our assertions. […] Moreover, if Habermas is right that moral judgments also have cognitive content and are not mere expressions of taste or disguised prescriptions, it also provides a theory of truth for issues of morality and of legitimate political authority. Moral norms are justified if they are what we would still uphold at the end of an ideal process of argumentation. […] The analysis of language-use can thus, Habermas believes, be expanded into a broader theory of communicative action, defined as action oriented by and toward mutual agreement. […] The theory of communicative action, then, underpins a communication theory of morality, law, and democracy, and it is these aspects which have dominated Habermas's most recent work (Outhwaite in Ritzer, 2003:229-230).

As I will discuss in detail in the following section, Habermas used Jean Piaget's stages of cognitive development and Lawrence Kohlberg's stages of moral development to build a universal argument for the theory of communicative action. Through this evolutionary account of moral and interactive competencies, Habermas justified his use of the "ideal speech situation" and "universal pragmatics" (Habermas, 1979) as species-wide, ideal foundations for communicative rationality and communicative action.[6] In the first volume of the *Theory of Communicative Action* (1984), Habermas discussed Evan-Pritchards investigations on the Azande tribe's beliefs on witchcraft, as part of his attempt

[6] Although Habermas has always recognized the inherent idealism of this concepts, he has also argued that the ideal speech situation is "counterfactually presupposed […] by our everyday practice of communication, which is made meaningful by the real or hypothetical prospect of ultimate agreement" (Outhwaite in Ritzer, 2003:229). Moreover, the point of including Piaget's cognitive psychology in his argument is to bring empirical evidence in favour of universal pragmatics and the ideal speech situation as real possibilities for human communication and action.

to defend the universality of the evolutionary argument about the rationalization of wordviews through learning processes. He insisted that the idea of rational worldviews behind the notion of communicative rationality should not be understood in any substantial sense. On the contrary, Habermas followed Piaget arguing that "cognitive development signifies in general *the decentration of an egocentric understanding of the world*" (Habermas, 1984:69, his emphasis), and explained this by making reference to the differentiation of subjective, objective and social worlds. According to Piaget, the growing child draws a "demarcation through the construction of the universe of objects and of the internal universe of the subject" (Piaget cited in Habermas, 1984:68). Furthermore, the external universe is split between the world of objects, on the one hand, and the world of "normatively regulated interpersonal relations on the other" (Habermas, 1984:68), that is, the social world. In sum, the individual's subjective world also becomes aware of other individuals' subjective worlds, making it possible "to adopt in common the perspective of a third person or a non-participant" (1984:69). It is in this context where Habermas introduces the concept of the *Lebenswelt* or lifeworld:

> Subjects acting communicatively always come to an understanding in the horizon of a lifeworld. Their lifeworld is formed from more or less diffuse, always unproblematic, background convictions. This lifeworld background serves as a source of situation definitions that are presupposed by participants as unproblematic. In their interpretive accomplishments the members of a communication community demarcate the one objective world and their intersubjectively shared social world from the subjective worlds of individuals and (other) collectives. The world-concepts and the corresponding validity claims provide the formal scaffolding with which those acting communicatively order problematic contexts of situations, that is, those requiring agreement, in their lifeworld, which is presupposed as unproblematic (Habermas, 1984:70).

Habermas then goes on to identify four formal properties that cultural traditions must have for communicative rationality to operate: a) the cultural tradition must provide differentiated formal concepts, validity claims and basic attitudes for each of the three worlds; b) it "must permit a reflective relation to itself", it should not be fixed as dogma but allow critical revisions; c) it must be open to feedback "in specialized forms of argumentation", fostering the emergence of cultural subsystems "in which traditions take shape that are supported by arguments rendered fluid through permanent criticism but at the same time professionally secured"; and d) the cultural tradition "must interpret the lifeworld in such a way that action oriented to success [...] can be uncoupled from action oriented to reaching understanding", making possible "a societal institutionalization of purposive-rational action for generalized goals, for example, the formation of subsystems, controlled through money and power, for rational economics and rational administration" (Habermas, 1984:71-72). This last condition thus establishes the distinction between the lifeworld and the subsystems through which society operates in a daily basis or, in other words, the distinction between instrumental and communicative action.

Through this line of reasoning, Habermas provides a universalist argument for a procedural concept of communicative rationality without falling into the trap of building itself upon a transcendental cornerstone, that is, a substantive notion of rationality. Such an argument would then maintain the idea of a progression in human societies from pre-conventional to post-conventional forms of morality associated with the development of higher forms of rationality.[7] However, Habermas's critics claim that he does not succeed, and that the developmentalist imprint in Habermas's account of discursive rationality ends up refurbishing a transcendental standpoint that limits human agency to individuals situated in post-conventional societies.

> The consideration of *praxis* recedes from the center of Habermas' thought not because of specific theoretical intentions, but rather as an unintended consequence of his decisions about other theoretical issues: in particular, his effort to avoid lapsing into any form of technocratic instrumentalism by doggedly maintaining the absolute distinction between instrumental and communicative action, and his attempt to establish a "quasi-transcendental" normative foundation for critical theory through his model of an ideal speech situation and his rational reconstruction of the evolution of forms of social integration (Livesay, 1985:69).

[7] The expression "post-conventional societies" comes from Lawrence Kohlberg's explanation of a post-conventional morality: "He described development in terms of conventional moral thinking (the morality of maintaining social norms because they are the way we do things) shifting to postconventional thinking (the morality that rules, roles, laws, and institutions must serve some shareable ideal of cooperation)" (Rest et al., 1999:2). Along these lines, Habermas equated the individual development of moral autonomy to the system's morality at the collective level. Societies would have followed a similar trajectory travelling from pre-conventional to conventional and post-conventional forms of morality: "The disenchantment of religious worldviews not only has the destructive consequence of undermining the 'two kingdoms' of sacred and secular law, and with this the hierarchical subordination to a higher law. It also leads to a reorganization of legal validity, in that it simultaneously transposes the basic concepts of morality and law to a postconventional level. With the distinction between norms and principles of action, with the idea that norms should be generated from principles and by voluntary agreement (*Vereinbarung*), with the concept of the lawmaking power of privately autonomous legal persons, and so on, there develops a notion of norms as positively enacted and hence changeable, yet at the same time criticizable and in need of justification" (Habermas, 1996:72).

Along these lines, Anthony Giddens's most poignant argument against Habermas is the one concerning how he reduces the scope of human agency by not taking into account how lay actors can be knowledgeable and reproduce or transform structures in ways other than discourse. A second criticism (closely related to the former) that Giddens raised against Habermas had to do with his "taxonomic fervour" and the "puritanical formalism" of his writing (1987:242). This criticism goes beyond nitpicking his writing style but has to do with a constant theoretical strategy in Habermas's works, where he recurs to classifications that might be useful in a conceptual level but push his theory further away from empirical evidence. The problems associated with the clear-cut differentiation between, on the one hand, 'labour' and 'interaction', and, on the other, purposive-rational action and communication, in Habermas's epistemological argument (1971) were once again reproduced in the TCA through the subsequent differentiation of, first, instrumental and communicative action and, later on, the economic and political subsystems and the lifeworld. In Giddens's words:

> I am unhappy with your distinction between system and life-world —as I was with the differentiation between 'labour' and 'interaction' which appeared prominently in your earlier work. If, as you say, the separation between system and life-world is methodological, how can it also operate as a substantive distinction within modernized societies? Moreover, your use of systems theory, of notions such as 'steering mechanisms' and so on, seems to do scant justice to the active struggles of individuals and groups out of which history is made (Giddens, 1987:250).

The uncoupling of system and lifeworld: its consequences for an explanation of social transformation

In the second volume of *The Theory of Communicative Action* (1987), Habermas developed a three-part model of modern societies, where they appear divided into political, economic and societal subsystems or arenas. According to Cohen and Arato (1992), the latter could —generally and with some conceptual restrictions— be equivalent to the concepts of lifeworld, civil society, or the Parsonian notion of social community. They insist that such translation cannot be made without distortion, and propose to differentiate two distinct levels within the concept of lifeworld: one that refers to the "reservoir of implicitly known traditions" and background assumptions (1992:428), and another that recognizes the "reproductive processes of cultural transmission, social integration, and socialization" (1992:428). The main point in their argument is that this differentiation of the lifeworld is the result of the modernization process that fostered the emergence of institutions specialized in the aforementioned reproductive processes. Habermas pointed out that the differentiation of the economy from the state in bourgeois societies had been mirrored by a similar differentiation in the lifeworld, "taking the shape of private and public spheres, which stand in a complementary relation to one another" (Habermas, 1987:319-320). In this way, the three-part model turns into a four-part model of society, in which the lifeworld stands as an analytical category —with some minor caveats— at the same level as the economic and political subsystems.

Habermas's notion of the lifeworld comes from Schutz's and Luckman's definition of this concept: "the unquestioned ground of everything given in my experience and the unquestionable frame in which all the problems I have to deal with are located" (Schutz and Luckman cited in Habermas, 1987:131). The lifeworld thus represents a taken-for-granted, inter-subjectively shared background, which includes cultural experiences and communicative interactions, temporarily stabilizing certain presuppositions in order to allow mutual understanding through meaningful utterances (Habermas, 1987:131). Individuals use the lifeworld as the unproblematized —"until further notice", as Luckman says (in Habermas, 1987:130) — basis from which life experiences are interpreted and conceptualized. Therefore, in the words of Simone Chambers:

> Communication is the way we transmit and reproduce our lifeworld. More particularly, we can identify three activities that function as transmitters of the lifeworld: cultural reproduction, through which traditions and cultural meanings are passed down; social integration, through which we recognize norms of cooperation and interaction; and finally, socialization, through which we acquire identities both as collectives and as individuals. These three functions are symbolically mediated. We pass on cultural understandings, learn to live together under certain rules, and form our identities, by talking and communicating with one another (1995:241-242).

Chambers (1995) has analyzed how this shared background is constantly regenerated and transformed. She speaks about situations when cultural justifications are no longer convincing to support a specific normative claim, and thus the claim needs to be regenerated or replaced through the communicative practice of providing reasons to support or dismiss that norm (Chambers, 1995:242). For our purposes, this issue is significant because it illuminates the process through which the lifeworld is constructed and regenerated in a day-to-day basis:

> According to Habermas this process often takes place unreflectively, in what he calls the "negotiation of a new situation definition." The negotiation is informal and partial, and is characterized by a "diffuse, fragile, continuously revised and only momentarily successful communication in which participants rely on problematic and unclarified presuppositions and feel their way from one occasional commonality to the next" (Chambers, 1995:242).

Once again, we see how the lifeworld is mostly constituted through microscopic bargaining and discussion. Partial understandings are created through symbolic interaction in an attempt to solve the normative questions that arise around specific issues. As Habermas and Chambers indicate, these are not discursive practices, but only partially reflective forms of everyday communication where the "force of the better argument" is constantly overridden by threats, bribes, manipulation, lies, etc. (Chambers, 1995:243). However, the negotiation of new situation definitions reproduces the validity of the norms that support the symbolic system of the lifeworld.

Certainly, cultural reproduction, social integration and socialization, happen in every realm of social life and every subsystem of society, especially in the interactions that individuals experience in their daily routines. These transmitters of the lifeworld get into play and are reinforced every time two people meet and interact with each other, regardless if it is in a private household, a church meeting, the stock market, or the parliament house. Therefore, the lifeworld cannot be understood as a synonym for civil society or the societal sphere (not even an imperfect one), since it is an undercurrent of non-concretized rules, norms, practices, conventions and shared meanings, that cuts across all the subsystems despite the specialized languages that in some of them allow the unburdening of communicative practices. This is the reason why the distinctions between labour and interaction, and between instrumental and communicative action, are conceptual categories that do not correspond to reality. If the lifeworld is reproduced within the subsystems, the problem is not just that the subsystems (or the specialized languages of particular subsystems) have colonized the lifeworld (giving primacy to purposive-rational action, for example) but that they are also the grounds in which the lifeworld is reproduced through daily interactions creating the incentives to perpetuate specific behaviour. These spaces foster networks of routinized social interaction where the regeneration of the lifeworld unfolds through situations of co-presence and all sorts of face-to-face interaction, including forms of partial communication. Despite differences in the overarching forms of morality, in the degree of industrial development, in the depth of the separation between the economic and the political spheres, or, if we accept the Habermasian evolutionary model, in their transition from pre-conventional to post-conventional forms of normative justification, all societies provide routinized situations of interaction through which the lifeworld is reproduced.

In sum, Giddens's second critique of Habermas points towards certain limitations that the distinction between the lifeworld and the system suffers when grasping the way in which cultural reproduction, social integration and socialization occur in a daily basis. The inaccuracy of the distinction cannot be solved by acknowledging an overlap among the subsystems or by referring to the colonization of the lifeworld by the economic or political spheres. The lifeworld and the subsystems cannot be seen as different categories at the same level of analysis, rather the subsystems are part of the scenario, of the space where the lifeworld operates. In a way, their relationship might be more accurately described as one between continent (subsystems) and content (lifeworld). Even if the specialized subsystems of the market and public administration develop particular languages to unburden the communicative tasks and allow purposive-rational action for generalized goals, the lifeworld is still being reproduced through the daily interactions that occur within them. This not only challenges Habermas's explanation of social reproduction in different stages of the evolution of societies, but fails to consider how human actions have an effect on social transformation. In other words, the two parallel distinctions (instrumental / communicative action and lifeworld / system) make human agency recede from view in less-than-ideal societies. His account of the uncoupling of lifeworld and system ends up presenting social evolution as a unidirectional path in which every new level of system differentiation is achieved regardless of human action. It is precisely for these reasons that the TCA faces important obstacles to explain the complex dynamics occurring in social traps. Conversely, a more accurate explanation needs a more inclusive conception of praxis than individual agency as action coordination through discourse, and incorporate something similar to what Giddens called "practical consciousness" through the "reflexive monitoring of action" (1984:xxiii). I will come back to this point in the fourth section of this paper.

The “virtuous circle of discourse”

Before moving forward and explaining how a reductionist view of praxis as discourse affects the TCA’s explanatory leverage of social traps, I want to address a second feature in Habermas’s work that also undermines its capacity to engage with less-than-ideal situations. Habermas’s model of discursive democracy seems to work in a self-referencing way, where legitimate norms, laws and institutions derive from deliberative processes and, at the same time, those deliberative processes nurture the cognitive and affective capacities that the citizens require to engage willingly in deliberation. This tautological argumentation refers to what I have called *the virtuous circle of discourse* or *the two functions of deliberation*: (1) providing legitimacy to norms in post-conventional societies, and (2) encouraging the self-transformation of the individual towards more autonomous forms of moral reasoning.

Normativity in post-conventional legal orders

In *Between Facts and Norms*, Habermas (1996) put forward the idea that deliberation provides a source of legitimacy for the construction of legal norms in post-conventional societies. He reversed “Max Weber’s query about how political legitimacy can result from legality”, asking instead “how we can justify the legitimacy of legality” (Benhabib, 1996:725). From this standpoint, Habermas saw human beings as embedded in structures of communication through which they could build together the systems of norms to rule their actions. For the results of those processes of deliberation to be normatively valid, they should have been achieved through the discussion of validity claims that were accepted or rejected depending on “the force of the better argument” (Habermas, 1984). In other words, “just those norms deserve to be valid that could meet with the approval of those potentially affected, insofar as the latter participate in rational discourses” (Habermas, 1996:127).

Essential to his argument is the question: “do we obey the law because of the threat of sanction backed by the state or because we see the law as reasonable and legitimate?” (Dodd, 1997:329). Habermas insists that individuals should be simultaneously subjects and participants in the construction of the law that rules their societies. From that discursive participation would then arise the normative component of law, which, combined with the facticity element (the capacity of the state for coercion), would create a situation where law and politics precede and influence each other.

Habermas follows Talcott Parsons in the belief that actors have behavioural expectations about each other's actions. Therefore, any social order needs to rely on mechanisms of action coordination (1996:139) to stabilize those expectations and prevent anomic action. Those action coordination problems involve either issues of interpersonal conflicts or the pursuit of collective goals, and in both cases that coordination can by driven by value-oriented or interest-governed motivations (1996:139-140). This constructs a two-by-two matrix that offers four types of conflict resolution strategies and collective will formation: consensus and arbitration, as forms of regulation of interpersonal conflicts; and decision by authority and compromise, as strategies for collective goal setting. This allows Habermas to assemble an almost social-evolutionary account on the "the co-original constitution of binding law and political power" (1996:141) that operates in two stages. First, in pre-modern societies a leader who enjoys some kind of transcendental normative authority displays the functions of a judge-king and concentrates the exclusive attributions to interpret the norms of the community, as well as the resources to factually enforce those laws. In the second stage that normative legitimacy and the factual political power are institutionalized into the administrative system: the state. "Not only does law now legitimate political power, power can make use of law as a means of organizing political rule" (1996:142). Nonetheless, in post-conventional societies, there is no longer the transcendental source of authority that had originally legitimized the judge-king to regulate interpersonal conflicts through arbitration or to decide on collective goals through authoritarian decisions. Habermas insists here that modern societies need then the motivating force that comes from consensus, from "discursively produced and intersubjectively shared beliefs" to provide law with the normative content that had once been invested through transcendental sources (Habermas, 1996:149). For Habermas, only through discourse, through communicative power, the transition from sacred law to positive law can maintain that normative component, allowing action coordination to be more than simple compromise and to integrate value-oriented motivations in its daily operation.

Self-transformation of the individual

Consensus through discourse becomes the motivational force that upholds norms and laws in post-conventional societies. "At the same time, Habermas argues that only in a discursive context can moral capacities develop fully" (Warren, 1993:218). It is widely assumed by participatory democrats that engaging in democratic participation is a social experience "likely to develop [in the individual] just those values and capacities that democracy needs to be a viable, thriving, and vibrant system of government" (Warren, 1993:210). In the following, I will present the second function of deliberation through one of Habermas's early attempts to explain how discursive practices operate as sources of individual self-transformation. This explanation was later refined, but not substantially changed, in his *Theory of Communicative Action* (Habermas, 1984; 1987).

In "Moral Development and Ego Identity", Habermas (1979) correlated the ontogenesis of the ego with the notion of interactive competence, trying to explore the cognitive capacities required to engage in communicative action. As Thomas McCarthy mentions in his introduction to *Communication and the Evolution of Society* (Habermas, 1979), that essay is part of a project to distinguish the different dimensions of human development (linguistic, cognitive, interactive, and ego), in order to provide a general explanation of the interdependence of personality structures —in the form of species-wide competences achieved and ordained in a hierarchical sequence of stages— and social structures (McCarthy in Habermas, 1979:xx).

First of all, Habermas recognized three theoretical traditions that had treated the issue of the development of ego identity: "in analytic ego psychology (H. S. Sullivan, Erikson), in cognitive developmental psychology (Piaget, Kohlberg), and in the symbolic interactionist theory of action (Mead, Blumer, Goffman, et al.)" (Habermas, 1979:73). Although these intellectual trends differ from each other in important aspects —and for Habermas none of them "has as yet led to an explanatorily powerful theory of development" (1979:75)—, they share important commonalities that represent the departure point to eventually equate full moral autonomy with the capacities to function within a universal ethics of speech.

Among the commonalities that Habermas identified, three are especially relevant for the purposes of this paper:

> 4. The developmental direction of the formative process is characterized by increasing autonomy. By that I mean the independence that the ego acquires through successful problemsolving, and through growing capabilities for problemsolving in dealing with—
>
> > a) The reality of external nature and of a society that can be controlled from strategic points of view;
> >
> > b) The nonobjectified symbolic structure of a partly internalized culture and society; and
> >
> > c) The internal nature of culturally interpreted needs, of drives that are not amenable to communication, and of the body.
>
> 5. The identity of the ego signifies the competence of a speaking and acting subject to satisfy certain consistency requirements. […] It consists rather in a competence that is formed in social interactions. Identity is produced through *socialization*, that is, through the fact that the growing child first of all integrates itself into a specific social system by appropriating symbolic generalities; it is later secured and developed through *individuation*, that is, precisely, through a growing independence in relation to social systems.
>
> 6. The transposition of external structures into internal structures is an important learning mechanism. […] With this mechanism is connected the further principle of achieving independence —whether from external objects, reference persons, or one's own impulses— by actively repeating what one has at first passively experienced or undergone (Habermas, 1979:74).

These three assumptions are in the core of the idea that a universal ethics of speech can only occur among individuals that have reached the highest stage of moral autonomy and viceversa. Moreover, they already point at the importance of social interactions: first through the integration of external structures and symbolic generalities; and later on with the differentiation of the self from those social systems.

The next couple of steps in Habermas's account of the formation of ego identity are, first, to equate Kohlberg's stages of moral consciousness with an action-theoretic framework (Habermas, 1979:82-83), and, secondly, to integrate in that framework the "reciprocity requirement" that bridges action structures with the stages of moral consciousness (1979:88-89). By doing this, he tried to prove that the cognitive mastery of "general levels of communication" and the "ability to give one's own needs their due in these communication structures" (1979:78) are essential in the development of ego identity. Let me elaborate on each of these moves separately:

Habermas defines moral consciousness as one aspect of ego development, the cognitive ability to make moral judgments (1979:78). Only action conflicts that are capable of consensual resolution (excluding force and "cheap" compromises) classify as morally relevant under this account. Therefore, following Kohlberg, there are six moral stages divided in three hierarchical levels (preconventional, conventional and post-conventional). Each one of these stages is defined by the consequences of action that determine the outcome of moral judgments, going from physical stimuli (avoiding physical punishment and looking for hedonistic rewards) in the first stage, to the capacity of making moral judgments based on abstract ethical principles in the sixth and final stage.

The second move in the argument involves comparing the stages of moral development with general qualifications for role behaviour (1979:82). In this context, behavioural expectations also evolve in three levels, travelling from individual and concrete expectations and actions, to the emergence of social roles adopted from the surrounding symbolic universe, and finally to the capacity to question those social roles. The purpose here is to emphasize how reaching the third level of role behaviour entails the possibility for the individual to distance herself from her own cultural tradition, abandon a role-dependent position, and observe roles and systems of norms as objects of discursive will-formation.

Habermas adds to the symbolic universes of role behaviour and moral judgments the need for specific abilities (interactive competence) to move within those structures (1979:86). Moral consciousness will then be "the ability to make use of interactive competence for consciously processing morally relevant conflicts of action" (1979:88). Besides the degree of moral autonomy achieved in Kohlberg's post-conventional level, agents in the last stages of interactive competence still require certain structural conditions to be able to solve those conflicts of action through consensus. At this point, Habermas argues that reciprocity embodies the structural conditions capable of shaping possible interactions (1979:88). The requirement of reciprocity (whether it is complete or incomplete, that is, whether two actors expect the same thing from each other or hold different expectations based on unequal power-relations) leads to each one of the stages of moral consciousness when it is applied to the levels of role competence.

Habermas's argument has hitherto only focused on cognitive aspects of the ontogenesis of the ego. Nonetheless, he was careful to warn that a general theory of ego development should include affective and motivational development (1979:91). In the last part of his essay, he briefly addressed the motivational side of moral consciousness, the psychodynamics of superego formation and defence mechanisms that can explain "the discrepancies between moral judgment and moral action" (McCarthy in Habermas, 1979:xxi).

> The correlation between levels of interactive competence and stages of moral consciousness […] means that someone who possesses interactive competence at a particular stage will develop a moral consciousness at the same stage, insofar as his motivational structure does not hinder him from maintaining, even under stress, the structures of everyday action in the consensual regulation of action conflicts" (Habermas, 1979:91).

I have described thus far how Habermas draws a logical connection between higher stages of interactive competence and the development of moral autonomy (and thus capacities to engage in discursive participation). However, it remains unclear how the self-transformation function of discourse operates. How exactly do discursive practices, particularly participation in discourse in political settings, develop capacities of autonomy? Warren (1993) unpacks the answer to this question by distinguishing three different levels of analysis in Habermas's argument: (1) "the potentials within social relations for the development of autonomy"; (2) "the specific moral competencies required by situations of political conflict"; and (3) "the motivational force of speech in the direction of autonomy" (Warren, 1993:216).

The first of these levels of analysis relates to the common assumption in all the aforementioned traditions of ego psychology on the social constitution of identity. The individual develops interactive competencies and autonomy through *socialization*, by her integration into a social system and the appropriation of its symbolic generalities; and then through *individuation*, by adopting the ability to distance herself from those social systems and reconstruct her ego identity by discursive means (Habermas, 1979:74). These two processes can only be fully undertaken through social relations.

The second level of analysis is grounded on Habermas's claim that moral development in Kohlberg's theory is already embedded in social relations and "general structures of interaction" (Warren, 1993:218). At this point, reciprocity plays an important role in the construction of the self, since it develops out of mutual recognitions which, depending on whether they are complete or incomplete, push forward the ontogenesis of the ego through the different stages of moral consciousness. Warren argues that since individuals are unlikely to challenge themselves, they must be exposed to others and to the need of discursively justifying their needs and interests (1993:219). This is why, in the context of a discursive model of democracy, moral capacities have to go past Kohlberg's six stages into a seventh moment that Habermas calls "discourse ethics", where generalizable norms of action can only be built discursively (Habermas, 1979:90; Warren, 1993:218-129).

Finally, the third level of analysis poses the question about why individuals should resolve conflicts by means of democratic discourse. Warren follows Habermas in arguing that speech frames cognitive capabilities in the context of the social relationships (that is, the levels of role behaviour) in which the individual is embedded.

> [S]peech intrinsically relates cognitive competence and motivation. To the extent that we deal cognitively with the relations that situate us in the world, we do so through the medium of language. But since language is not private, since it is learned and sustained intersubjectively, we are also motivated to come to understandings with others about the validity of our claims about these relations. Cognitive veracity depends on intersubjective validity (Warren, 1993:220).

Further on, Warren elaborates on the gaps in Habermas's attempt to flesh-out the assumptions of the self-transformation thesis, addressing the necessity to include the affective dimensions of the self in any theory that tries to explain the psychological impact of discursive democracy on the individual (Warren, 1993:221). He begins by bringing up the premise in Habermas's argument that individuals, besides being morally autonomous, would also be willing to solve conflicts through discourse even when other means "might be more satisfying from the perspective of maintaining a psychodynamic balance of desires and impulses" (1993:221). In politics, ideologies might be examples where such an assumption is hard to uphold. Warren, commenting on Habermas's *Knowledge and Human Interests* (1971), compares ideologies with the effects that neurosis has on individuals, where unconscious resistances hinder discourse because it would challenge the identity of the self. Under those circumstances, the cognitive approach to the self-transformation thesis seems to be insufficient, and thus Warren introduces the therapeutic models of communication as a complementary element to Habermas's account on the influence of discursive forms of democracy on the ontogenesis of the ego (Warren, 1993:223).

> Described politically, therapeutic critique involves dispelling the kind of false consciousness that is a matter of cognitive incapacity. *Political enlightenment*, even in Habermas's terms, will mean not just developing cognitive capacities but also transforming character structures in ways that allow individuals to engage in discourse that is not 'distorted' by self-defeating psychodynamics. Discursive democracy, if it is to be workable in a world where character structures are not ideal, will of necessity involve a therapeutic dimension (Warren, 1993:225).

The question is whether therapeutic communication can be directly included in democratic theory since politics are not very similar to a psychoanalyst's office. Discursive democracy is only workable under at least minimum symmetric conditions of complete reciprocity between morally autonomous individuals. These conditions would then provide enough incentives to engage in deliberation. The logical conclusion suggests that the pre-requisites to engage in discourse need to be developed before democratic participation can reach undistorted levels of communication. It is essential to mention here that Habermas is aware of these obstacles and therefore regards therapeutic *critique* not as a form of discourse but as its "presupposition", as an antecedent condition to solve problems of distorted communication before autonomous individuals can engage in discursive forms of normative justification (Warren, 1993:225). At this point appears an inevitable question for any theory concerning false consciousness and cognitive barriers. Rephrasing this question in his on terminology: how can a society move from structures of distorted communication to discourse as the source of communicative power?

The previous paragraphs have attempted to describe the reciprocal effects that the two functions of deliberation have upon each other. Furthermore, low levels of distortion in communication are a necessary prerequisite for societies to gain access into this self-reinforcing form of democratic practices, where: (1) discursive procedures are the source of legitimacy for laws and institutions, and (2) autonomous citizens feel comfortable without transcendental justifications for norms of action, for whom instrumental calculation is not necessarily the prevalent motivation to live up to their collective agreements (the element of coercion and the facticity of law), and who are willing to engage in discourse to define those norms of action without perceiving communication as a threat to their particular identities.

I want to re-describe now the self-reinforcing model of discursive democracy in a different way. Quoting Simone Chambers's remarks on the circularity of the Habermasian account of discursive legitimacy:

> The institutional form of democratic will formation must itself meet with standards of discursive validity. This appears to lead to a circle: The institutional arrangements that make discourse possible must be justified by a discourse. If the mandate to set up a discourse can only be conferred in a discourse, we are left with no means of justifying the initial establishment of discourse (1995:241).

Indeed, I have introduced a further complication by analyzing in parallel the two functions of deliberation. It is not just that "the institutional arrangements that make discourse possible must be justified by a discourse", but also that *the moral and interactive competences that make discourse possible must be developed through discourse*. In a way, what this implies is that structural arrangements under which social interaction occurs constrain the range and form that those interactions can adopt. Socialization processes that depend on how those social interactions happen are thus tilted in certain directions that may or may not stimulate the development of ego identities compatible with discursive justification.

In other words, the circular relationship between individual's identities and competences and discursively constructed norms and institutions can also be seen as the ideal relationship between agents and structures that are reciprocally constituted through discourse. The Habermasian understanding of ego ontogenesis suggests the appearance of individual agency only in the latter stages of moral development, when the individual differentiates herself from those constraining structures and analyzes them in a self-reflective manner. This necessarily poses the question about the point of entry to those dynamics favourable to the establishment of a (discursively) democratic rule of law. Once we observe the negative cases, those where discursive dynamics are not present, we face a typical example of a social trap: structures that promote forms of asymmetrical communication would tend to hinder the progression towards higher stages of moral consciousness among those involved. The scenario that would ideally predate the development of the virtuous circle of deliberation actually seems to move in the opposite direction than that expected by the Habermasian evolutionary argument. It displays a path-dependence where asymmetrical structures, mostly caused by vertical power relations and fragmented social groups, develop constraints that reinforce cognitive barriers and non-discursive habits for action coordination. These dynamics in turn inhibit the competences and motivations that agents would require to discursively modify those structures.

Habermas's ideal explanation of law is challenged when someone tries to observe less-than-ideal societies that have not developed a discursive form of social organization and coordination for collective action. On the contrary, many of these societies rely on the factual element of law (strong coercive power by the state) or have improvised non-democratic mechanisms to provide some sort of normative content to precarious forms of order, mostly based on asymmetrical power relations and authority: i.e., narco-culture in Latin America, patronalism among Mafia organizations, or factional identities in opposition to the antagonistic *other* in situations of ethnic, religious or class conflict. For this reason, in order to observe how the virtuous circle of discourse reproduces the deadlock of social traps, we need to follow Habermas's argument, but instead of looking at the ideal cases where discursive democracy is solid and self-reinforcing, we need to focus on the negative situations where this assumption does not hold. The purpose of Habermas's theory of law is to show how discursive democracy can perpetuate itself once it is established, but the very same reasons that allow this to happen are the ones that inhibit the development of a democratic rule of law in less-than-ideal societies.

The interdependence of personality and social structures that Habermas explains through his thesis of ego ontogenesis starts from the commonalities shared by analytic ego psychology, cognitive developmental psychology, and the symbolic interactionist theory of action. The three points[8] mentioned before refer to one of the central conclusions that Rothstein draws from the logic of social traps (2005:8):

[8] 1) Moral autonomy as growing capabilities for problemsolving when engaging with the object-world, with the symbolic structures of society, and with the internal nature of interpreted needs; 2) the cognitive developmental process travelling from socialization to individuation; and 3) the separation of external and internal structures.

that individuals base their actions on the expectations about others' behaviour. Nevertheless, this kind of strategic behaviour does not necessarily imply the kind of discursive consciousness that Habermas discovers in the last stages of ego development, where the individuation process has occurred and external structures have been internalized and challenged. However, strategic behaviour does entail a degree of self-reflexivity that manifests itself in a different way, under something that would be closer to the idea of subjective rationality (Rothstein, 2005:35) or "practical consciousness" (Giddens, 1984:xxiii).

If we include in our critique the motivational and affective sides of the ego besides the cognitive one, we also observe that strategic behaviour under non-discursive conditions will actually prevent the development of discourse, enhancing the social trap in which nobody is willing to cooperate with the rest. Warren's point on the "general structures of interaction" and the reciprocity requirement through which they shape social relations is particularly useful for this task (Warren, 1993:218). The conditions of reciprocity will affect the trust in the behaviour of others, to the extent that negative or incomplete reciprocity (tainted by power relations, by past experiences of deceit and betrayal or by negative perceptions of the *other*) might push the individual to withdraw from discourse and action coordination and pursue her own self-interests. In the long run, if these reciprocity requirements are not fulfilled, the transition to latter stages of moral development and the acquisition of discursive capacities will be hindered, reinforcing the individual's "taste" for self-interested non-cooperative behaviour.

Lastly, Warren (1993) also asked why individuals would prefer to go through the costly process of discursive forms of conflict resolution, even if they have achieved the higher levels of moral autonomy, and not other strategies that might be less threatening to their personal identities: *i.e.*, bargaining, intimidating, accommodating, or walking out of the discussion. Indeed, the structural conditions that determine social interaction might not provide the incentives to engage in discursive practices. Furthermore, in a social trap situation they would tend to reinforce instrumental calculation as part of strategic action, turning normative questions irrelevant and maintaining order only through the factual element of law. In these situations, order becomes particularly precarious and oscillates between this calculation of costs and benefits regarding the authority's capacity for coercion and the constant attempt to overrun those constraints to maximize private interests. In other words, when we apply Habermas's line of reasoning to less-than-ideal societies, we discover that the same elements that interconnect the virtuous circle of discourse are the ones that make social traps so difficult to solve. Before the virtuous circle of discourse can be achieved, those damaging circularities need to be transformed.

Praxis, the virtuous circle of discourse, and social traps

In the two preceding sections I have tried to point at some features of the TCA that make it incapable of engaging successfully with topics such as social traps. Moreover, these features are all interrelated because they grow from the basic assumptions without which the discourse principle cannot survive. In this section I will tie up my argument by recapitulating how those assumptions lead to three characteristics of the TCA that make it blind to the complexities of social traps.

I began this essay by citing Luhmann's critique of the unexplained conditionality in the discourse principle: "Those norms for action are valid, to which all potentially affected persons could agree as participants in a rational discourse" (Habermas, 1996:107). The conditional introduced through the word *could* is explained by Habermas through the developmentalist process of ego ontogenesis and moral autonomy. Individuals can participate in rational discourse once they achieve the last stages of moral and cognitive development and are capable to situate themselves within a universal ethics of speech (Habermas, 1979:89). I argue that by grounding the possibility of communicative action upon cognitive, moral and interactive competences that are achieved in a unidirectional and univocal path the TCA acquires three characteristics that unfold sequentially:

First, the evolutionary imprint in Habermas's attempt to establish discourse as a normative goal and as the last step of cognitive development produces a deterministic view of social transformation. Habermas's attempts to universalize his model fell again into a quasi-transcendental first philosophy, in which societal change occurs despite human agents. "Instead of finding progressive structures in history, [Habermas] tries to find them in pre-history, in 'anthropologically deep-seated general structures'" (Couzens Hoy in Couzens Hoy & McCarthy, 1994:158). If there is a path in the form of cognitive development towards collective systems of post-conventional morality, we should expect a long-term evolutionary pattern in which actors don't really have a say.

The determinism in Habermas's theory is also symptomatic of the transcendentalism that lingers underneath his adaptation of Piaget's distinction between competences and interactions:

> Since language precedes humans, the structure of language conditions human interaction. Historiography can chart these *interactions*, but only theory, according to Habermas's adaptation of Piaget, can reach more deeply and describe the development of universal *competences* that underlie these interactions (Cozens Hoy in Couzens Hoy & McCarthy; 1994:159).

This shift towards competences seems to help Habermas move away from a transcendentalist set of speculations of universal history and at the same time maintain a cornerstone for his TCA. Nonetheless, this theoretical move only conceals the way in which he still relies upon some kind of univocal notion of "truth" that anyone *could* agree upon under the right conditions. The implications that this move towards *competences* has for the deterministic imprint of the TCA is one of the first reasons why it is not capable to provide a convincing explanation of social traps. An interesting way of developing this point further is by drawing parallels with Douglass North's account of social traps and its inbuilt determinism (1990). In the first section of this paper I mentioned how the notion of subjective rationality can be understood in two different senses: one, along North's argument, that sees it as the result of incomplete information resulting from imperfect feedback processes and human's limited computational capacity; or, secondly, as the strategic manoeuvring of knowledgeable agents within certain structural constraints. If we follow the logical implications of North's line of reasoning, we discover that Habermas falls into similar problems when trying to avoid the determinism of instrumental views of human action. For North, subjective models of rationality are the result of long-term processes of repeated interaction between actors. North argues that:

> institutions basically alter the price individuals pay and hence lead to ideas, ideologies, and dogmas frequently playing a major role in the choices individuals make. […] [T]he subjective and incomplete processing of information plays a critical role in decision making. It accounts for ideology, based upon subjective perceptions of reality, playing a major part in human beings' choices (1990:22-23).

For North, the difference between informal and formal institutions is one of degree; they stretch along a continuum where societies, as they become more complex, formalize the rules of the game (1990:46). He argues that informal constraints reduce uncertainty as compared to a world of no institutions, and are the immediate source of choice in daily interactions, even if there are underlying formal rules (1990:36). These informal constraints are part of the cultural heritage (maybe the result of previous formal rules that have been internalized). The solutions they offered to exchange problems in the past carry over into the present (1990:37) and are recurrent in how people interact. To sum it in a nutshell: “actors make choices based on subjectively derived models that diverge among individuals and the information the actors receive is so incomplete that in most cases these divergent subjective models show no tendency to converge” (North 1990: 17).

A one-dimensional understanding of human motivation and action remains in North’s work, regardless how culturally-sensitive it attempts to be. From this perspective of “subjective models of reality” or rationality, cultural differences and ideologies are only the result of divergent past experiences that contaminate with wrong information purely instrumental calculations. The assumption is that if feedback processes were complete and actors received full information, their subjective models would converge completely because they would also replicate an objective law-like universe.

Habermas’s whole project as a critical theorist has indeed taken the shape of a strong critique against a one-dimensional account of human action: communicative rationality is built precisely as a description of human motivation and action that includes normative and aesthetic principles in an attempt to go beyond mere instrumental motives. However, the developmentalist argument and the distinction between instrumental and communicative rationality seem to operate in a similar way as North’s notion of incomplete and complete information. If North upholds a conventional understanding of rationality and objective truth that overlap under perfect (ideal) conditions, the concepts of universal pragmatics and the ideal speech situation place Habermas in a similar position. For North, the difference between incomplete and complete information and the resulting models of reality is one of degree, while Habermas insists that the difference between instrumental and communicative rationality is one of kind. Habermas’s hesitations to adopt a full-fledged post-foundationalist view of truth and rationality makes him ground his theory upon the ideal speech situation and the notion of “the force of the better argument”, which in turn are supported by the developmental explanation of ego ontogenesis. Under this framework, strategic action guided by instrumental rationality is the result of intermediate stages of cognitive, moral and interactive development, or of adverse motivational structures, whereas communicative action successfully guides collective will-formation and conflict resolution in the last stage of a universal ethics of speech and under ideal conditions of undistorted communication. This is where Habermas, despite all the obvious differences, is closer to North than he would be willing to acknowledge: in the end of the day, how different is a claim “to which all potentially affected persons could agree as participants in a rational discourse” from the idea of increasingly convergent subjective models of reality that result from more complete processes of feedback information? When North speaks about ideal conditions (complete information) does he mean something substantively different from Habermas when he speaks about ideal conditions (undistorted communication)? Are not both talking about the cognitive capacities and the access to information / communication that actors require to be able to self-reflect upon their interests and preferences?

An important difference along these lines needs to be raised: North believes only in incremental change and does not recognize it as the result of human agency. This could only happen if individual actors had access to complete information and an extraordinary computational capacity to use that information to predict every possible unintended consequence. This information should then be used to self-reflectively observe how previous institutions can be tainting their lines of reasoning. Needless to say, such an expectation is unrealistic and thus only an ideal category for North. Habermas, instead, believes that human actors do have a say once they engage in discursive practices. For him, such a computational capacity is not necessary, because there is no objective information from the real world that needs to be sorted out, but utterances that need to be interpreted, modified, dismissed or adopted only through the force of the better argument. Instead, certain moral and cognitive competences are the prerequisites for agency. Habermas insists that the ideal conditions for communicative action, and thus human agency, are achievable or, at least, can be approximated. North doesn’t.

Nevertheless, Habermas still keeps a strong deterministic imprint when he establishes those as necessary conditions for individual agency by understanding human praxis only as discursive participation. In other words, for him, individuals’ capacity to transform structural processes is obstructed until they reach the last levels of moral autonomy and interactive competence, when they can finally intervene self-reflectively upon those structural conditions. The question, once we talk about social traps, remains unresolved: how do we build those conditions of undistorted communication in the first place? If there is so little room for human agency when motivational structures hinder discursive practices, then how can actors change those motivational structures? Moreover, such a poor consideration of *praxis* and individual agency makes Habermasian Critical Theory blind to situations like social traps, where less-than-ideal societies that are not deeply embedded in communicative practices utilize different mechanisms of collective decision-making and conflict-resolution. Under these circumstances, individual actors make use of their agency without recurring to discourse but entrenching themselves in secured positions that are perpetuated in their daily interactions. At the same time, social transformation occurs as part of the incremental changes produced by the unintended consequences of actors that strategically appeal to the interlocked systems of incentives and beliefs to guide their action.

David Couzens Hoy has discussed how Habermas's evolutionary theory fails to move away from the transcendentalism of universal history and the implications that this has for the TCA (in Couzens Hoy & McCarthy, 1994:158-1964). By following Piaget in the distinction between interactions (contingently charted by history) and competences (as statable rules that predate interaction), Habermas's theory tries to reconstruct those rules that intervene in ethical reasoning. I have already discussed how, by making this move and recurring to the developmentalist argument of ego ontogenesis, Habermas abandons "one set of empty speculations, those of universal history, for another, those of his evolutionary vision" (Couzens Hoy in Couzens Hoy and McCarthy, 1994:158-159). However, Couzens Hoy adds in his critique an interesting point to explain why this transcendentalist influence makes Habermas miss the mark of how ethical judgments actually occur. Following Bernard Williams, Couzens Hoy argues that moral judgment does not operate like a linguistic rule that precedes human interaction, but involves something like the Aristotelian concept of *phronesis* or practical wisdom. This means that judgments depend more on social and historical factors rather than universal principles, since ethical action requires more than the knowledge of moral principles and demands "something like the skill of seeing what is called for in a practical situation", which can only be obtained by the individual's situatedness in a specific community (1994:160).

The absence of a relevant consideration of *phronesis* as individual action in the TCA leads us to the second problem that it faces when observing social traps. The dichotomous view of human rationality and action —where instrumental rationality is linked with false-consciousness and the unreflective pursue of perceived self-interests, and communicative rationality is linked with discursive competence and individual agency— pushes Habermas to draw the distinction between lifeworld and systems. Furthermore, the idealized alignment of economic and political subsystems with instrumental action, and of the public sphere with communicative action, results from the absence of a concept like "practical consciousness" or "subjective rationality". A better consideration of *phronesis* would explain how the reproduction and transformation of the intangible rules and resources that guide individual action operate across all the different subsystems of society.

Although Habermas abandoned the distinction between labour and interaction of his first phase because it was too simplistic and incapable to grasp practical situations (Outhwaite in Ritzer, 2003:232), the influence of that line of reasoning remained palpable after his turn to language (Livesay, 1985:67; Giddens, 1987). This is particularly important in Habermas's distinction between instrumental and communicative action, since this dichotomy is still based upon conceptual distinctions that face limitations to explain how social reproduction and transformation work. The three activities that function as transmitters of the lifeworld (cultural reproduction, social integration, and socialization) are put in practice constantly every time two individuals interact with each other. Even those interactions that occur within the economic and political subsystems carry on these tasks through processes of partial communication (Chambers, 1995:242). An attempt to distinguish motivational structures between systems of incentives that guide instrumental action and systems of beliefs that would be contested within the public sphere does not manage to integrate both as complementary factors that shape human action and determine social reproduction and transformation. Not even pointing at the colonization of the lifeworld is enough to achieve this task, since it still assumes that individuals constantly have to choose between these two systems of motivations in their daily lives as two separate sets of structural conditions. In less-than-ideal situations (and I would argue that even in those societies that lead their action coordination through discourse) these two systems are intricately intertwined and operate conjunctively in every realm of society through micro-processes of interaction and daily ethical judgments. If we accept the idea that human action is guided by expectations about others' behaviour, we can then follow Peyton-Young's definition of praxis, which Rothstein borrows (2005:37), to show how systems of incentives and systems of beliefs combine to shape those expectations. This is a fundamental issue that must be considered in any observation of self-reinforcing phenomena such as social traps, since it is the result of that combination what hinders the possibility of cooperation and collective action. A general description of the economic or political subsystems in advanced capitalist societies as reified spheres dominated by purposeful-rational action can only be accurate at the collective level. Individuals that act within those systems are always challenging and bending them through strategic actions that require an understanding of subjective rationality through which people use the cultural repertoire as a tool-box in their benefit (Giddens, 1984; Rothstein, 2005).

> Of course, the realms of work and interaction are not absolutely separable since: a) human labor ordinarily takes place through the medium of social relations which are normatively structured, b) technical rules themselves have a conventional quality, and c) communicative action ordinarily involves not just formal attempts to understand and be understood by others, but also efforts to influence or control their subsequent action (Livesay, 1985:67).

Regardless of how we call this competence of human agents to guide their lives in specific social and historical contexts —practical consciousness, *phronesis*, subjective models of reality, culture as a tool-box—, we need to acknowledge this level of agency if we want to provide a consistent explanation of social reproduction and transformation, which in turn has to be at the core of any approach to social traps.

The claim that civil society (Benhabib, 2002) or the public sphere (Habermas, 1996) are the ideal places where systems of beliefs can be contested, and where, for example, "the resolution of multicultural dilemmas [can be undertaken] through processes of will- and opinion-formation" (Benhabib, 2002:106), should not be justified by drawing strong distinctions between the kind of competences and action-motives that guide behaviour in different realms of social action. An illuminating example of the consequences that this has for the empirical applicability of the TCA has been advanced by the feminist critiques of Habermas:

> Now what are the critical insights and blindspots of this model? Let us attend first to the question of its empirical adequacy. And let us focus, for the time being, on the contrast between "the private sphere of the lifeworld" and the (official) economic system. Consider that this aspect of Habermas' categorical divide between system and lifeworld institutions faithfully mirrors the institutional separation of family and official economy, household and paid workplace, in male-dominated, capitalist societies. It thus has some *prima facie* purchase on empirical social reality. But consider, too, that the characterization of the family as a socially-integrated, symbolic reproduction domain and of the paid workplace, on the other hand, as a system-integrated material reproduction domain tends to exaggerate the differences and occlude the similarities between them (Fraser, 1985:106-107).

Finally, the determinism in the TCA is further enhanced by the circularity of the two functions of deliberation. From this standpoint, there seems to be an insurmountable abyss between, on the one hand, post-conventional societies that successfully establish their forms of collective regulation and engage in conflict-resolution and collective will-formation through discursive practices, and, on the other, those societies that operate through a stable but inefficient equilibrium in which actors are entrenched in self-interested, non-cooperative behaviour. As I pointed out before, the virtuous circle of discourse provides an explanation of how discursive practices can be perpetuated once they are established, at the cost of providing little insight into how situations of distorted communication can be transformed. The recursive influence between social structures and individual agents that permits, in certain cases, the circular relationship between consensus based on discourse and a citizenry composed of morally autonomous individuals, also closes the trap between generalized social distrust and individual non-cooperative behaviour in less-than-ideal situations. It is for this reason that Critical Theory requires a more coherent and unified theory of *praxis* if it wants to address social traps and explain social transformation. Going back to Rothstein (2005), the circular dynamics of social traps, through which individual actors determine their behaviour in relation to their expectations about others' actions, demand a detailed explanation of the recursive mechanisms that connect social structures and individual actions. Why is it that "individual rationality can be collective irrationality"? Under which conditions are expectations stabilized to grant social order? How exactly does collective memory shape individual preferences and resources to determine what will count as an individual rational action in specific contexts? Most importantly: In the cases that have successfully escaped from a social trap, which individual practices and through which unintended consequences pushed social transformation in that direction?

Conclusions

Rothstein's critique of culturalist and rationalist approaches to social traps had to do with three common problems of both traditions: a) their unrealistic assumptions of human action, b) the determinism in their conclusions, and c) their inability to explain social transformation. From my previous discussion of the TCA, I argue that these problems are also present in Habermas's work although for different reasons. The first one, the unrealistic assumptions of human action, appears in Habermas's reduction of human agency to discursive practices and the dichotomous conceptualization of human behaviour as either instrumental or communicative action, which forces something like practical consciousness to recede from view. The absence of such a concept does not become a problem for Habermas to address the recursive dynamics between social structures and individual action in ideal situations through the virtuous circle of discourse. However it does become a problem if he wants to explain the transitions from pre-discursive scenarios to post-conventional societies with moral and legal systems grounded on discourse. Interestingly, Habermas's TCA builds a very strong idealization of human agency through the notion of individual autonomy, but fails to consider how human actions can push a society through the developmental trajectory from pre-conventional to post-conventional forms of morality. The TCA's (a) unrealistic assumption of human action (not considering individuals as knowledgeable agents that strategically deploy culture as a tool-box) leads to its (b) underlying determinism. Such a deterministic imprint in the TCA can adopt two possibilities: one, closer to Habermas's expectations, would suggest that the "anthropologically deep-seated general structures" of language and cognition that predate the individual would push societies towards post-conventional forms of morality regardless of individual action. However, a second and contradictory possibility is raised by his account of the recursive dynamics between agents and structures. This circularity, combined with the absence of a thorough consideration of practical consciousness, seems to generate a path-dependence in the opposite direction, where non-discursive practices and strategic action would be self-reinforcing, closing an inescapable social trap. This is where the third problem, (c) the TCA's inability to explain social transformation, becomes evident and undermines its emancipatory potential as a critical theory.

Stephen Bronner has pointed at the problems that Habermasian fundamental categories face once questions about social or historical specificity are raised, as well as the implications that this has for the project of Critical Theory as a whole:

> A real "postmetaphysical" philosophy of a critical sort would concern itself not with the deconstruction or reconstruction of "truth" but with the specification of those *material* constraints on its pursuit. Neither the legal nor the linguistic theory of Habermas can link the prerequisites for communicative competence or the stages of moral evolution with the reality of compromise, violence, and the structural imbalance of power (2002:209).

An implicit objective of this paper has been to suggest some paths other than the TCA that Critical Theory can follow to achieve a closer connection between its theoretical categories and their applicability in non-ideal situations. If we follow the critique of the TCA that I have presented so far, two questions remain for further research: 1) What can a critical approach contribute to the literature on social traps? And, 2) What should a critical approach to social traps consider in order to offer a better explanatory leverage and hold a stronger emancipatory potential?

I will leave the first of these questions unaddressed, only suggesting the importance of maintaining on the one hand, praxis and social transformation, and on the other, power relations and non-idealistic accounts of human action and social dynamics, in the center of concern. The self-reflexivity that rests at the core of Critical Theory should be its major distinction from the current literature on social traps. Regarding the second of these questions, the previous discussions should have provided a few points to take into account if one wants to develop a critical methodology for the study of social traps:

1. Avoiding quasi-transcendental or evolutionary arguments, which might produce analytical categories that lose their explanatory power once they are contrasted with social and historical specificity.

2. Providing a thorough explanation of the recursive interplay between structure and agency to avoid fallacious assumptions of human action, whether cultural determinism, voluntarism, or atomistic descriptions of unencumbered selves.
3. Building a wide understanding of praxis that includes instrumental, practical and discursive rationality. This is essential to be able to provide an explanation for social transformation in non-ideal situations, or, in other words, to explicate the immanent processes that may or may not lead towards those ideal scenarios.
4. Developing a typology of institutions that includes informal cultural practices and worldviews as determinant factors for human action. This is a major point of discussion in Rothstein's work (2005:40). Although he is aware of the need of providing a definition of institutions as inclusive as possible to be able to incorporate the impact of informal practices in shaping political outcomes, he still believes that maintaining categorical distinctions between political and other kinds of institutions is necessary. At this point, his emphasis on formal institutions (the welfare state in Sweden, for example) makes him overlook other kinds of repetitive behaviour that through daily interactions reproduce or transform social and political structures. A critical approach, not only for purposes of explanatory leverage, but also to be as self-reflexive as possible, needs to develop theoretical strategies to cope with those institutionalized patterns of behaviour without conflating them with formal political institutions.

References

Antonio, R. J. (1981). “Immanent critique as the core of critical theory: its origins and developments in Hegel, Marx and contemporary thought” in *British Journal of Sociology*, Vol. 32, No. 3 (Sept., 1981), pp. 330-345.

Bates, R. H., De Figueiredo, R. J. P., and Weingast, B. R. (1998). “The Politics of Interpretation: Rationality, Culture, and Transition” in Politics & Society, Vol. 26, pp. 221-256.

Baxter, H. (1987). “System and Life-World in Habermas’s ‘Theory of Communicative Action’” in *Theory and Society*, Vol. 16, No. 1 (Jan., 1987), pp. 39-86.

Benhabib, S. (1996). “Review of *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*” in *The American Political Science Review*, Vol. 91, No. 3 (Sep., 1997), pp. 725-726.

Benhabib, S. (2002). *The Claims of Culture: Equality and Diversity in the Global Era*. Princeton: Princeton University Press.

Bohman, J. (1994). “Complexity, Pluralism, and the Constitutional State: On Habermas’s *Faktizität und Geltung*” in *Law & Society Review*, Vol. 28, No. 4 (1994), pp. 897-930.

Bohman, J. (2007). *Democracy across Borders: From Dêmos to Dêmoi*. Cambridge: The MIT Press.

Brandom, R. B. (2000). *Rorty and His Critics*. Malden: Blackwell.

Bronner, S. (2002). *Of Critical Theory and Its Theorists*. New York: Routledge.

Chambers, S. (1995). “Discourse and democratic practices” in White, S. K. (ed.), *The Cambridge Companion to Habermas*. Cambridge: Cambridge University Press, pp. 233-260.

Chriss, J. J. (1998). “Review of *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*” in *Theory and Society*, Vol. 27, No. 3 (Jun., 1998), pp. 417-425.

Cohen, J. L. & Arato, A. (1992). *Civil Society and Political Theory*. Cambridge: The MIT Press.

Couzens Hoy, D. and McCarthy, T. (1994). *Critical Theory*. Oxford: Blackwell.

Craib, I. (1992). *Modern Social Theory: From Parsons to Habermas*. New York: Harvester / Wheatsheaf.

Dodd, N. (1997). “Review of *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*” in *The British Journal of Sociology*, Vol. 48, No. 2 (Jun., 1997), pp. 329-330.

Dryzek, J. S. (1994). *Discursive Democracy: Politics, Policy, and Political Science*. Cambridge: Cambridge University Press.

Dryzek, J. S. (2002). *Deliberative Democracy and Beyond. Liberals, Critics, Contestations*. Oxford: Oxford University Press.

Elster, J., (ed.). (1998). *Deliberative Democracy*. Cambridge: Cambridge University Press.

Fraser, N. (1985). “What’s Critical about Critical Theory? The Case of Habermas and Gender” in *New German Critique*, No. 38 Special Issue on Jürgen Habermas (Spring-Summer, 1985), pp. 97-131.

Geuss, R. (1981). *The Idea of a Critical Theory. Habermas and the Frankfurt School*. Cambridge: Cambridge University Press.

Giddens, A. (1977). *Studies in Social and Political Theory*. New York: Basic Books.

Giddens, A. (1984). *The Constitution of Society*. Berkeley and Los Angeles: The University of California Press.

Giddens, A. (1987). *Social Theory and Modern Sociology*. Oxford: Polity Press.

Goodin, R. E., and Niemeyer, S. J. (2003). "When does deliberation begin? Internal reflection versus public discussion in deliberative democracy" in *Political Studies* Vol. 51, No. 4.

Gutmann, A. and Thompson, D. (2004). *Why Deliberative Democracy?* Princeton: Princeton University Press.

Giddens, A. (1987). *Social Theory and Modern Sociology*. Stanford: Stanford University Press.

Habermas, J. (1971). *Knowledge and Human Interests*. Boston: Beacon Press.

Habermas, J. (1979). *Communication and the Evolution of Society*. Boston: Beacon Press.

Habermas, J. (1984). *The Theory of Communicative Action, vol. 1*. Boston: Beacon Press.

Habermas, J. (1987). *The Theory of Communicative Action, vol. 2*. Boston: Beacon Press.

Habermas, J. (1989). *Jürgen Habermas* on *Society and Politics: A Reader*. Boston: Beacon Press.

Habermas, J. (1996). *Between Facts and Norms. Contributions to a Discourse Theory of Law and Democracy*. Cambridge: The MIT Press.

Habermas, J. (1998). *The Inclusion of the Other: Studies in Political Theory*. Cambridge:MIT Press.

Held, D. (2006). *Models of Democracy*. Stanford: Stanford University Press.

Horkheimer, M. and Adorno, T. (2002). *Dialectic of Enlightenment*. Stanford: Stanford University Press.

Kitching, G. N. (1988). *Karl Marx and the Philosophy of Praxis*. London:Routledge.

Lichbach, M.I. & Zuckerman A.S. (2006). *Comparative Politics: Rationality, Culture and Structure*. Cambridge: Cambridge University Press

Livesay, J. (1985). "Normative Grounding and Praxis: Habermas, Giddens, and a Contradiction within Critical Theory" in *Sociological Theory*, Vol. 3, No. 2 (Autumn, 1985), pp. 66-76.

Luhmann, N. (1998). "*Quod Omnes Tangit*: Remarks on Jürgen Habermas's Legal Theory" in Rosenfeld, M., and Arato, A. (eds), *Habermas on law and democracy: Critical exchanges*. Berkeley and Los Angeles: University of California Press.

Mantzavinos, C., North, D. C., and Shariq, S. (2003). "Learning, Institutions and Economic Performance". Preprints of the Max Planck Institute for Research on Collective Goods, Bonn, 2003/13.

Marcuse, H. (1991) *One-Dimensional Man: Studies in the Ideology of Advanced Industrial Society*. Boston: Beacon Press.

May, T. (1996). *Situating Social Theory*. Buckingham: Open University Press.

Nagel, M. (2008). “What if Habermas Went Native?” in *Peace Studies Journal*, Vol. 1, No. 1 (Fall, 2008), pp. 1-12.

Niemeyer, S. J., and Dryzek, J. S. (2007) "The Ends of Deliberation: Metaconsensus and Intersubjective Rationality as Deliberative Ideals" in *Swiss Political Science Review,* Vol. 13, no. 4, pp. 497–526.

North, D. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.

North, D. C., Summerhill, W., and Weingast, B. (2000). “Order, Disorder and Economic Change: Latin America vs. North America” in Bueno de Mesquita, B., and Root, H. (eds.), *Governing for Prosperity*. New Haven: Yale University Press.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge University Press

Ostrom, E. (2008). “Review of *Social Traps and the Problem of Trust*” in Political Psychology, Vol. 29, No. 1 (Feb., 2008).

Platt, J. (1973). “Social Traps” in *American Psychologist*, Vol. 28, pp. 641-651.

Prasad, P. (2005). *Crafting Qualitative Research. Working in the Postpositivist Traditions*. New York: M. E. Sharpe.

Putnam, R. (1993). *Making Democracy Work. Civic Traditions in Modern Italy*. Princeton: Princeton University Press.

Rest, J. *et al.* (1999). *Postconventional Moral Thinking. A Neo-Kohlbergian Approach*. Mahwah and London: Lawrence Erlbaum Associates.

Ritzer, G. (2003). *The Blackwell Companion to Major Contemporary Social Theorists*. London: Blackwell Publishing.

Rothstein, B. (2005). *Social Traps and the Problem of Trust*. Cambridge: Cambridge University Press.

Seidman, S. (ed.). (1989). *Jürgen Habermas on Politics and Society: A Reader*. Boston: Beacon Press.

Tucker, R. C. (ed.). (1978). *The Marx-Engels Reader*. New York: W. W. Norton & Company.

Van Evera, Stephen. (1997) *Guide to methods for students of political science*. Ithaca: Cornell University Press.

Warren, M. E. (1992). “Democratic Theory and Self-Transformation” in *The American Political Science Review*, 86:8-23.

Warren, M. E. (1993). “Can Participatory Democracy Produce Better Selves? Psychological Dimensions of Habermas’s Discursive Model of Democracy” in *Political Psychology*, Vol.14, No. 2 (Jun. 1993), pp. 209-234.

Warren, M. E. (1995). “The self in discursive democracy” in White, S. K. (ed.), *The Cambridge Companion to Habermas*. Cambridge: Cambridge University Press, pp. 167-200.

Warren, M. E. (2000). *Democracy and Association*. Princeton: Princeton University Press.

Willet, C. (2001). *Soul of justice : social bonds and racial hubris*. Ithaca: Cornell University Press.

Wiggershaus, R. (1994). *The Frankfurt School. Its History, Theories, and Political Significance*. Cambridge: The MIT Press.

Wren, T.E. (1990). *The Moral Domain. Essays in the Ongoing Discussion between Philosophy and the Social Sciences*. Cambridge: The MIT Press.

Zanetti, L. A. (1997). "Advancing Praxis: Connecting Critical Theory with Practice in Public Administration" in *The American Review of Public Administration*, Vol. 27, No. 2 (June, 1997), pp. 145-167.